mathematics

MDPI

Review

Large Language Models for Multimodal Fact-Checking: Methodological Advances and Persistent Challenges

Tiantian Zhu [1,2], Zhicong Wang [1], Xiongnan Jin [1,2], Qingcai Chen [3,4], Jun Zhang [1,2], Yumeng Liu [5], Zexuan Zhu [1,2,*] and Jianqiang Li [1,2,*]

[1] School of Artificial Intelligence, Shenzhen University, Shenzhen 518060, China; tiantianzhu@szu.edu.cn (T.Z.); 2410673032@mails.szu.edu.cn (Z.W.); xiongnanjin@szu.edu.cn (X.J.); junzhang@szu.edu.cn (J.Z.)
[2] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China
[3] Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China; qingcai.chen@hit.edu.cn
[4] Peng Cheng Laboratory, Shenzhen 518052, China
[5] College of Big Data, Shenzhen Technology University, Shenzhen 518118, China; liuyumeng@sztu.edu.cn
* Correspondence: zhuzx@szu.edu.cn (Z.Z.); lijq@szu.edu.cn (J.L.)

Abstract

Fact-checking aims to assign truth values to claims within a specified context. Unimodal fact-checking methods, due to modality fragmentation and inadequate adaptiveness, struggle in complex scenarios where multimodal misinformation is interconnected. The rapid advancement of generative large language models (LLMs) has provided a novel technological paradigm for multimodal fact-checking. With their capabilities in deep semantic understanding and cross-modal alignment, LLMs can significantly enhance the accuracy, interpretability, and generalizability of misinformation detection. Synthesizing representative studies from 2018 to 2026 retrieved via a multi-dimensional keyword search, two-step screening, and citation chaining, we comprehensively survey the subtask arrangements, LLM-based methodologies, and relevant datasets for both unimodal and multimodal fact-checking tasks. Following the standard fact-checking pipeline, we comprehensively review existing methods across four key stages, namely *Claim Detection and Extraction*, *Evidence Retrieval*, *Verdict Prediction*, and *Justification Production*, providing a structured perspective for method comparison and evaluation. Our synthesis indicates that while multimodal large language models (MLLMs) enhance cross-modal understanding and verification flexibility, fine-grained visual–textual consistency and evidential sufficiency remain critical challenges. We provide a comprehensive survey of fact-checking research and offer insights into the development of next-generation fact-checking systems expected to be more robust, scalable, and adaptable across diverse information environments.

Keywords: multimodal; fact-checking; large language models; survey

MSC: 68T50; 68T07; 68T45; 68P20

Check for updates

Academic Editor: David N. Olivieri

Received: 9 September 2026
Revised: 29 October 2026
Accepted: 4 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The exponential growth of the Internet has enabled information dissemination at unprecedented velocity and volume. This advancement, however, has triggered widespread disinformation from human and algorithmic sources, fostering societal discord. Although manual verification maintains high reliability, its labor-intensive nature makes it inadequate

to address the dynamics of information explosion. These constraints have consequently stimulated substantial academic interest in automated fact-checking mechanisms.

Fact-checking establishes a truth-value assignment framework for contextualized claims through evidentiary reasoning [1]. The core mechanism primarily involves systematic retrieval and logical evaluation of evidentiary support to determine the verdict of a claim. Conventional manual fact-checking is a multi-stage workflow that involves evidence aggregation, source credibility evaluation, and comparative analysis. Due to its labor-intensive nature and low verification coverage, this manual method is ultimately futile in the face of the explosive growth of information [2]. In response to the explosive growth in information volume, academia has proposed semi-automated fact-checking systems that automate ranking, retrieval, and visualization [3–6]. However, they still depend on human judgment for final verdicts [7,8]. After semi-automated systems, pre-trained language models (PLMs) enabled fully automated fact-checking by learning claim–evidence representations in a supervised manner [9]. However, these approaches still relied heavily on task-specific training data and exhibited limited generalization [10].

Breakthroughs in generative large language models (LLMs) have catalyzed intensive research efforts, particularly in developing LLM-driven fact-checking frameworks that leverage neural reasoning capabilities to assess information credibility. LLM-based methods perform well in zero-shot or few-shot settings and do not require retraining [2,11,12], while also getting rid of the limitations of manual processing and human–computer interaction. Such methods can support end-to-end claim parsing and evidence synthesis [13], while also combining retrieval-enhanced evidence with explicit intermediate reasoning steps [14]. These steps make the output easier to inspect, but they do not necessarily explain why the model made the prediction. The model may produce different reasoning steps for the same verdict, omit key evidence, or generate a reasoning chain shaped more by the prompt format than by the evidence that actually influenced the decision [15]. In addition, LLM-based methods significantly outperform manual detection in both processing efficiency and evidence retrieval capability, as they can rapidly sift through vast amounts of information [13,16].

Early approaches to fact-checking predominantly focused on unimodal frameworks operating within textual evidence-verification paradigms, while approaches since 2021 have increasingly explored multimodal fact-checking. Given the inherent multimodality and dynamic evolution of contemporary information ecosystems, unimodal approaches prove fundamentally inadequate for cross-modal misinformation mitigation [17]. Figure 1 presents an example involving a Facebook post (Facebook post: https://perma.cc/HS6S-B9SJ, accessed on 4 June 2026) that claimed to show a flock of sheep "freshly dyed" in multiple colors by a farmer in Scotland. The claim can be checked by conducting a reverse image search, which identifies the original image on an agricultural website (Agricultural website: https://www.ars.usda.gov/oc/images/photos/apr12/k4166-5/, accessed on 4 June 2026). In the original source image, the same flock appears white rather than multicolored. This cross-source comparison supports the conclusion that the circulated version was digitally manipulated and that the Facebook claim is not supported by the visual evidence (Source: https://factcheck.afp.com/doc.afp.com.9M32U2, accessed on 4 June 2026). Multimodal content is perceived as more credible by the public than purely textual content [18], and multimodal information has a stronger visual impact and spreads faster than plain text [19]. The emergence of multimodal large language models (MLLMs) has enabled cross-modal reasoning [20], catalyzing methodological innovations in multimodal fact-checking methods. Some studies have further evaluated MLLMs as real-world fact-checkers, analyzing their capabilities in verdict prediction, explanation generation, confidence estimation, and failure modes [21]. However, understanding these

limitations and deploying generative models in real-world verification requires looking beyond individual benchmark metrics to consider the complete verification lifecycle.

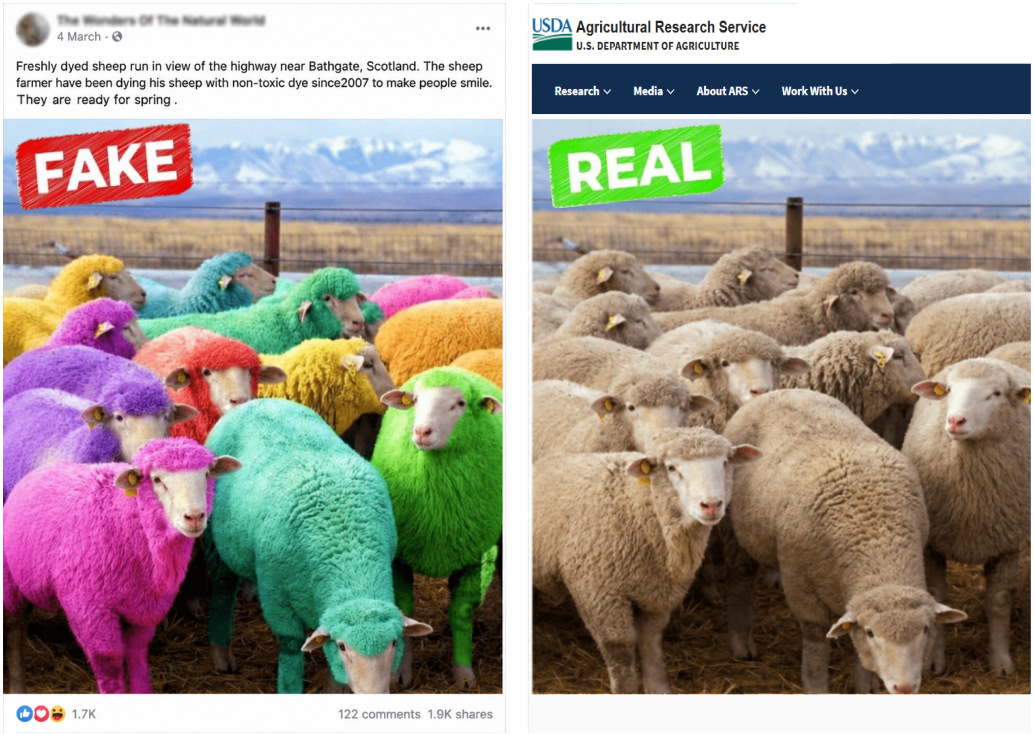


Figure 1. An example of a fake claim: the (**left panel**) shows a Facebook post claiming that the sheep were dyed in Scotland, while the (**right panel**) shows the original image from an agricultural website, where the same flock appears white rather than multicolored.

We comprehensively investigate LLM-based fact-checking methodologies through three principal contributions: (1) a descriptive review of fact-checking strategies published between 2018 and 2026 across the four key stages of the pipeline, compiling benchmark datasets and state-of-the-art methods (a GitHub repository is available at https://github.com/zhicong992-droid/LLMs-for-Multimodal-Fact-Checking (accessed on 3 November 2026), including links to the datasets and SOTA methods discussed in this survey); (2) an analytical synthesis of unimodal, multimodal, and LLM/MLLM-based approaches, clarifying their evolving roles and grounding mechanisms; and (3) an original empirical case study across four representative benchmarks. This evaluation directly quantifies latency, token expenditure, and accuracy trade-offs, providing actionable insights for practical deployment.

2. Related Surveys

2.1. Comparison with Existing Surveys

Zellers et al. [22], Lee et al. [23] are among the earliest works to explore the potential of PLMs for fact-checking tasks. Subsequent studies, including surveys such as Saleh et al. [24], extensively review the use of natural language processing (NLP) algorithms for automated fact-checking. Since 2023, the rapid development of LLMs has driven significant progress in this field. Several surveys have investigated the application of LLMs in this field. The survey by Vykopal et al. [25], focusing on LLMs in automated fact-checking, is closely related to ours but leaves the multimodal domain unexplored. Beyond general surveys, some studies focus on specific perspectives of LLM-based fact-checking. For instance, Quelle and Bovet [26] find that providing contextual information enhances the performance of LLM-based agents when used for fact-checking. Eldifrawi et al. [27]

present a taxonomy of the justification production process and examine the capabilities of LLMs in this context, which focuses solely on a subtask of fact-checking. Moreover, several surveys [28–33] extend prior unimodal coverage by considering additional information modalities. However, they do not provide a comprehensive overview that simultaneously organizes both multimodal approaches and LLM-based methods, as presented in this survey.

As highlighted in Table 1, although prior surveys examine various aspects of fact-checking, they remain fragmented, as none simultaneously covers the complete four-stage pipeline, multimodal data, and LLM/MLLM paradigms. They either focus on isolated subtasks such as verdict classification alone [24,34] and justification in isolation [27], or restrict their scope to unimodal textual pipelines [11,25]. Crucially, this fragmentation leads existing reviews to overlook inter-stage information flow and error cascading across the four pipeline stages. In practice, ambiguous claim extraction misdirects evidence retrieval, incomplete or conflicting evidence skews verdict classification, and flawed verdicts inevitably produce unfaithful justifications. Consequently, three fundamental scientific issues remain insufficiently addressed across the prior literature: (1) *Cross-Stage Error Cascading*: how uncertainties and alignment errors introduced during multimodal claim extraction propagate through retrieval to compromise downstream verdict and justification faithfulness. (2) *Evidence Provenance vs. Hallucination*: how models can ground reasoning in incomplete or conflicting multimodal evidence rather than confabulating plausible rationales from parametric memory. (3) *Verification Complexity vs. Latency*: how to balance the fidelity of cross-modal reasoning against the substantial computational and time costs of MLLMs. Rather than evaluating subtasks in isolation, our work introduces a unified analytical framework to investigate these cross-stage dependencies, review emerging mitigation strategies, and analyze the practical trade-offs between verification rigor and computational efficiency.

Table 1. Comparison of related surveys on automated fact-checking. TimeCov. denotes temporal coverage; Pipe. denotes covered fact-checking pipeline stages (CD: Claim Detection; ER: Evidence Retrieval; VP: Verdict Prediction; JP: Justification Production); Multi. denotes multimodal coverage; LLM denotes coverage of LLMs or MLLMs; SOTA denotes whether state-of-the-art performance is compared; Data denotes the number of compiled datasets; Refs denotes the approximate number of references. The symbols $\surd$ and $\times$ indicate that the corresponding aspect is covered and not covered, respectively.

Survey	Year	TimeCov.	Pipe.	Multi.	LLM	SOTA	Data	Refs
Guo et al. [11]	2022	2014–2021	CD + ER + VP + JP	$\times$	$\times$	$\times$	53	∼260
Alam et al. [30]	2022	2015–2022	ER + VP	$\surd$	$\times$	$\times$	38	∼210
Tufchi et al. [28]	2023	2017–2023	ER + VP	$\surd$	$\times$	$\times$	20	∼170
Akhtar et al. [29]	2023	2018–2023	CD + ER + VP + JP	$\surd$	$\times$	$\times$	52	∼150
Saleh et al. [24]	2024	2018–2024	VP	$\times$	$\times$	$\times$	$\times$	∼30
Yang et al. [32]	2024	2018–2023	ER + VP	$\surd$	$\times$	$\times$	9	∼70
Vykopal et al. [25]	2024	2020–2024	CD + ER + VP + JP	$\times$	$\surd$	$\times$	64	∼130
Eldifrawi et al. [27]	2024	2018–2024	CD + ER + VP + JP	$\surd$	Partial	$\times$	12	∼70
Lin et al. [34]	2025	2019–2025	VP	$\surd$	$\surd$	$\times$	66	∼430
Ours	2026	2018–2026	CD + ER + VP + JP	$\surd$	$\surd$	$\surd$	136	∼470

2.2. Scope and Literature Selection

This work presents a structured survey rather than a Systematic Literature Review (SLR) or a Systematic Mapping Study (SMS). We do not adopt an SLR/SMS framework because studies in this field vary widely in their task definitions, modalities, datasets, model architectures, and evaluation protocols. Moreover, the field is evolving rapidly

with the development of LLMs and MLLMs. Together, these factors make direct statistical synthesis difficult.

We regarded a study as seminal if it introduced a foundational task formulation, dataset, benchmark, or methodological framework that was subsequently adopted in related research. To retrieve the candidate literature available through June 2026 across Google Scholar, DBLP, and major academic databases, we combined search terms across four dimensions, covering verification tasks ("fact-checking", "claim verification"), the four pipeline stages ("claim detection", "claim extraction", "evidence retrieval", "verdict prediction", and "justification production"), modalities ("multimodal", "vision-language", "video", and "audio"), and model families ("LLMs" and "MLLMs"). We conducted a two-step screening over titles, abstracts, and full texts to select peer-reviewed and high-impact papers addressing explicit pipeline stages. Records were excluded based on four criteria: (1) thematic divergence, such as rumor propagation or standalone stance detection without veracity verification; (2) methodological mismatch, such as generic retrieval or vision studies lacking a fact-checking objective; (3) dataset mismatch, such as benchmarks evaluating tasks outside the fact-checking pipeline stages; and (4) redundant preprints superseded by peer-reviewed publications. To ensure balanced coverage of principal methods and benchmarks, keyword searches were complemented by forward and backward citation chaining from seminal works and benchmark leaderboards. In total, this process captured all primary model paradigms across 136 compiled datasets and approximately 470 representative studies published between 2018 and 2026. Although the selection is not exhaustive, it covers the principal research directions within the scope of our survey.

3. Task Formulation

To ensure consistency, we standardize key terminology across the survey. Multimodal fact-checking assesses claim veracity across multiple modalities. We distinguish text-only LLMs from cross-modal MLLMs. Claim detection identifies check-worthy content, claim extraction formulates explicit verifiable propositions, and evidence retrieval operationalizes evidence gathering.

This section introduces the subtask arrangement of automated fact-checking, which sequentially encompasses *claim detection and extraction* [35], *evidence retrieval* [36], *verdict prediction* [37], and *justification production* [38].

3.1. Claim Detection and Extraction

A claim is a statement that asserts a factual proposition about the world and can be verified against evidence. Claim detection aims to find claims that are *checkable* and *check-worthy* [39]. Here, checkable refers to factually verifiable claims, while check-worthy denotes those important factual claims that merit verification efforts. Input space:

A content item x containing text, image, video, audio, metadata, or their combinations. Output space: A set of claim-bearing units, such as text spans, image regions, video/audio segments, metadata fields, or cross-modal units. Success criteria: The detected units should be factual, checkable, and check-worthy [39]. Evaluation regime: Claim detection is evaluated with precision, recall, and F1 at the item, span, segment, or region level. However, these metrics evaluate text spans and image regions in isolation. They measure whether a bounding box or text span is flagged, but cannot verify whether the visual cue and textual claim actually refer to the same real-world event.

Claim extraction complements claim detection by converting the selected content into explicit, checkable claims [40]. In multimodal scenarios, claim extraction is the process of identifying the underlying claim expressed by multimodal content. This includes cases where the claim is implicit or distributed across multiple modalities [41,42]. Input space:

The selected claim-bearing content and its original context. Output space: Explicit natural-language claims. Success criteria: The extracted claims should be faithful to the source, unambiguous, and verifiable [42]. Evaluation regime: Claim extraction is evaluated by semantic equivalence, factual consistency, checkability, and clarity. Yet, because these metrics only compare text strings, they cannot detect when an extracted claim ignores key visual context, such as background landmarks, or invents visual details not present in the original media. Representative methods include text-only claim detection and extraction [42,43], LLM-based claim rewriting and decomposition [44], early multimodal fusion [45,46], Transformer-based multimodal fusion [47,48], and LLM/MLLM-based multimodal claim formulation [29].

3.2. Evidence Retrieval

Evidence retrieval aims to collect materials related to the claim from available resources to support subsequent claim verification. These materials include text, tables, knowledge bases (KB), and relevant metadata. In multimodal scenarios, evidence retrieval goes beyond text and also gathers visual evidence such as images and video frames [49]. This requires retrieving information across modalities and matching textual content with visual content. Input space: A claim and available evidence resources, which may include text corpora, tables, knowledge bases, images, videos, metadata, or their combinations. Output space: A ranked or selected set of evidence items related to the claim. Success criteria: The retrieved evidence should be relevant, reliable, and sufficient for subsequent claim verification [29]. Evaluation regime: Evidence retrieval is evaluated with retrieval metrics such as precision@k, recall@k, mean reciprocal rank, and evidence recall [50]. In practice, these metrics reward surface visual similarity rather than factual proof. A model can achieve high recall simply by retrieving near-duplicate images or matching video scenes, even when those files provide zero factual evidence to confirm or refute the claim. Major improvement strategies include adaptive query construction [51–55], cross-modal retrieval for visual or audiovisual evidence [56], retrieval-augmented generation (RAG) for iterative evidence seeking [57,58], and evidence consolidation with conflict resolution [59,60].

3.3. Verdict Prediction

Verdict prediction evaluates a claim in its context, judges it against the available evidence, and assigns a truthfulness label from a predefined label set [39]. Depending on the scenario or dataset, the label set may be binary or multi-class. Because many claims cannot be conclusively resolved with the information at hand, datasets often include an explicit uncertain label [51]. RealFactBench [21] further introduces an "Unknown Rate" metric to evaluate whether models appropriately abstain when the evidence is insufficient.

Verdict classification. Given a claim and retrieved evidence, this task determines whether the evidence supports the claim, refutes it, or is insufficient for verification [51]. Input space: A claim, its context, and the available evidence. Output space: A truthfulness label, such as supported, refuted, uncertain, or dataset-specific variants. Success criteria: The label should be consistent with the claim and supported by the available evidence. Evaluation regime: Verdict prediction is evaluated with accuracy, precision, recall, or macro-F1. Crucially, these metrics cannot determine whether a model actually examined the visual evidence. A model can achieve high accuracy simply by memorizing textual patterns or guessing from the claim alone, completely ignoring manipulated or out-of-context images. Typical methods include supervised claim–evidence classifiers [51,61], direct LLM-based verification [62,63], and indirect reasoning with chain-of-thought (CoT), predefined question answering (QA), or verification planning [14,64,65].

In multimodal settings, misinformation can stem from the claim, the multimodal content, or their combination. Accordingly, task formulations often complement verdict labeling with manipulation and out-of-context detection [66,67]. *Manipulation detection.* Manipulation detection aims to assess whether the multimodal content has been synthesized or substantially altered in a way that changes its evidentiary content [66]. Input space: Multimodal content and metadata. Output space: A label indicating whether the content is authentic or manipulated. Success criteria: The prediction should correctly identify synthetic or altered content. Evaluation regime: Manipulation detection is evaluated with accuracy, precision, recall, or macro-F1. Typical methods include modality-specific forensic detectors [68–71], as well as MLLM-based manipulated content detection and localization [72].

Out-of-context detection. Out-of-context detection identifies whether authentic multimodal content is reused or paired with an incorrect context, such as a different time, location, or event, to support a misleading narrative [67]. Input space: Multimodal content with its claim, caption, metadata, or context. Output space: A label indicating whether the content–context pairing is correct or out of context. Success criteria: The prediction should correctly judge whether the content matches the stated context. Evaluation regime: Out-of-context detection is evaluated with accuracy, precision, recall, or macro-F1. Typical methods include fusion-based cross-modal consistency modeling [45,67] and MLLM-based intermediate relation reasoning [73].

3.4. Justification Production

Justification production aims to generate an explicit, evidence-grounded explanation for a predicted verdict label, clarifying why a claim is judged as supported, refuted, or uncertain based on the available evidence [11]. In multimodal fact-checking, justification production should specify evidence from different modalities and explain how cross-modal cues support or contradict the claim [74]. Input space: A claim, the available evidence, and a predicted or gold verdict label. Output space: An explicit natural-language justification linking the verdict to supporting, refuting, or insufficient evidence. Success criteria: A high-quality justification should be complete, coherent, concise, and impartial [75,76]. Evaluation regime: Justification production is evaluated with faithfulness, verdict consistency, completeness, coherence, conciseness, and human judgments of explanation quality [75]. Standard metrics only evaluate the wording and fluency of the generated text. They cannot check whether the explanation accurately cites the underlying image region or video frame, frequently rewarding fluent explanations that hallucinate visual evidence. Major methodological directions include evidence localization with highlighting or structured rationales [77–79], LLM/MLLM-based evidence selection [80], and natural-language justification generation with post-hoc, pre-verdict, or subclaim-level designs [81–83].

Because these four subtasks are conventionally organized in a sequential pipeline, upstream errors cascade across stages and compound final verification failures [11]. From extraction to retrieval, omitting visual context such as timestamps, background landmarks, or text overlays produces decontextualized claims, which misdirect search queries toward superficially similar media. From retrieval to verdict prediction, this evidential starvation leaves models without decisive proof. Lacking abstention mechanisms, models treat deficient evidence as exhaustive and fall back on parametric memory, assigning incorrect labels. From verdict prediction to justification, models conditioned on these false verdicts engage in post-hoc rationalization, hallucinating visual citations and fabricating evidence links to defend the flawed decision [27]. These cascading failures reveal the bottleneck of linear workflows, necessitating the alternative architectures and iterative verification paradigms examined in Section 5.5.

4. Datasets

4.1. Claim Detection and Extraction

The input to the claim detection and extraction system includes original content [84], which may contain text and structured metadata. In multimodal scenarios, the input also includes non-text streams such as images [84], videos [85], or audio [86]. Claim detection is commonly formulated as a classification or ranking problem over candidate units in the input content [84]. Its output is typically a compact label or score, such as a binary checkable label or a check-worthiness score for prioritization [45,87]. The claim extraction task extracts one or more explicit claims from a selected text unit [40], rather than a decision label. Claim extraction typically outputs explicit claims in two formats. One format is an extractive text span that delineates the claim's scope in the source document [88–91]; the other is a rewritten, standalone claim statement [40,42,92,93]. Tables 2 and 3 summarize representative claim detection and extraction datasets.

Table 2. Claim detection datasets. The table summarizes text-based and multimodal benchmarks, including input modalities Img (I), Txt (T), Vid (V), Aud (A), and Meta, available context or evidence, output format as label set size or ranking, dataset scale, language coverage, and source platform. For reproducibility, Avail. indicates whether a stable public dataset link is currently available; Access denotes whether the dataset is directly accessible or requires application/registration, with "–" used when unavailable; and Split indicates whether the official release provides fixed train/dev/test or evaluation partitions, where "–" means that no official fixed split is provided. # denotes a count; #Input denotes the number of input instances, and #Output denotes the number of output labels, with Ranking used for ranking tasks. In the Avail. column, ✓ indicates that a stable public dataset link is available, whereas ✗ indicates that no stable public link was identified.

Dataset	Input	Context	#Output	#Input	Lang	Source	Avail.	Access	Split
CredBank [94]	T	Meta	5	1049	En	Twitter	✓	Direct	–
Weibo [95]	T	Meta	2	5656	En/Zh	Twitter/Weibo	✓	Direct	–
PHEME [96]	T	T/Meta	3	330	En/De	Twitter	✓	Direct	–
RumourEval19 [97]	T	T/Meta	3	446	En	Twitter/Reddit	✓	Direct	Fixed
DAST [98]	T	T/Meta	3	220	Da	Reddit	✓	Direct	–
Suspicious [99]	T	–	2/5	131,584	En	Twitter	✓	Direct	–
CheckThat20-T1 [100]	T	–	Ranking	8812	En/Ar	Twitter	✓	Direct	Fixed
CheckThat21-T1A [101]	T	–	2	17,282	5 langs	Twitter	✓	Direct	Fixed
Debate [39]	T	–	3	1571	En	Transcript	✗	–	–
ClaimRank [102]	T	–	Ranking	5415	En	Transcript	✓	Direct	–
CheckThat18-T1 [103]	T	–	Ranking	16,200	En/Ar	Transcript	✓	Direct	Fixed
CitationReason [104]	T	Meta	13	4000	En	Wikipedia	✓	Direct	–
PolitiTV [105]	T	–	7	6304	En	Transcript	✗	–	–
Bingcheck(raw) [106]	T	T/Meta	2	396	En	BingChat	✓	Direct	–
IndiaWApp [107]	I	Meta	2	2500	4 langs	WhatsApp	✗	–	–
DisinfoMeme [108]	I	–	2	1170	En	Reddit	✗	–	–
TikTok [87]	V	T/Meta	2	891	En	TikTok	✗	–	–
COVID-VTS [109]	V	T/A	2	10,000	En	Twitter	✓	Direct	–
FakeSV [110]	V	T/Meta	2	3654	Zh	TikTok/Kuai	✓	Apply	Fixed
WhatsApp [86]	A	Meta	2	42,689	Pt	WhatsApp	✗	–	–
Podcast-FC [86]	A	–	2	300	En/De	Podcast	✓	Apply	–
Weibo [45]	T/I	Meta	2	9528	Zh	Weibo	✓	Direct	Fixed
FauxBuster [111]	T/I	T/Meta	2	917	En	Twitter/Reddit	✗	–	–
Exfaux [112]	T/I	T	2/4	263	En	Twitter/Reddit	✗	–	–
MuMiN [113]	T/I	Meta	3	12,914	41 langs	Twitter	✓	Direct	Fixed
MMClaims [84]	T/I	–	4	3400	En	Twitter	✓	Direct	Fixed
ContrastFaux [114]	T/I	–	2	1841	En	Twitter/Reddit	✗	–	–
CLEF2023 [43]	T/I	–	4	6000	3 langs	Twitter	✓	Direct	Fixed
MR2 [115]	T/I	T/I/Meta	3	14,700	En/Zh	Twitter/Weibo	✓	Direct	–
CheckMate [116]	T/I/V	Meta	3	2200	Hi	ShareChat	✓	Direct	–
MisDissem [117]	T/I/V/A	Meta	2	121,781	Pt	WhatsApp	✗	–	–

For BingCheck [106], we distinguish two forms of the dataset: BingCheck (raw) corresponds to the original responses used in claim detection, while BingCheck (full) corresponds to the extracted claims used for evidence retrieval. For datasets covering

more than three languages, we report only the number of languages covered in the list; the specific languages included can be found in Appendix B, Table A2.

The main bottleneck of claim detection and extraction datasets is the subjective definition of check-worthiness and claim boundaries. Datasets built from public media may bias models toward highly visible events rather than generalizable check-worthiness signals. Heterogeneous outputs, such as labels, rankings, and rewritten claims, also make datasets under the same task difficult to compare. This problem is harder for implicit or multimodal claims, where claim boundaries may be unstated and relevant information may be distributed across text, images, or videos.

Table 3. Claim extraction datasets. Output denotes the target claim representation. Claim_Span refers to the extractive token-level span of claim-worthy content within the input text. Reformulated_Claim refers to a textual claim reformulated from the input, covering normalized, decontextualized, or atomic formulations. # denotes a count; #Input denotes the number of input instances. In the Avail. column, ✓ indicates that a stable public dataset link is available.

Dataset	Input	Output	#Input	Lang	Source	Avail.	Access	Split
CURT [88]	T	Claim_Span	7500	En	Twitter	✓	Direct	Fixed
X-CLAIM [89]	T	Claim_Span	7040	6 langs	Fact-check sites	✓	Direct	Fixed
HECSI [90]	T	Claim_Span	14,097	En/Hi	Twitter	✓	Direct	Fixed
Multilingual-Claim-Span [91]	T	Claim_Span	4988	En/Hi/Bn	Social media	✓	Apply	–
AVeriTeC-DCE [42]	T/Meta	Reformulated_Claim	1231	En	Web articles	✓	Direct	–
FEVERFact [40]	T	Reformulated_Claim	4400	En	Wikipedia	✓	Direct	Fixed
CLAN [92]	T	Reformulated_Claim	6388	En	Social posts	✓	Apply	–
mCLAN [93]	T	Reformulated_Claim	28,012	20 langs	Fact-check sites	✓	Direct	Fixed

4.2. Evidence Retrieval

The input to the evidence retrieval system is a retrieval query derived from the previous stage. It may include the claim itself as well as auxiliary cues such as decomposed sub-queries [118], entities [119], and contextual metadata [29]. Evidence retrieval is commonly formulated as a ranking problem, where the system returns a top-k list of candidate evidence items [51]. Depending on the setting, these items can be text passages or non-textual evidence such as images, videos, or audio segments. These retrieved candidates serve as supporting context for downstream verification models to assess the claim and produce evidence-grounded explanations. Retrieval is often evaluated using a predefined evidence collection. When the evidence collection is fixed, retrieval performance can be compared fairly because all systems search the same corpus, even in experimental settings with different evidence modalities, data sources, or languages [120]. Table 4 summarizes representative evidence retrieval datasets.

Table 4. Datasets for evidence retrieval in fixed evidence corpora, covering text-only and multimodal settings where retrieval ranks candidates from a predefined corpus. Evi. denotes the evidence modality. # denotes a count; #Input denotes the number of retrieval queries or input instances. In the Avail. column, ✓ indicates that a stable public dataset link is available.

Dataset	Input	Output	Evi.	Retrieval Source	#Input	Lang	Avail.	Access	Split
BEIR [121]	T	Top-k Txt	T	Multiple text corpora	49–13,145	En	✓	Direct	Fixed
MS MARCO [122]	T	Top-k Txt	T	Web	532,761	En	✓	Direct	Fixed
mMARCO [123]	T	Top-k Txt	T	Web	532,761	14 langs	✓	Direct	Fixed
MIRACL [124]	T	Top-k Txt	T	Wikipedia	53,274	18 langs	✓	Direct	Fixed
Mr. TyDi [125]	T	Top-k Txt	T	Wikipedia	69,707	11 langs	✓	Direct	Fixed
DuReader [126]	T	Top-k Txt	T	Web	92,395	Zh	✓	Direct	Fixed
BingCheck(full) [106]	T	Top-k Txt	T	Web	3840	En	✓	Direct	–
WIT [127]	T	Top-k Img	I	Wikipedia	37,600,000	108 langs	✓	Direct	Fixed
TRECVID [128]	T/I	Top-k Vid	V	IACC	30	En	✓	Apply	Fixed
INRIA Holidays [129]	I	Top-k Img	I	–	1491	–	✓	Direct	Fixed
Stanford I2V [130]	I	Top-k Vid	V	TV channels	229	–	✓	Direct	Fixed

Table 4. *Cont.*

Dataset	Input	Output	Evi.	Retrieval Source	# Input	Lang	Avail.	Access	Split
NTU-VOI [131]	I	Top-k Vid	V	YouTube	10	–	✓	Direct	Fixed
FIVR-200K [132]	V	Top-k Vid	V	YouTube	100	–	✓	Direct	Fixed
VCDB [133]	V	Top-k Vid	V	YouTube/MetaCafe	28	–	✓	Direct	Fixed
CC_WEB_VIDEO [134]	V	Top-k Vid	V	YouTube/Google Video	24	–	✓	Direct	Fixed
EVVE [135]	V	Top-k Vid	V	YouTube	620	–	✓	Direct	Fixed

The central limitation of evidence retrieval datasets is the gap between retrieval relevance and evidential sufficiency. A relevant document, image, or video may still be insufficient to support or refute a specific claim. This gap is especially clear in multimodal benchmarks, which often emphasize visual similarity, entity overlap, or video instance retrieval rather than claim-conditioned evidence. Therefore, high retrieval scores may not translate into reliable fact-checking performance. Open-web settings further complicate evaluation because evidence may change, disappear, or become inaccessible.

4.3. Verdict Prediction

The input to verdict prediction typically consists of a claim and a set of candidate evidence items [51], where the evidence may be provided by the dataset or retrieved from external resources. The output of the verdict prediction is a structured decision that assigns a truthfulness label to a claim based on a predefined set of labels [136].

Out-of-context datasets. Out-of-context datasets test a system's ability to detect multimodal content reuse, where the media is genuine but is presented with an incompatible context that makes the overall message misleading. They are constructed from authentic image–text pairs. Negative examples are created by controlled context swapping, for example, by pairing the same real image with different but plausible titles or metadata such as time, place, or event [137–139].

Manipulation datasets. Manipulation datasets evaluate whether a system can detect synthesized or substantially altered multimodal content. Image datasets are built by collecting real images as a base set, creating manipulated versions of a subset through operations such as splicing [140], copy-move forgery [141], inpainting [142], or AI-based editing [143], and labeling real and manipulated instances accordingly. Video datasets generate deepfakes from real video, using strategies such as face swaps [144] and reenactments [145]. Audio datasets are based on real recordings, and fake samples are created through text-to-speech synthesis or speech conversion [146]. Table 5 summarizes representative out-of-context datasets and manipulation datasets.

Table 5. Datasets for out-of-context and manipulation detection. In the Task column, O denotes out-of-context detection, and M denotes manipulation detection. # denotes a count; #Input and #Output denote the numbers of input instances and output labels, respectively. In the Avail. column, ✓ indicates that a stable public dataset link is available, whereas ✗ indicates that no stable public link was identified.

Dataset	Input	Evidence	#Output	Task	#Input	Lang	Source	Avail.	Access	Split
News400 [147]	T/I	I	2	O	400	En/De	News	✓	Direct	–
Ookpik [148]	T/I	–	2	O	545	En	Social media	✗	–	–
InfoSurgeon [149]	T/I	KB/Meta	2	O	30,000	En	VoA	✓	Direct	–
TNews [147]	T/I	I	2	O	72,561	En	News	✓	Direct	–
NeuralNews [150]	T/I	–	4	O	128,000	En	Grover/GoodNews	✓	Direct	Fixed
MEIR [151]	T/I	Meta	2	O	140,096	En	Flickr	✓	Direct	–
COSMOS [137]	T/I	–	2	O	201,700	En	News/Snopes	✓	Direct	Fixed
MAIM [138]	T/I	Meta	2	O	239,968	En	Flickr	✗	–	–
Twitter-COMMs [152]	T/I	Meta	2	O	884,000	En	Twitter	✓	Direct	Fixed
NewsCLIPpings [139]	T/I	–	2	O	988,283	En	CLIP/VisualNews	✓	Direct	Fixed
MMOOC [153]	T/I/V/A	–	2	O	455,000	–	News	✗	–	–

Table 5. *Cont.*

Dataset	Input	Evidence	#Output	Task	#Input	Lang	Source	Avail.	Access	Split
In-the-Wild [154]	I	Meta	2	M	201	–	Reddit/Onion	✓	Direct	–
MediaEval [140]	I	T/Meta	2	M	13,924	–	Twitter	✓	Direct	Fixed
FakingSandy [155]	I	T/Meta	2	M	16,117	–	Twitter	✓	Direct	–
PS-Battles [156]	I	T/Meta	2	M	103,028	-	Reddit	✓	Direct	–
DGM [157]	I	T	2	M	230,000	–	News	✓	Direct	Fixed
SID-Set [158]	I	–	3	M	300,000	–	Social media	✓	Direct	Fixed
GenImage [143]	I	–	2	M	2,681,167	–	ImageNet/Generators	✓	Direct	Fixed
VTD [159]	V	–	2	M	33	En	YouTube	✓	Direct	–
DeepfakeDetect [160]	V	–	2	M	600	En	Vid Webs./HOHA	✗	–	–
FaceForensics [145]	V	-	2	M	1004	En	YouTube	✓	Apply	Fixed
Celeb-DF [161]	V	–	2	M	6229	En	YouTube	✓	Apply	Fixed
DeeperForensics-1.0 [162]	V	–	2	M	60,000	En	Recorded	✓	Apply	Fixed
DFDC [144]	V	–	2	M	128,154	En	Recorded	✓	Direct	Fixed
DF-Platter [163]	V	–	2	M	133,260	En	YouTube	✓	Direct	Fixed
KoDF [164]	V	–	2	M	237,942	Ko	Recorded	✓	Apply	Fixed
ASVspoof [165]	A	–	2	M	16,375	En	SAS	✓	Direct	Fixed
DeepSonar [166]	A	–	2	M	18,614	En/Zh	TTS Systems/VCC	✗	–	–
ADFF 2024 [167]	A	–	2	M	26,976	En	TTS/VC Systems	✗	–	–
Phonespoof [168]	A	-	2	M	34,407	En	ASVspoof	✗	–	–
FoR [146]	A	–	2	M	53,868	En	TTS Systems	✓	Direct	Fixed
HAD [169]	A	–	3	M	88,035	Zh	AISHELL-3	✓	Direct	Fixed
ASVspoof 5 [170]	A	–	2	M	1,004,081	En	Crowdsourced	✓	Apply	Fixed
Deepfake-Eval-2024 [171]	I/V/A	–	2	M	5831	52 langs	In-the-wild	✓	Direct	Fixed
FakeAVCeleb [172]	V/A	–	4	M	20,000	En	VoxCeleb2	✓	Direct	Fixed
AV-Deepfake1M [173]	V/A	–	4	M	1,146,760	En	Recorded/VoxCeleb2	✓	Direct	Fixed

Verdict classification datasets. Verdict classification datasets are used to evaluate whether fact-checking systems can assign the correct verdict label to a claim under a specified evidence access mechanism [174]. Text-only datasets draw claims from sources such as Wikipedia [136], political speeches [175], or fact-checking websites [36], and derive labels from the corresponding references and verdict reports. When evidence is provided, it is curated by recording the specific passages or documents used to support or refute the claim. Image-based datasets collect real-world posts that pair text with images [74,110,176]. Labels are assigned by linking posts to existing fact-check articles when available. Otherwise, annotators verify the claim using reliable sources and label the post accordingly. Evidence is drawn from the post context and associated media, and some benchmarks [74,177,178] additionally annotate external sources that support or refute the claim [21,179]. Video-based datasets gather claims grounded in clips or programs and label them by verification against trusted sources. Evidence may include aligned transcript segments or video snippets, together with external reports used during verification [83,180]. Table 6 summarizes representative verdict classification datasets across modalities.

Table 6. Statistics of unimodal and multimodal fact-checking datasets. "Mod." denotes modality, "Evi." indicates whether evidence is provided, and "Lang" denotes language. # denotes a count; #Claim and #Label denote the numbers of claims and verdict labels, respectively. In the Avail. column, ✓ indicates that a stable public dataset link is available, whereas ✗ indicates that no stable public link was identified.

Dataset	Mod.	#Claim	Source	Evi.	#Label	Lang	Avail.	Access	Split
FEVER [51]	T	185,445	Wikipedia	Y	3	En	✓	Direct	Fixed
LIAR [175]	T	12,836	PolitiFact	Y	6	En	✓	Direct	Fixed
FEVEROUS [181]	T	87,026	Wikipedia	Y	3	En	✓	Direct	Fixed
VitaminC [182]	T	163,180	Wikipedia	Y	3	En	✓	Direct	Fixed
HoVer [183]	T	26,171	Wikipedia	Y	2	En	✓	Direct	Fixed
AVeriTeC [184]	T	4568	Fact-check sites	Y	4	En	✓	Direct	Fixed
X-FACT [185]	T	31,189	Fact-check sites	Y	5	25 langs	✓	Direct	Fixed
LIAR2 [186]	T	22,962	PolitiFact	Y	6	En	✓	Direct	–

Table 6. *Cont.*

Dataset	Mod.	#Claim	Source	Evi.	#Label	Lang	Avail.	Access	Split
CFEVER [187]	T	30,012	Wikipedia	Y	3	Zh	✓	Direct	Fixed
CREAK [188]	T	12,918	MTurk	N	2	En	✓	Direct	Fixed
FaVIQ [189]	T	188,376	NQ/AmbigQA	Y	2	En	✓	Direct	Fixed
DanFEVER [190]	T	6407	Wikipedia	Y	3	Da	✓	Direct	Fixed
CoVERT [191]	T	300	Tweet	Y	3	En	✓	Direct	–
Stanceosaurus [192]	T	28,033	Fact-check sites	Y	3/5	En	✓	Direct	Fixed
SciFact [193]	T	1409	Journal	Y	2	En	✓	Direct	Fixed
TabFact [194]	T	118,275	Human	Y	2	En	✓	Direct	Fixed
WatClaimCheck [195]	T	33,721	Fact-check sites	Y	3/86	En	✓	Direct	Fixed
WikiFactCheck [196]	T	124,821	Wikipedia	Y	3	En	✓	Direct	–
CHEF [197]	T	10,000	Fact-check sites	Y	3	Zh	✓	Direct	Fixed
ChatGPT-FC [198]	T	22,337	PolitiFact	N	6	En	✗	–	–
MultiFC [36]	T	36,535	Fact-check sites	Y	2/40	En	✓	Direct	Fixed
FakeNewsNet [199]	T	1,423,073	Open channels	N	2	En	✓	Direct	–
MultiClaim [200]	T	28,092	Fact-check sites	Y	2	39 langs	✓	Direct	Fixed
BingCheck [106]	T	3840	Web Search	Y	3	En	✓	Direct	–
MOCHEG [74]	T/I	16,607	PolitiFact/Snopes	Y	3	En	✓	Direct	Fixed
VERITE [176]	T/I	1000	Snopes/Reuters	N	3	En	✓	Direct	–
Factify [177]	T/I	50,000	X	Y	3	En	✓	Direct	Fixed
Factify 2 [178]	T/I	50,000	X/PolitiFact/Snopes	Y	3	En	✓	Direct	Fixed
MM-Claims [84]	T/I	3400	X	Y	2/3	En	✓	Direct	Fixed
STVD-FC [201]	T/V	1200	TV programs	N	–	Fr	✓	Direct	–
FakeSV [110]	T/V	3654	Douyin/Kuaishou	Y	2	Zh	✓	Apply	Fixed
FakeTT [180]	T/V	1991	Snopes	Y	2	En	✓	Direct	Fixed
FakeVV [202]	T/V	102,000	News Accounts	Y	2	Zh	✓	Direct	Fixed
MAD [203]	T/A	4915	PolitiFact	N	2	En	✓	Direct	Fixed
MAD2 [204]	T/A	3368	PolitiFact	N	2	En	✓	Direct	Fixed
CheckMate [116]	T/I/V	2200	ShareChat	Y	3	Hi	✓	Direct	–
TRUE [83]	T/I/V	2925	Snopes	Y	2/8	En	✓	Direct	Fixed

Beyond standard dataset statistics such as modality, scale, evidence availability, and label cardinality, we additionally report the SOTA for each benchmark together with its main evaluation scores, including accuracy (Acc) and macro-F1 (F1), as summarized in Table 7.

Table 7. SOTA performance on representative fact-checking datasets. Results are from different papers and are not directly comparable due to differences in splits, evidence settings, and metrics. Mod. (Modality): T = text, I = image, V = video, and A = audio. Input: Clm = claim, Evi = evidence, Web_Evi = web-retrieved evidence, and Meta = metadata. Metric: Acc = accuracy, Macro-F1 = macro-averaged F1, FO_S = FEVEROUS score, F_S = FEVER score, AC_S = AVeriTeC score, and N/R = not reported. Split: real/syn_Test = real/synthetic test sets, 3/5 Labs = 3-class/5-class evaluation subsets. Year is the publication year of the reported SOTA method.

Dataset	Mod.	SOTA	Split	Input	Metric		Year
					Macro-F1/Joint	Acc	
FEVER [51]	T	CorXFact [205]	Test	Clm/Evi	93.5	93.6	2025
LIAR [175]	T	Aslam et al. [206]	Test	Clm/Meta	91.4	89.8	2021
FEVEROUS [181]	T	Varnosfaderani et al. [207]	Test	Clm/Evi	32.8 (FO_S)	62.0	2024
VitaminC [182]	T	VitaminC [182]	real/syn_Test	Clm/Evi	N/R	88.9/94.0	2021
HoVer [183]	T	MRR-FV [208]	Dev	Clm	71.7	N/R	2025
AVeriTeC [184]	T	HerO [209]	Test	Clm/Meta	57.0 (AC_S)	75.2	2024
X-FACT [185]	T	Singhal et al. [210]	Mixed Set	Clm	N/R	59.0	2024
LIAR2 [186]	T	FDHN [186]	Test	Clm/Meta	69.6	70.2	2024
CFEVER [187]	T	Lin et al. [187]	Test	Clm/Evi	64.8 (F_S)	69.7	2024
CREAK [188]	T	Onoe et al. [188]	Test	Clm	N/R	85.1	2021
FaVIQ [189]	T	QaDialMoE [211]	FaVIQ-R Test	Clm/Evi	N/R	86.0	2022
DanFEVER [190]	T	Nørregaard and Derczynski [190]	Test	Clm/Evi	90.2	N/R	2021
CoVERT [191]	T	CliVER [212]	Test	Clm/Evi	92.0	92.5	2024

Table 7. *Cont.*

Dataset	Mod.	SOTA	Split	Input	Metric		Year
					Macro-F1/Joint	Acc	
Stanceosaurus [192]	T	Zheng et al. [192]	Test(3/5 Labs)	Clm/Meta	66.8/61.0	N/R	2022
SciFact [193]	T	Kosprdic et al. [213]	Test	Clm/Evi	88.0	89.0	2024
TabFact [194]	T	ARTEMIS-DA [214]	Test	Clm/Meta	N/R	93.1	2025
WikiFactCheck [196]	T	Sathe et al. [196]	Test	Clm/Evi	69.5	68.0	2020
CHEF [197]	T	Hu et al. [197]	Test	Clm/Web_Evi	62.3	65.3	2022
MultiFC [36]	T	DRIVE [215]	Test	Clm/Meta	63.2	59.7	2025
FakeNewsNet [199]	T	Al-obaidi and Çağlıkantar [216]	Test	Clm/Meta	99.9	99.9	2024
BingCheck [106]	T	Self-Checker [106]	Full Set	Clm/Web_Evi	N/R	63.4	2024
MOCHEG [74]	T/I	DEFAME [55]	Test	Clm/Web_Evi	53.8	59.5	2024
VERITE [176]	T/I	DEFAME [55]	Full Set	Clm/Web_Evi	N/R	84.5	2024
Factify [177]	T/I	Dhankar et al. [217]	Test	Clm/Evi	74.8	N/R	2022
Factify 2 [178]	T/I	Suryavardan et al. [218]	Test	Clm/Evi	81.8	N/R	2023
MM-Claims [84]	T/I	Cheema et al. [84]	Evaluation Set	Clm/Evi	81.1	81.4	2022
FakeSV [110]	T/V	Fact-R1 [202]	Test	Clm/Meta	74.7	75.6	2025
FakeTT [180]	T/V	Fact-R1 [202]	Test	Clm/Meta	72.7	74.4	2025
FakeVV [202]	T/V	Fact-R1 [202]	Test	Clm/Meta	80.3	81.2	2025
MAD [203]	T/A	Chun et al. [203]	Test	Clm/Meta	82.7	87.0	2025
MAD2 [204]	T/A	Chun et al. [204]	Fusion Test	Clm/Meta	76.4	85.2	2026
TRUE [83]	T/I/V	Niu et al. [83]	15% Test	Clm/Web_Evi	79.6	78.9	2025

FakeNewsNet is a closed-domain news verdict classification benchmark. Models are evaluated on article text and social context features, such as user engagement, without open-domain evidence retrieval. This makes the task substantially easier than FEVER-style fact-checking, explaining the near-perfect accuracy. Joint verification and retrieval scores are reported in the Macro-F1/Joint column when standard Macro-F1 is unavailable, where FO_S denotes the FEVEROUS score, F_S denotes the FEVER score, and AC_S denotes the AVeriTeC score. Note: All performance metrics in this table are compiled from the respective original published papers cited in the SOTA column. Because benchmarks differ fundamentally in evaluation metrics, data partitions, evidence frameworks, input modalities, and classification labels, these scores serve strictly as dataset-specific historical reference points rather than a standardized cross-benchmark leaderboard.

Several quantitative insights emerge from this compilation. First, the evidence acquisition strategy remains a dominant factor. Claim-only settings range from 59.0 to 85.1 in accuracy, while dataset-provided evidence settings generally reach 62.0 to 94.0. Web-evidence settings are lower and more variable (59.5–84.5), reflecting the instability of retrieved web evidence. Second, modality matters. Text-only datasets (e.g., FEVER, LIAR) tend to have higher reported scores than multimodal datasets (e.g., MOCHEG, VERITE), partly because multimodal tasks require joint reasoning across heterogeneous signals. Finally, year-over-year improvement is visible. Datasets introduced since 2023 (e.g., AVeriTeC) show notable gains over earlier ones, driven by advances in LLM-based reasoning and better evidence annotation. Although results are not directly comparable across datasets, these trends suggest that curated evidence reduces uncertainty, multimodal tasks remain challenging, and recent progress is unevenly distributed across pipeline components.

The biggest challenge in verdict prediction datasets is cross-benchmark comparability, as label schemes, evidence settings, access conditions, and modality coverage vary substantially [32]. Fine-grained labels may distinguish exaggeration, partial truth, missing context, or unverifiable claims, whereas coarser labels merge these cases. Claim-only settings may reflect prior knowledge or artifacts, evidence-based settings test evidence-conditioned reasoning, and web-retrieval settings jointly test retrieval and reasoning. Multimodal verdict datasets often lack diagnostics for whether LLMs actually use visual or audio evidence [20]. Benchmark leakage also deserves attention because public claims and fact-checking articles may have been exposed during LLM pretraining or instruction tuning [219].

4.4. Justification Production

Justification production explains why a system assigns the verdict to a claim by grounding the explanation in the available evidence. The input to this stage minimally includes the claim and the evidence items selected from the retrieval stage or provided by

the benchmark [27]. In some pipeline designs, the predicted verdict label is also used as an input to generate the justification consistent with the final decision [27]. In multimodal settings, the input can also contain any non-textual stream information that can be used as evidence [11]. The output of justification production is typically defined in two main forms. The first form is a natural language justification generated by a model [74]. This textual explanation is the most directly interpretable to readers because it states how the evidence supports or contradicts the claim, or why the evidence is insufficient to reach a definitive conclusion. When the evidence includes visual or audio content, the justification usually incorporates a description and analysis of salient cues, clarifying how the multimodal content contributes to the reasoning and how it interacts with the textual evidence. The second form is structured or highlighted output, which indicates which parts of the input are crucial without providing a full text explanation [220]. Typical outputs include highlighting or emphasizing specific tokens or spans in the claim and evidence, and representing key facts as subject–verb–object triples that capture the proposition underlying the conclusion. In multimodal settings, some systems further localize relevant regions in images [221] or identify informative segments in video [222] or audio [223], helping users directly inspect the portions of the evidence that support the system's decision. Table 8 summarizes representative justification datasets.

Table 8. Datasets for justification production in automated fact-checking. Input abbreviations: Clm denotes claim, Evi denotes evidence, and Lab denotes label. Output abbreviations: NL denotes natural language, HL denotes highlight, and Tri. denotes triples. Just. denotes whether the gold justification is manually written or model-generated. # denotes a count; #Input denotes the number of input instances. In the Avail. column, ✓ indicates that a stable public dataset link is available, whereas ✗ indicates that no stable public link was identified.

Dataset	Mod.	Input	Output	#Input	Lang	Source	Just.	Avail.	Access	Split
FEVER [51]	T	Clm/Evi	HL	185,445	EN	Wikipedia	Manually	✓	Direct	Fixed
HoVer [183]	T	Clm/Evi	HL	26,171	EN	Wikipedia	Manually	✓	Direct	Fixed
SciFact [193]	T	Clm/Evi	HL	1409	EN	Scientific papers	Manually	✓	Direct	Fixed
ExClaim [224]	T	Clm/Evi	NL	6951	EN	Fact-checking Websites	Manually	✓	Direct	–
FactEx [2]	T	Clm/Evi	NL	12,150	EN	FullFact website	Manually	✗	–	–
LIAR-PLUS [225]	T	Clm/Lab	NL	12,836	EN	PolitiFact	Manually	✓	Direct	Fixed
LIAR++ [226]	T	Clm/Lab	NL	6451	EN	PolitiFact	Manually	✓	Direct	Fixed
FullFact [226]	T	Clm/Evi	NL	1838	EN	FullFact website	Manually	✓	Direct	Fixed
PUBHEALTH [227]	T	Clm/Lab	NL	11,832	EN	Fact-checking Websites	Manually	✓	Direct	Fixed
AVeriTeC [184]	T	Clm/Evi/Lab	NL	4568	EN	ClaimReview	Manually	✓	Direct	Fixed
EX-FEVER [228]	T	Clm/Lab	NL	61,265	EN	Wikipedia	Manually	✓	Direct	Fixed
e-FEVER [229]	T	Clm/Evi/Lab	NL	67,687	EN	Wikipedia	Generate	✗	–	–
FlawCheck [230]	T	Clm/Evi	NL	33,721	EN	Fact-checking Websites	Generate	✓	Direct	Fixed
FEVEROUS [181]	T	Clm/Evi	HL	87,026	EN	Wikipedia	Manually	✓	Direct	Fixed
ChartCheck [231]	T/I	Clm/Evi/Lab	NL	10,500	EN	Wikimedia Commons	Manually	✓	Direct	–
MOCHEG [74]	T/I	Clm/Evi/Lab	NL	15,601	EN	PolitiFact/Snopes	Manually	✓	Direct	Fixed
RW-Post [232]	T/I	Clm/Evi	Tri.	1771	EN	Snopes	Generate	✓	Direct	–
TRUE (ECR) [83]	T/V/A	Clm/Evi	NL	2925	EN	Fact-checking Websites	Manually	✓	Direct	Fixed
TRUE (LSR) [83]	T/V/A	Clm/Evi	NL	2925	EN	Fact-checking Websites	Generate	✓	Direct	Fixed
MMM-Fact [233]	T/I/V	Clm/Evi	NL	125,449	EN	PolitiFact/Snopes	Manually	✓	Direct	–

The main pain point of justification datasets is the trade-off between annotation quality and dataset scale. Manual justifications are more faithful but costly, limiting data quantity, language diversity, and modality coverage [27]. Generated justifications increase coverage but may inherit model biases or hallucinations. Datasets also define justification differently, ranging from evidence snippets and highlighted spans to rationales and abstractive explanations, making results difficult to compare. In multimodal settings, high-quality justifications further require grounding in image regions, optical character recognition (OCR) text, video frames, audio segments, or cross-modal relations, which most datasets still lack.

4.5. Evaluation Metrics

Existing fact-checking benchmarks commonly report classification metrics, such as accuracy and macro-F1, together with dataset-specific scores, including FEVER score, FEVEROUS score, and AVeriTeC score. Accuracy measures the proportion of instances whose predicted labels match the gold labels:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i), \tag{1}$$

where N is the number of test instances, y_i is the gold label, $\hat{y}_i$ is the predicted label, and $\mathbb{I}(\hat{y}_i = y_i)$ equals 1 if the prediction is correct and 0 otherwise. For each class c, precision, recall, and F1 are computed as

$$P_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \quad R_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}, \quad \text{F1}_c = \frac{2P_cR_c}{P_c + R_c}, \tag{2}$$

where TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives for class c, respectively. Macro-F1 gives equal weight to each class by averaging the class-wise F1 scores:

$$\text{Macro-F1} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\text{F1}_c, \tag{3}$$

where $\mathcal{C}$ denotes the set of verdict labels.

Evidence-grounded benchmarks further require systems to provide supporting evidence, and therefore use joint metrics that evaluate both verdict correctness and evidence adequacy. In FEVER [51], the FEVER score is computed as

$$\text{FEVER_score} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i \wedge \text{EvOK}_i), \tag{4}$$

where $\hat{y}_i$ and y_i are the predicted and gold verdict labels, and EvOK_i indicates whether the retrieved evidence satisfies the gold evidence requirement. For supported and refuted claims, $\text{EvOK}_i = 1$ only if at least one complete gold evidence set is retrieved; for claims labeled as *Not Enough Information*, no gold evidence is required. Thus, the FEVER score measures joint label accuracy and evidence sufficiency. FEVEROUS [181] extends the FEVER-style metric to heterogeneous evidence, including textual sentences and tabular elements. Its score is computed similarly:

$$\text{FEVEROUS_score} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i \wedge \text{HetEvOK}_i), \tag{5}$$

where HetEvOK_i indicates whether the retrieved textual and/or tabular evidence covers at least one complete gold evidence set. Therefore, the FEVEROUS score measures whether a system can make the correct verdict while retrieving sufficient evidence from both unstructured and structured sources. AVeriTeC [184] evaluates real-world claim verification with evidence represented as question–answer pairs. The AVeriTeC score can be written as

$$\text{AVeriTeC_score} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i \wedge \text{EvScore}_i \geq \tau), \tag{6}$$

where EvScore_i is the official evidence score measuring the quality of the generated question–answer evidence against the reference evidence, and τ is the official evidence-quality threshold. The metric therefore measures verdict accuracy conditioned on sufficient evidence quality.

Although these metrics are widely used, they are insufficient for realistic fact-checking evaluation. First, false positives and false negatives have asymmetric costs, and their severity depends on the application context [234]. False-positive errors may incorrectly flag legitimate content, whereas false-negative errors may allow harmful misinformation to spread in high-risk domains such as public health, elections, and crisis response. Thus, single-point accuracy and macro-F1 can obscure precision–recall trade-offs and should be complemented with precision–recall curves [235], operating-point analysis [236], cost-sensitive metrics [234], and calibration measures [237].

Second, label-level metrics do not assess whether generated justifications are complete, coherent, concise, impartial, or faithful to the evidence. A model may predict the correct verdict while relying on incomplete, irrelevant, or unfaithful reasoning. Therefore, evaluation should include ROSCOE-style reasoning metrics [238], FActScore-style atomicity checks [239], faithfulness protocols [240], and hallucination metrics [241]. Human-centered evaluation rubrics [226] remain important for assessing explanation quality, evidential adequacy, and whether generated justifications are understandable and properly grounded in evidence.

Third, benchmark contamination is a concern for LLM-based fact-checking because public datasets such as FEVER [51], LIAR [175], and MultiFC [36] may overlap with pretraining corpora [242]. High performance on static benchmarks may therefore reflect memorization rather than genuine verification. Robust protocols should include temporal splits [243], contamination audits [242], and evidence sources unlikely to have appeared during pretraining [244,245].

Fourth, conventional metrics are inadequate for evaluating multimodal verification, severely limiting cross-modal comparability. Standard accuracy and macro-F1 cannot distinguish genuine cross-modal reasoning from reliance on a single modality, allowing models to score high by exploiting textual cues while completely ignoring accompanying visual or acoustic signals. As a result, high benchmark scores may reflect unimodal bias rather than true verification capability, making comparisons between text-only and multimodal settings misleading [20,176]. Furthermore, joint metrics such as the FEVER score and FEVEROUS score evaluate evidence sufficiency only for text sentences and table cells, providing no criteria to assess whether visual regions, video clips, or audio streams adequately support the claim [51,181]. Consequently, the field lacks standardized evaluation protocols that jointly measure multimodal evidence adequacy and verdict correctness.

Reported benchmark performance should be interpreted with caution, as interpreting performance differences requires distinguishing genuine methodological improvements from differences caused by experimental conditions [32]. Specifically, differences in experimental conditions primarily involve the following factors. First, evidence access substantially influences outcomes, as models evaluated on curated gold evidence typically achieve higher scores than those navigating open-web retrieval [51,74,182,197]. Second, model configurations also confound comparisons, where score gains often arise from scaling backbone models rather than algorithmic advances [246]. Third, variations in label granularity and modality settings can complicate cross-benchmark comparisons. High accuracy on binary tasks does not necessarily translate to fine-grained classifications [36,175], while multimodal gains may sometimes stem from unimodal verification rather than genuine cross-modal reasoning [20,176]. In addition, potential benchmark exposure during pretraining can lead to score inflation through memorization rather than actual verification [219,242]. In contrast, genuine methodological improvements stem from algorithmic advances, such as adaptive query reformulation, cross-modal alignment, and iterative evidence cross-examination. These methods yield consistent gains under controlled baselines

with identical settings. Therefore, performance differences reflect methodological progress only when experimental conditions are strictly matched.

5. Methods

5.1. Claim Detection and Extraction

In this section, we discuss representative modeling paradigms for claim detection and extraction, including unimodal claim processing, early multimodal fusion, Transformer-based multimodal fusion, and LLM-based multimodal approaches. Figure 2 illustrates the structure of this section.

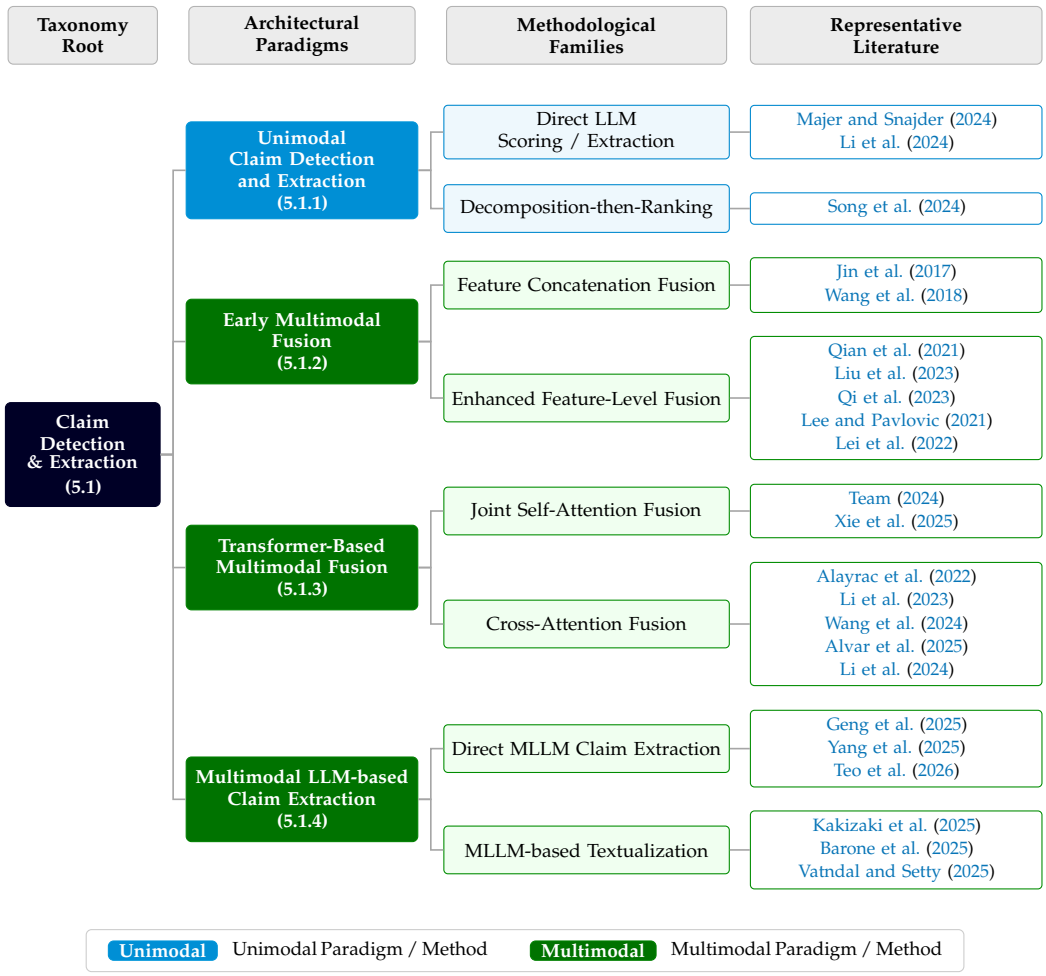


Figure 2. Taxonomy of claim detection and extraction methods. Blue nodes denote unimodal methods, whereas green nodes denote multimodal methods. The figure groups methods by how claims are obtained from text-only or multimodal inputs: LLM scoring or decomposition, feature-level fusion, Transformer-based attention fusion, and MLLM-based extraction or textualization. Top header boxes delineate the taxonomic progression across architectural paradigms, methodological families, and the representative literature.

5.1.1. Unimodal Claim Detection and Extraction

Traditional methods primarily model this stage as a classification problem, determining whether the original content contains a claim and whether the claim is checkable or check-worthy [43]. Then, in a separate step, they extract text segments related to the claim [42]. However, this fixed pipeline relies on task-specific classifiers and extractors, making it less flexible for implicit or context-dependent claims. In contrast, unimodal approaches use LLMs as more adaptive claim-processing modules, with their specific role varying across different subtasks. For check-worthiness, LLMs serve as claim selectors,

performing binary classification or producing ranking scores for prioritization. A common approach is prompt-based check-worthiness estimation. The worthiness criteria are distilled from annotation guidelines and encoded in zero-shot or few-shot prompts, and the LLM's confidence scores are used to rank candidate claims for downstream verification [247]. Complementary work fine-tunes open-source LLMs for supervised check-worthiness estimation. FactFinders [248] is one example, combining fine-tuning with training-data pruning to reduce cost while keeping performance competitive. In contrast, VERISCORE [44] does not predict check-worthiness scores. It uses an LLM for claim extraction and decomposition, converting long text into atomic, verifiable propositions for downstream verification. This is implemented with a sliding-window context, which helps generate standalone claims and excludes non-verifiable content such as advice, fiction, or subjective opinions.

Unimodal methods are efficient and controllable for explicit claims, making them suitable for low-cost filtering [42,43]. But they struggle with implicit, long, or context-dependent claims. LLM-based methods are more flexible, yet they are costly, prompt-sensitive, and may produce unfaithful outputs [247]. A key gap is cost-aware processing with stronger faithfulness checking and uncertainty estimation.

5.1.2. Early Multimodal Fusion

To extend claim detection and extraction from text inputs to real-world posts that also include images and videos, subsequent work investigates how to fuse information across modalities. Early multimodal approaches relied on feature-level modeling. Features from different modalities are extracted independently and then fused for claim detection. The most common strategy is to directly concatenate visual and textual features [45,46]. Later work adopted stronger fusion modules, including attention-based fusion [249–251], shared latent representation learning [252], and graph-based interaction modeling [253], to better capture cross-modal dependencies. These methods primarily integrate images and text to make a binary decision, often failing to preserve the fine-grained cross-modal details necessary to compose a complete claim [254–256], and the resulting claims may omit key information.

Direct concatenation is simple and efficient, but it mainly captures coarse modality co-occurrence. Attention-based fusion improves cross-modal alignment, though it increases training and computation costs and may still rely on shallow correlations [254]. Future work should move toward finer visual–textual relations and verifiable claim components.

5.1.3. Transformer-Based Multimodal Fusion

Since 2024, Transformer-based vision–language modeling has shifted multimodal fusion from pooled feature combinations to token-level interactions within the Transformer [257]. This shift lays the foundation for MLLM-based systems that support both claim detection and claim extraction through explicit token-level cross-modal grounding. Transformer-based multimodal fusion is commonly realized in two ways: joint self-attention fusion and cross-attention fusion. Concretely, a vision encoder converts an image or sampled video frames into a sequence of visual tokens, and a lightweight projection or adapter maps these tokens into the embedding space of the LLM. In joint self-attention fusion, the model constructs a unified token sequence by combining visual tokens with text tokens as the Transformer input. Multi-head self-attention is then applied over the entire sequence, allowing each visual token and each text token to attend to all other tokens. This fully connected attention pattern supports fine-grained alignment between linguistic units and localized visual information, and helps preserve token-level image–text correspondences throughout multimodal representation learning [47,258]. In cross-attention fusion, the model maintains separate token sets for vision and language. The vision encoder produces

a sequence of visual tokens that is separate from the text tokens. The LLM updates text representations through self-attention over the text sequence, and incorporates visual information through cross-attention. In this cross-attention, text tokens act as queries, while visual tokens provide keys and values. This design injects relevant visual information into the language representations. Because cross-modal interaction is limited to a dedicated cross-attention layer, the LLM can integrate high-dimensional visual inputs with greater computational efficiency. Meanwhile, cross-attention fusion preserves access to fine-grained visual cues, thereby enhancing the fidelity and informativeness of multimodal representations for downstream tasks [48,259]. To make these strategies practical on high-resolution images or long videos, systems often use frame sampling and visual token compression or pruning [260,261]. For posts containing multiple visuals, models can encode multiple images or sampled video frames within a single context together with the post text. This setting allows the fusion mechanism to aggregate complementary visual cues across inputs. The model then produces outputs for claim detection or claim extraction [262].

Joint self-attention supports rich image–text interaction [47], but becomes expensive as visual tokens grow. Cross-attention is more scalable for high-resolution images or long videos [48], but may miss subtle dependencies. A key challenge is distinguishing claim-relevant cues from background noise and spurious correlations [257].

5.1.4. LLM-Based Multimodal Claim Detection and Extraction

Multimodal claim detection can be formulated as check-worthiness prediction over image–text claim pairs. HINTSOFTRUTH [263] introduces a dataset of real-world and synthetic image and claim pairs. It defines multimodal check-worthiness with standard criteria and image-conditioned questions. These questions assess whether the image supports the claim and whether it adds information beyond the text. An MLLM is then prompted with the image and the accompanying text to directly predict the check-worthiness label. When claims are embedded in images, videos, or audio, a key bottleneck is claim formulation, namely turning heterogeneous signals into a set of explicit, verifiable propositions [29]. Earlier pipelines typically convert each modality into text using OCR [116,264], automatic speech recognition (ASR) [265], or captioning [87], and then apply text-only claim spotting. However, these conversion components are error-prone in real-world conditions [266] and may produce incorrect or incomplete descriptions. They may omit or distort key details such as entities, numbers, and qualifiers, which directly constrain what the downstream claim extractor can recover. Works on LLM/MLLM-based multimodal claim extraction mainly follow two setups. First, one line of work prompts an MLLM to read a multimodal social media post and directly generate the main claim as a standalone sentence. M4FC [41] and RAMA [267] formalize this as visual claim extraction, where the model generates a claim from a screenshot, so the target is an explicit proposition. Some approaches further evaluate whether the generated claim is fully faithful to the original multimodal input and reject those that are not [268]. Second, textualizing multimodal inputs into structured textual representations and then prompting an LLM to synthesize that representation into one or more explicit claim statements. MAFT [269] follows this paradigm and directly relies on an MLLM such as GPT-4 to both textualize multimodal inputs and extract claim text that is ready for downstream retrieval and checking. ShortCheck [270] further applies this textualization strategy to short-form videos, extracting speech transcripts, OCR text, and video captions before feeding them into an LLM for claim extraction and check-worthiness detection. In contrast, CER [271] extracts the main textual content from web pages by parsing HTML, and converts videos into text by transcribing the audio stream with an ASR system. It then prompts an LLM to select or extract check-worthy claim candidates from the resulting text.

Direct MLLM extraction works well when claims depend on image–text or video–text relations [29], but it is costly and can be weakly grounded. Textualization-based extraction is easier to integrate with text pipelines, but errors in OCR, ASR, or captioning can hurt downstream performance [264]. The main gap is validating cross-media grounding for subtle or implicit claims.

5.1.5. Trends and Evolution of Claim Detection and Extraction

This section contrasts several paradigms for claim detection and extraction. Unimodal methods are efficient and controllable, but they mainly handle explicit textual claims and remain weak for implicit, long-context, or context-dependent cases [43]. Early multimodal fusion extends coverage to image–text inputs, yet often captures only coarse cross-modal correlation. Transformer-based fusion improves token-level alignment, but aligned cues are not necessarily claim-relevant or checkable. In comparison, MLLM direct reasoning is more flexible for multimodal claim formulation, but it is more expensive, prompt-sensitive, and may produce weakly grounded outputs. LLM-based rewriting and decomposition further convert content into atomic, standalone, and verifiable propositions [44], but they may introduce semantic drift. Taken together, these methods indicate an evolution from efficient claim spotting to more flexible claim formulation, but with increased difficulty in grounding and verification. The future direction is to retain this flexibility while improving faithfulness, traceability, and meaning preservation.

5.2. Evidence Retrieval

Evidence retrieval in fact-checking generally follows two strategies. One relies solely on the claim and its context [222,272], which is lightweight but has limited ability to assess real-world factuality [154]. The other incorporates external evidence such as web search [147], external corpora [51,61,193], fact-checking articles [4,273], knowledge bases [149,274], or reverse image search for visual claims [147,274,275]. Traditional retrieval pipelines rely on manually designed query flows and fixed cross-modal transformations [274,276], which can miss relevant evidence or introduce irrelevant information [277]. Following a typical retrieval pipeline, this section introduces existing methods according to the main stages of the retrieval process, including query construction, cross-modal retrieval, RAG, and evidence consolidation. Figure 3 illustrates the structure of this section.

5.2.1. Query Construction

This motivates LLM-based approaches that make query construction adaptive via claim decomposition and query rewriting. In LLM-based evidence retrieval, query formulation can be viewed as four increasingly complex strategies: (1) Claim-as-query uses the original claim as the retrieval query, serving as a low-cost baseline in early fact-checking pipelines [51]. (2) Subclaim-as-query uses an LLM to decompose a complex claim into atomic subclaims used as retrieval queries. The LLM decontextualizes each subclaim by completing omitted subjects, replacing pronouns with explicit entities, and adding the minimal temporal or locative context needed for the subclaim to be interpretable on its own [52]. This approach typically improves retrieval coverage when a single claim bundles multiple verifiable propositions, but it can introduce decomposition noise and discard cross-sentence context [53]. (3) Claim rewriting applies an LLM to reformulate the claim into a more effective search query by extracting key entities and relations, adding missing constraints such as time or location, and removing non-informative wording [54,55]. This rewriting can also be combined with claim decomposition, where the claim is first split into atomic subclaims, and each subclaim is then rewritten as a query. Such retrieve-and-rewrite loops iteratively refine queries to target missing evidence, improving recall and precision for long or multi-hop claims [278–281].

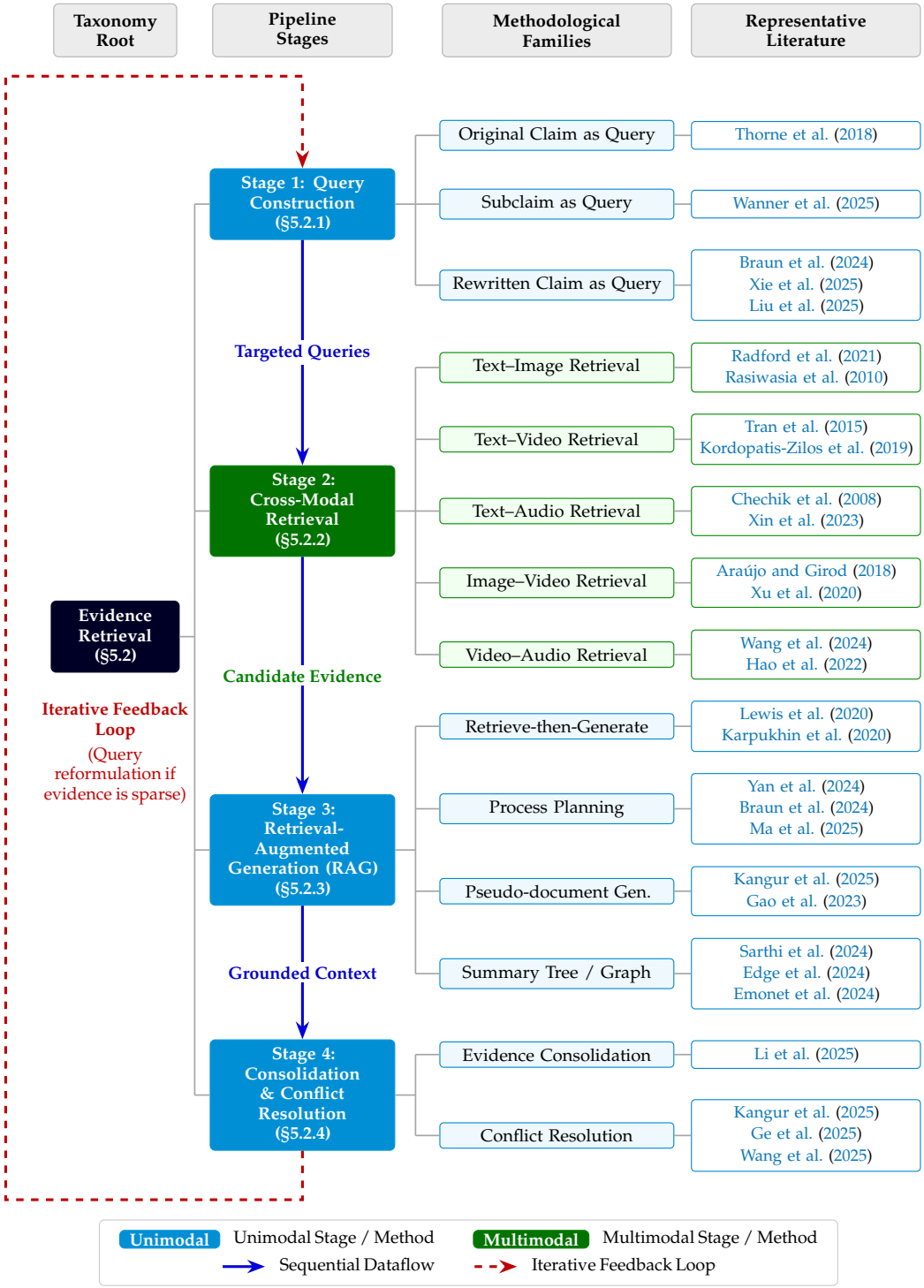


Figure 3. Taxonomy of evidence retrieval methods. Blue nodes denote unimodal methods, whereas green nodes denote multimodal methods. The figure organizes retrieval methods by pipeline stage: query construction, cross-modal retrieval, retrieval-augmented generation, and evidence consolidation with conflict resolution. It further lists the corresponding method families, including query formulation, cross-modal matching, RAG, evidence consolidation, and conflict resolution. Top header boxes denote the taxonomic progression, blue arrows indicate the sequential dataflow between stages, and the dashed red arrow traces the iterative feedback loop for query reformulation.

Query construction balances efficiency, coverage, and faithfulness. Claim-as-query is the cheapest and most transparent, but it is brittle for complex claims [51]. Subclaim decomposition improves coverage, and rewriting improves searchability, yet both add cost

and may introduce errors or drift [52–55]. Combining the two is most adaptive for long or multi-hop claims, but also the most expensive and error-prone [278–280].

5.2.2. Cross-Modal Retrieval

This subsection reviews general cross-modal retrieval methods as background. Although many of these methods were not developed specifically for fact-checking, they form the retrieval backbone of multimodal fact-checking systems that seek evidence across text, images, videos, and audio. For text–image retrieval, the general approach is to learn a shared embedding space. Text is embedded by a sentence encoder, and the image is embedded by a visual encoder. The similarity between the embeddings in the same embedding space is then calculated [56,282]. For text–video retrieval, the standard approach is to represent a video as a sequence of frame/clip embeddings. Video frames/clips are encoded by feeding them into a video encoder such as a 3D CNN [283] to produce a fixed-dimensional embedding [283]. Relevance is computed by matching the query to all segments and aggregating their similarities, emphasizing the best-matching moments rather than a global average. Thus, a few relevant segments can determine the final score even if most of the video is irrelevant [283,284]. For text–audio retrieval, a common approach is window-level audio embedding. Audio is divided into short windows, each converted into a time–frequency representation and encoded by an acoustic model to produce a fixed-dimensional embedding [285]. The query is matched with each window, and the resulting similarities are aggregated into a clip-level score [285,286]. For image–video retrieval, a typical method is to match the query image with video segment embeddings. Each video is first divided into shots or short clips, which are encoded and indexed. Given a query image embedding, the system computes similarities with all segment embeddings. The video is then scored by the maximum similarity or by aggregating its top segment matches, and videos are ranked accordingly [287,288]. For video–audio retrieval, the common approach is segment-level cross-modal matching between video clips and audio windows. The video is split into short clips, and the audio into aligned windows, and each unit is encoded into features. It scores all segment–window pairs and ranks by the top matching moments [289,290].

Cross-modal retrieval broadens evidence coverage beyond text, but increases encoding, indexing, and latency costs [56]. Embedding similarity alone cannot guarantee that matched images or videos truly support the claim [283]. When retrieval is weak, systems should reformulate queries, search more precisely, or report uncertainty instead of forcing a verdict.

5.2.3. Retrieval-Augmented Generation

In LLM/MLLM-based evidence retrieval, query construction and retrieval are often framed as RAG, where a retriever fetches relevant passages from an external corpus and the generator conditions on them to produce the answer [57,291]. At the simplest level, the pipeline follows a retrieve-then-generate paradigm: a retriever fetches top-k evidence, and the LLM then reasons over the retrieved evidence [57,292]. More advanced variants use an LLM to adaptively control the retrieval process, deciding when to retrieve, how many rounds to run, and whether additional evidence is required [58,293]. Query formulation can be improved by using an LLM to generate a hypothetical pseudo-document as the retrieval query, which is then embedded and used to retrieve nearest-neighbor passages [294,295]. For long or heterogeneous sources, retrieval can be strengthened by building a hierarchical summary tree index: documents are split into leaf chunks, related chunks are summarized into linked parent nodes, and retrieval descends from top-level summaries to leaf chunks containing supporting spans [296]. LLMs can improve retrieval by imposing structure on the corpus. They extract entities and relations from documents to build a graph, group

densely connected entities into clusters, and use an LLM to summarize cluster passages as higher-level graph nodes for coarse-to-fine retrieval. At query time, the system retrieves the most relevant subgraph or group summaries and uses the linked original text spans as evidence for generation [297,298].

LLM-based retrieval improves flexibility by reformulating information needs [57], but it reduces transparency and adds audit burden. Retrieve-then-generate is efficient when top-k evidence is reliable, while LLM-controlled retrieval better handles ambiguous claims but costs more and may propagate errors [291]. Hierarchical or graph-based RAG suits long corpora [296], but stronger uncertainty handling is still needed. A key practical risk is retrieval failure: irrelevant evidence may introduce noise and bias the verdict, while missing evidence reflects limited retrieval coverage or evidence availability.

5.2.4. Evidence Consolidation and Conflict Resolution

After retrieval, the LLM can consolidate evidence by summarizing search results and filtering irrelevant or duplicated materials. To reduce redundancy, one approach guides LLM evidence summarization with claim-relevant knowledge-graph relation paths [299]. It first extracts key entities from the claim and uses them to retrieve multi-hop relation tuples from a knowledge graph, forming candidate relation paths. It then selects the paths most relevant to the claim and uses them as guidance when the LLM summarizes the retrieved paragraphs, retaining only claim-relevant content while removing redundant text [299].

Sometimes, retrieved evidence is conflicting, meaning some paragraphs support the claim while others refute it [59]. To handle conflicting evidence, an effective approach is to use an LLM as an evaluator to score each evidence item [59,300]. This approach assigns an entailment stance label, such as fully supported, partially supported, unsupported, or contradictory, and predicts utility signals that reflect verification value, including relevance to the claim, evidential usefulness/coverage, source authority, and timeliness. Then, it selects the evidence with the highest score. Conflicting evidence can be mitigated through multi-agent deliberation over retrieved documents [60]. Given a claim and a set of retrieved documents, each document is assigned to an LLM-based agent, which generates an intermediate answer based solely on its own document. An aggregator then collects all drafts and produces a unified answer with an explanation that reconciles inconsistencies and preserves multiple plausible interpretations under ambiguity. In subsequent rounds, agents receive the aggregator's summary and revise their drafts by considering counter-evidence from other agents. After a fixed number of rounds, the aggregator outputs the conclusion based on the converged discussion [60].

LLM-based consolidation can better interpret semantic evidence and partial conflicts than fixed heuristics [51,119], but it is less deterministic and harder to audit. Summarization reduces redundancy [59], yet may omit minority signals. The key challenge is traceable aggregation that preserves provenance and separates real conflicts from retrieval noise [60]. This distinction is non-trivial because apparent conflicts may arise from temporal change, source unreliability, or retrieval noise, each requiring different handling strategies.

5.2.5. Trends and Evolution of Evidence Retrieval

Evidence retrieval in fact-checking has evolved from fixed relevance matching toward more adaptive evidence seeking, organization, and verification. Early retrieval pipelines largely relied on direct claim-to-corpus matching, which is efficient and transparent but brittle when claims are complex, ambiguous, or cross-modal [51]. LLM-based methods extend this stage by reformulating queries, enabling iterative retrieval, and consolidating evidence more intelligently [57,58,278]. Cross-modal retrieval further broadens access to text, images, videos, and audio [56,283]. However, these gains come with higher system

complexity, weaker traceability, and greater difficulty in handling insufficient or conflicting evidence. The central challenge is therefore to improve evidence sufficiency, preserve provenance, and manage conflicts without sacrificing transparency.

5.3. Verdict Prediction

Verdict prediction involves different inference settings depending on the target inconsistency and the available evidence. A system may detect manipulated content, determine whether multimodal elements are contextually aligned, or assess whether a claim is supported by external evidence. These settings rely on different cues, including forensic traces, contextual consistency, and evidence-based support or contradiction. MFC-Bench treats these as separate capabilities, evaluating large vision–language models separately on manipulation detection, out-of-context detection, and verdict classification [301]. Figure 4 illustrates the structure of this section.

5.3.1. Manipulation Detection

Although manipulation detection has traditionally relied on modality-specific forensic models, systems such as FORGERYSLEUTH [72] show that MLLMs can be adapted for manipulated-content detection and localization. This adaptation combines semantic reasoning, task-specific training, and forensic-oriented modules. However, MLLMs remain less reliable for low-level forensic analysis, where direct application can miss subtle artifacts and produce plausible but unsupported forensic rationales [302]. Therefore, subtle pixel, spectral, temporal, or speech artifacts often still require modality-specific forensic methods.

Following the taxonomy proposed by Lin et al. [34], we organize detection methods by modality. For text, common signals include embedded watermarks, residual generation traces, and statistical cues such as perplexity or cross-perplexity [68,303–308]. For images, detectors use generation watermarks, physical or physiological inconsistencies, model fingerprints, and statistical anomalies in texture, gradient, or spectral features [69,309–318]. For videos and audio, detection usually targets temporal inconsistencies, frame-level artifacts, and vocoder traces [70,71,319].

Text detectors are lightweight, but sensitive to rewriting and domain shift [68]. Image, video, and audio detectors can capture modality-specific artifacts [70,71,320], yet each has its own cost and robustness limits. A major gap is generalization across modalities when artifacts are subtle, compressed, or distributed.

5.3.2. Out-of-Context Detection

For out-of-context (OOC) detection, early systems framed the task as cross-modal semantic consistency checking and typically followed an encode, fuse, and classify pipeline. Text is encoded with TF-IDF n-gram features [45] or word2vec embeddings [321], while images are represented by global CNN features. Many methods further use region features from Mask R-CNN [322] to capture fine-grained cues [67,78,323]. Multimodal fusion in this period was dominated by the use of lightweight operators such as concatenation or dot products. Despite their simplicity, these pipelines often lack semantic grounding [324], making them fragile when captions are plausible yet contextually mismatched [153].

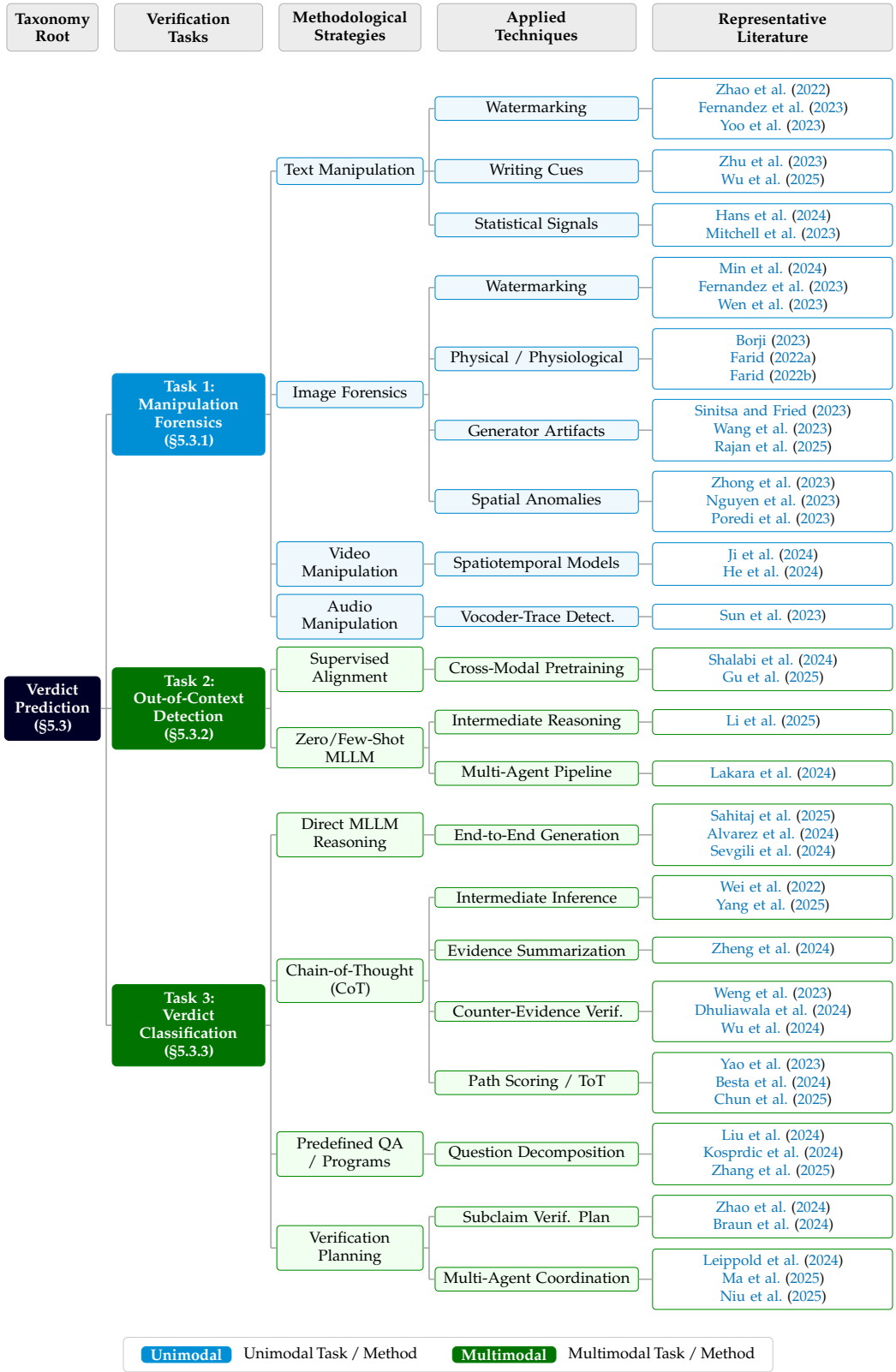


Figure 4. Taxonomy of verdict prediction methods. Blue nodes denote unimodal methods, whereas green nodes denote multimodal methods. The figure establishes a top-down hierarchy across three parallel tasks. Manipulation forensics identifies intrinsic media tampering, out-of-context detection uncovers authentic media paired with misleading claims, and verdict classification assesses claim veracity using external evidence. Each task then branches into corresponding methodological strategies and applied techniques.

MLLM methods improve OOC detection by leveraging cross-modal alignment learned from large-scale image–text pretraining [56]. A straightforward line of work adapts an MLLM to the OOC setting via task-specific fine-tuning [325,326]. Beyond direct fine-tuning, a line of work generates an intermediate "coexistence relationship" with an MLLM to explain why images and text may appear together, and uses this relationship to filter retrieved evidence [73]. Another line of work frames OOC detection as multi-agent deliberation grounded in multi-grained evidence retrieval. It first retrieves contextual evidence at multiple semantic levels, including entities, attributes, relations, and events, and then assigns LLM-based agents to verify different aspects of cross-modal consistency. The agents exchange intermediate findings and aggregate their judgments to produce the final decision [327].

Traditional fusion models are efficient for screening, but they mainly judge relatedness rather than real-world consistency. MLLM-based or agentic methods offer stronger entity, relation, and event reasoning, but they depend on external evidence and are slower. Lightweight models suit screening, while richer methods are better when provenance matters.

5.3.3. Verdict Classification

Early verdict classification systems typically modeled verification as reasoning over a claim and evidence. They encoded the claim and candidate evidence separately, computed claim–evidence relevance for evidence selection, and then fused the selected evidence with the claim representation for label prediction [51,61]. These text-only pipelines cannot use visual evidence when the factual inconsistency depends on objects, scenes, or other image-level cues. In contrast, multimodal verdict classification incorporates object-level region visual representations as additional evidence for verification [67]. However, these pipelines rely on shallow matching and fixed aggregation, which often leads to poor performance when evidence is long or indirect, when multi-hop reasoning is required [183], or when retrieval errors [277] and domain shifts [36,328] are present.

LLM-based fact-checking systems mitigate these issues by replacing fixed fusion and classifier heads with generative verification over natural-language evidence. Given a claim and a set of retrieved evidence passages, the model directly conditions on both as input and produces the verdict judgment through language generation [62,63,329–331].

To improve the reliability and interpretability of end-to-end reasoning, many works structure inference at test time. CoT prompts LLMs to generate explicit intermediate reasoning steps before producing a verdict label, making the verification process more reliable and interpretable [14]. In the verdict prediction stage, given a claim and a fixed set of evidence passages, LLMs can be used through four complementary reasoning paradigms. First, explicit step-wise inference requires the model to generate a fixed set of intermediate judgments, such as entity match, temporal consistency, or evidence coverage, each grounded in the evidence. The final verdict label is then derived from these intermediate judgments through a predefined aggregation rule [14]. Second, abstraction-first inference asks the model to first compress the evidence into a structured fact sheet containing key events, entities, times, locations, and numbers. The model then compares the claim against this sheet to determine whether it is supported, contradicted, or not sufficiently covered by the evidence [332]. Third, verification after the decision adopts a self-auditing procedure. Following Chain-of-Verification, the model first gives an initial verdict, extracts the key conditions behind it, turns them into verification questions, answers each question with retrieved evidence, and revises the verdict when a check fails [333]. For a claim such as "this image shows an event in city X in 2026," the audit criteria include whether the evidence supports the claimed location, whether the date is consistent with 2026, whether

the image has appeared in earlier sources, and whether any retrieved source contradicts the claimed time or place. The model revises the verdict if these checks reveal a mismatch, prior image use, ignored counter-evidence, or unsupported assumptions [334–336]. Fourth, verification can be framed as reasoning-path search over multiple possible verdicts. The system maintains separate reasoning paths for candidate verdicts and iteratively expands them with intermediate conclusions grounded in the evidence, finally returning the verdict associated with the strongest evidence-consistent path [337,338]. In multimodal settings, a common approach is to use an MLLM as an inspector that analyzes clues across modalities and produces an intermediate output. This output is then reused for prediction and evidence localization, enabling the system to return both the final verdict and the most relevant supporting visual region or segment [339]. Beyond visual grounding, MAD [203] and MAD2 [204] extend verdict classification to audio. They represent two common audio verification settings, with MAD first converting the spoken dialogue into text and then verifying the claim from the transcript, and MAD2 fusing audio and dialogue context before making the final verdict.

LLMs are not always used to predict verdict labels directly, but can instead answer a fixed set of verification questions about the message. These outputs are then mapped to symbolic values and aggregated by a decision module to produce the final verdict label [64,202,213]. When labeled data is scarce, LLMs can synthesize supervision for vision–text alignment by constructing intentionally mismatched image–text pairs. A common strategy rewrites real text using semantically similar candidates as anchors and selectively replaces key entities, such as names, locations, or numbers, so that the edited text remains fluent but becomes inconsistent with the image [202].

LLMs can also act as controllers with two execution paradigms. In one setting, the controller decomposes a claim into subclaims and executes a verification plan by dispatching each step to specialized reasoning modules and aggregating their outputs into a final verdict label [55,65]. In another setting, the controller orchestrates multiple specialized agents through structured critique or debate and derives the final verdict by reconciling their intermediate judgments [83,267,340,341].

From the perspective of evidence access, verdict prediction methods can be grouped into closed-book, fixed-evidence open-book, and retrieval-augmented regimes. Closed-book LLM verification directly prompts the model to predict a verdict from parametric knowledge, making it efficient but weakly grounded for current or source-specific claims [11,342]. Fixed-evidence open-book verification conditions on a given claim–evidence pair, including PLM classifiers, evidence-conditioned LLM generation, step-wise inference, fact-sheet abstraction, and predefined QA, but remains bounded by retrieval quality and evidence coverage [51,61,62]. Retrieval-augmented verification lets LLMs or agents reformulate queries, retrieve new evidence, self-audit or plan, and revise judgments, improving adaptiveness at the cost of higher latency and more failure points [55,65,333].

LLM-based verification makes reasoning more explicit through direct checking, CoT, predefined QA, or module coordination [55,65]. Direct verification is the cheapest, but it relies heavily on reliable evidence and is prone to unsupported judgments. CoT improves interpretability for multi-step reasoning, but it adds latency and may produce ungrounded rationales [14]. Structured QA is more controllable for recurring checks, while planning is more adaptive for complex claims [55,64,65,213]. A shared limitation is weak uncertainty calibration: LLMs may produce confident verdicts even with insufficient, ambiguous, or conflicting evidence [343]. Another limitation is the uneven coverage across modalities, as audio-based fact-checking remains far less explored than text-, image-, and video-based methods [344]. The key direction is controllable, verifiable reasoning with traceable checks and broader coverage of audio-based fact-checking.

5.3.4. Trends and Evolution of Verdict Prediction

This section reviews verdict prediction from a task-level perspective, showing a progression from local artifact inspection to contextual consistency checking and then to evidence-grounded decision-making. Manipulation detection relies on forensic cues and is least dependent on external evidence. Out-of-context detection shifts to cross-modal grounding and contextual alignment. Verdict classification requires integrating claims and evidence to produce explicit judgments, often through multi-step reasoning and calibration. As the abstraction level rises, methods become more flexible but also more costly, more dependent on retrieval quality, and harder to audit. The future direction is to combine robust local cues with traceable reasoning for verifiable and uncertainty-aware verdicts.

5.4. Justification Production

Following prior work that categorizes justifications by output form, fact-checking justifications are commonly expressed either as natural-language explanations or as evidence-oriented forms such as highlighted evidence spans and structured factual triples [27]. The corresponding generation paradigms differ depending on the form of justification to be produced. Figure 5 illustrates the structure of this section.

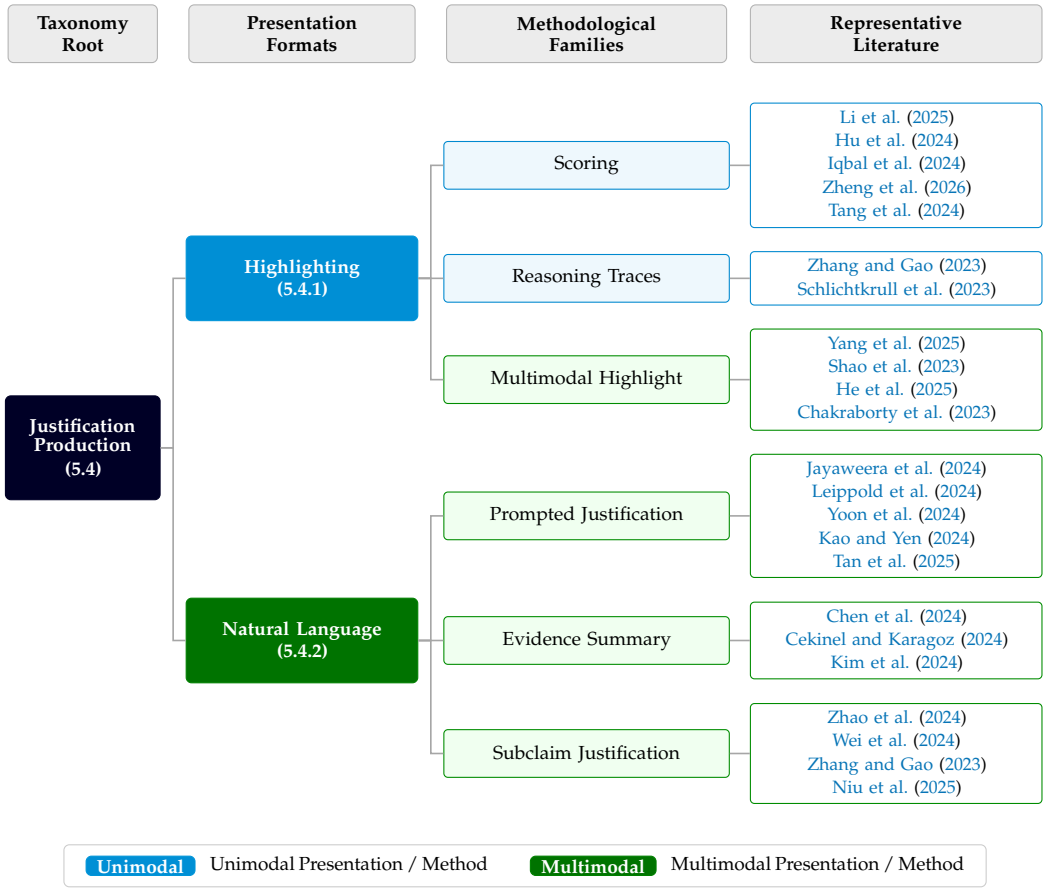


Figure 5. Taxonomy of justification production methods. Blue nodes denote unimodal methods, whereas green nodes denote multimodal methods. The figure separates justifications by presentation form, including highlighting-based justifications and natural-language justifications. Within each format, it further lists the corresponding production methods, such as scoring, reasoning traces, multimodal highlighting, prompted justification, evidence summarization, and subclaim-level justification. Top header boxes delineate the taxonomic progression across presentation forms, methodological families, and the representative literature.

5.4.1. Highlighting Justifications

The first category highlights the parts of the input that contribute to the models' results and uses them as justifications. Early methods typically treat justification as selection, using attention-based joint prediction to assign relevance weights to claim or evidence segments while predicting the verdict label [77,345–347]. In multimodal settings, this idea is extended by decomposing the input into modality-specific units, such as text spans, image regions, or video frames, and applying cross-modal attention to highlight relevant units [78,348,349]. Another line of work represents evidence as entity–relation graphs and uses selected subgraphs or subject–predicate–object triples as structured justifications, which can also be extended to multimodal evidence by unifying entities and relations across modalities [79,350–354]. However, attention-based highlights may reflect correlation rather than causal evidence, while graph-based justifications can miss fine-grained or implicit constraints because they depend on upstream extraction quality and predefined schemas [15,355–358].

LLM/MLLM-based methods address these limitations by framing justification as an explicit, evidence-grounded generation process [80,359,360]. A typical setup treats the LLM as a structured evidence selector: given a claim and retrieved evidence items, it assigns stance and utility labels to each item and returns selected evidence IDs or highlighted snippets in a structured format such as JSON [205,361]. This makes justification an explicit and auditable selection process rather than an implicit attention artifact [362–364]. A closely related paradigm produces subclaim-aligned structured justifications. The model decomposes a complex claim into atomic subclaims, retrieves evidence for each one, and assigns a stance label such as support, refute, or insufficient. The final output maps each subclaim to its evidence and stance in a structured form, making explicit which parts of the original claim are supported, refuted, or uncertain [365,366].

A second setup uses LLMs to produce procedural reasoning traces as structured justification. The model decomposes the claim, generates targeted verification questions, and answers them. The resulting subclaim list and question–answer trace form a structured justification that exposes missing premises and uncertainties while avoiding unsupported inferences [184,367].

In multimodal settings, these structured paradigms can be extended to grounded visual evidence. An MLLM can align visual rationale units with specific claim components by selecting frames, objects, or regions and linking them with the claim text. This yields a multimodal structured justification that explicitly states which visual cues support the claim and how they interact with the accompanying text [157,339,368,369].

Attention-based highlighting is fast and useful for locating relevant spans [77], regions [348], or frames [349], but it may reflect correlation rather than true evidence. Graph-based justification is more structured and interpretable [79], though it depends on extraction quality. LLM/MLLM-based justification is more flexible and auditable [80], but also more costly and sensitive to prompts and evidence quality.

5.4.2. Natural Language Justifications

The second category emphasizes providing short, concise, and less redundant natural language summaries. Traditional methods usually produce such justifications through extractive or abstractive summarization. Extractive methods first select relevant evidence sentences or spans and then organize them into a concise explanation through ordering, redundancy removal, and lightweight post-processing [74,81,370]. Abstractive pipelines further rewrite selected snippets with sequence-to-sequence models such as BART [371] and T5 [372]. For multimodal evidence, visual objects, scenes, temporal events, or cross-modal relations are often first converted into text through OCR, ASR, or cap-

tioning. Although these pipelines are efficient and readable, they may omit necessary context or implicit premises, and multimodal conversion errors can distort key entities or quantities [226,373–375]. These limitations motivate a shift from sentence selection or pipeline-based verbalization to explanation generation that can integrate context, evidence, and verdicts more explicitly.

For natural language justifications, LLM-based methods move beyond simple extraction or summarization by composing a coherent explanation from selected evidence and explicitly connecting it to the final verdict. They also allow finer control over the form of the explanation, while attribution and self-checking improve reliability by identifying and revising unsupported statements [80,359,360]. A simple and widely used setup is prompted justification generation. The input to the LLM typically includes the claim, the predicted label, and the top-k evidence passages, and the prompt asks for a short paragraph explaining why the evidence supports or contradicts the label [205,209,230,341,376,377]. In multimodal settings, this post-hoc pattern extends by adding multimodal inputs such as images, videos, or audio. The prompt then requires the model to explicitly reference cross-modal cues together with retrieved textual evidence when constructing the justification [55,202,378–380].

A second paradigm places justification before prediction and uses it as an intermediate representation for verdict prediction. Instead of generating justification for predicted labels, the LLM distills evidence into a claim-focused justification summary that retains only the information needed to validate the claim while restoring minimal missing context and removing irrelevant content. A downstream decision module then consumes the distilled justification to predict the verdict label [82,381]. In multimodal settings, the same pre-verdict distillation can be extended to multimodal evidence units. The system first summarizes what the visual and textual evidence establishes into a cross-modal justification, which is then used by the downstream decision module instead of the full raw multimodal context [382].

A third paradigm produces justifications at the subclaim level for complex or multi-proposition claims. The LLM aggregates subclaim-level justifications into a global verdict by first decomposing the claim into atomic subclaims and generating a local justification for each using the most relevant evidence [65,280,367]. This design improves coverage by verifying each atomic claim separately and making explicit which parts are supported, refuted, or uncertain. In multimodal settings, a related strategy generates justifications for different visual regions and then aggregates them into an overall explanation [83].

Post-hoc explanations are cheap, but they may be distorted by the predicted verdict [80,359]. Pre-verdict distillation can guide decisions [82], but errors in the explanation can mislead the model. Subclaim-based justification is best for complex claims [83], though it is the most expensive. A remaining issue is evaluating the faithfulness, completeness, and consistency of natural-language explanations. This is important because natural-language explanations may be verdict-consistent but still fail to reflect the evidence actually used or the reasoning process behind the decision.

5.4.3. Trends and Evolution of Justification Production

Justification production is moving from evidence selection toward explanation generation. Highlighting methods that localize supporting cues are fast and traceable, but they often rely on shallow correlations or predefined schemas, which limits coverage for implicit or cross-modal claims [77–79]. By contrast, LLM-based generation produces more fluent and context-aware explanations and is better suited to complex or multimodal evidence [371,372]. However, it is also more vulnerable to omission, distortion, and post hoc drift when grounding is weak [378]. The overall trend is toward more semantic and

explanatory justifications, with increasing emphasis on readability, coverage, and explicit grounding. The future direction is to combine explicit evidence linkage with faithful natural-language generation, so explanations remain both user-friendly and verifiable [27,383].

5.5. System Architectures Beyond the Linear Pipeline

The four-stage pipeline used in this survey is an analytical framework rather than a strict execution order. In real-world deployment, fact-checking systems rarely adopt a single architecture or paradigm, since traditional, LLM-based, and MLLM-based methods differ in latency, auditability, grounding ability, robustness, and hallucination risk. Traditional modules are efficient and traceable, LLMs enable flexible decomposition and evidence-conditioned reasoning, and MLLMs support visual or audiovisual grounding, but no paradigm dominates across all deployment requirements. A detailed stage-wise comparison is provided in Appendix B, Table A1. This motivates cost- and risk-adaptive architectures, which we summarize as monolithic single-shot verification, modular pipelines, and agentic iterative verification.

A monolithic single-shot system uses one model to produce a verdict, and sometimes a justification, in a single pass [62,63,329,377]. It has low orchestration overhead and is suitable when latency is the main constraint, such as rapid triage or low-risk screening. Its main weakness is poor error attribution: if the model misunderstands the claim, ignores evidence, or generates an unsupported rationale, it is difficult to identify which step failed. Thus, it is less appropriate when high accuracy, evidence provenance, or auditability is required.

A modular pipeline separates claim extraction, evidence retrieval, verdict prediction, and justification production into distinct components. This design supports component-level evaluation, caching, replacement, provenance tracking, and human review. It is appropriate when reliability, traceability, and auditability are more important than low-latency inference. However, feedforward modular pipelines lack feedback mechanisms to correct intermediate failures. When upstream claim formulation produces skewed representations, dense retrievers fetch superficially similar media rather than decisive counter-evidence. Lacking abstention mechanisms, downstream verdict models treat this deficient evidence as exhaustive, falling back on parametric memory or unimodal verification [20,176]. Subsequently, justification models produce post-hoc rationalizations, hallucinating evidence details to defend the erroneous verdict [27]. Because errors compound monotonically across module boundaries, rigid pipelines cannot self-correct without architectural interventions. Thus, modular systems require explicit fallback policies when evidence is missing or contradictory.

Agentic systems use planner-executor or critic-actor loops to adapt the verification process during inference. To break this feedforward error cascade, iterative verification paradigms transform fact-checking from a static pipeline into an adaptive closed-loop process. A planner-executor loop can decompose a claim, generate search queries, retrieve evidence, and assign subtasks to specialized modules. A critic-actor loop can check an initial verdict, identify unsupported reasoning, request additional evidence, and revise the output [333,335,336]. Through bidirectional feedback, iterative systems refine queries and retrieve additional evidence when initial findings are incomplete, preventing models from making unsupported guesses. In addition, critic modules verify that visual evidence truly supports the textual claim before finalizing the verdict, stopping the model from generating fabricated justifications. Multi-agent systems further distribute subclaims, evidence sources, or reasoning roles across agents and aggregate their judgments [83,340,341]. This collaboration allows visual forensic and evidence retrieval agents to cross-check their findings, preventing an error in one modality from misleading the final verdict. This

architecture is appropriate for complex, open-domain, multimodal, or contested claims where accuracy and evidence coverage are more important than latency [267]. Its main risks are higher cost, longer runtime, coordination errors, and weaker reproducibility. In deployment, the relative trade-off among these three architectures is governed by five operational factors. Under strict computational resources and low claim complexity, monolithic triage maximizes throughput. As claim complexity increases or visual information relevance demands cross-modal verification, systems must shift to modular pipelines to ensure auditability. Furthermore, when retrieval quality is noisy or evidence completeness is deficient, feedforward pipelines suffer from evidential starvation, requiring iterative agentic loops to reformulate queries and resolve conflicts. Together, these five factors guide architectural selection, balancing throughput, evidence traceability, and verification robustness according to task demands.

6. Challenges and Future Directions

This section outlines the challenges of LLM-based fact-checking and highlights promising research efforts to address them, suggesting future directions. To make the discussion more actionable, Table A4 in Appendix B provides a qualitative maturity assessment of the nine challenges based on fact-checking-specific benchmarks, mitigation methods, and deployment readiness.

6.1. Complexity Trade-Offs in Multimodal Fact-Checking

Misinformation on the internet is often disseminated not only in text, but also together with images, videos, or audio. Therefore, multimodal fact-checking systems are becoming increasingly important [384]. Compared to unimodal fact-checking, multimodal systems introduce additional steps such as visual feature extraction and cross-modal alignment. Some systems further incorporate real-time image retrieval capabilities [55]. These steps significantly increase computational costs. While multimodal fact-checking processes generally preserve the integrity of evidence, their high latency has rarely been studied. Therefore, understanding these efficiency trade-offs is crucial for designing scalable uni- and multimodal fact-checking systems.

6.1.1. Empirical Case Study

To evaluate the efficiency differences between unimodal and multimodal information processing for fact-checking tasks, we use DEFAME [55], a dynamic evidence-based fact-checking framework built on GPT-4o (2024-12-17) (because commercial API models may be updated over time, the reported latency and accuracy results are tied to the model snapshot used in our experiments and may not hold across future model upgrades) and LLaVA-OneVision-Qwen2-7B-ov-hf. We select AVeriTeC [184], FEVER [51], MOCHEG [74], and VERITE [176] to cover both unimodal and multimodal fact-checking. AVeriTeC, MOCHEG, and VERITE are directly supported by DEFAME with ready-to-use pipelines, covering open-web claim verification, textual-visual evidence fact-checking, and image-context verification. We further include FEVER as a widely used Wikipedia-based textual claim verification benchmark. To ensure consistency, the exact same fixed set of 500 claims per benchmark is selected via stratified random sampling with a fixed random seed of 42 to preserve the original class distribution, and evaluated across all configurations over five repeated runs. All experiments involving the open-source multimodal backbone are conducted on a single RTX 4090 GPU. The prompt templates used for verdict prediction are reported in Appendix A, Listing A1. Table 9 summarizes our results.

Table 9. Runtime, token consumption, and tool call counts under different experimental configurations on four datasets. "Ret." denotes whether retrieval is used (Yes/No). "Calls" denotes the number of LLM calls. "Google", "GVis", and "DDG" denote the numbers of retrieval calls to Google Search, Google Vision Search, and DuckDuckGo, respectively. All numeric results are reported as the mean ± standard deviation over five repeated runs. T denotes text-only input, and T/I denotes text-image input.

Dataset	Model	Input	Ret.	Time (s)	s/Claim	Calls	Input Tokens	Output Tokens	Google	GVis	DDG	Acc	F1
AVeriTeC	GPT-4o	T	No	4004.9 ± 312.5	8.01 ± 0.68	2042 ± 189	884,934 ± 67,823	116,696 ± 9127	0 ± 0	0 ± 0	0 ± 0	0.426 ± 0.007	0.341 ± 0.012
AVeriTeC	GPT-4o	T	Yes	18,650.3 ± 1520.1	37.30 ± 3.12	4410 ± 412	917,413 ± 74,582	326,746 ± 28,913	1004 ± 178	0 ± 0	25 ± 6	0.586 ± 0.018	0.401 ± 0.009
AVeriTeC	LLaVA	T	No	7354.2 ± 598.3	14.71 ± 1.21	1915 ± 168	890,628 ± 71,944	344,122 ± 29,675	0 ± 0	0 ± 0	0 ± 0	0.408 ± 0.003	0.307 ± 0.015
AVeriTeC	LLaVA	T	Yes	16,325.6 ± 423.5	32.65 ± 0.85	4202 ± 390	1,004,439 ± 88,306	429,083 ± 37,214	1013 ± 192	0 ± 0	23 ± 7	0.508 ± 0.022	0.347 ± 0.026
FEVER	GPT-4o	T	No	4435.1 ± 354.2	8.87 ± 0.74	2096 ± 185	913,752 ± 73,519	99,103 ± 8276	0 ± 0	0 ± 0	0 ± 0	0.642 ± 0.005	0.609 ± 0.010
FEVER	GPT-4o	T	Yes	16,210.2 ± 1350.5	32.42 ± 2.76	4219 ± 378	1,531,210 ± 127,438	269,093 ± 23,812	1044 ± 189	0 ± 0	33 ± 8	0.699 ± 0.014	0.664 ± 0.021
FEVER	LLaVA	T	No	6514.0 ± 542.7	13.03 ± 1.09	1885 ± 162	825,752 ± 67,905	304,896 ± 26,113	0 ± 0	0 ± 0	0 ± 0	0.518 ± 0.016	0.485 ± 0.008
FEVER	LLaVA	T	Yes	14,700.8 ± 1210.0	29.40 ± 2.48	4273 ± 389	1,442,697 ± 118,624	747,677 ± 62,557	933 ± 165	0 ± 0	16 ± 4	0.569 ± 0.025	0.513 ± 0.013
MOCHEG	GPT-4o	T	No	7349.3 ± 612.8	14.70 ± 1.27	1513 ± 136	973,563 ± 80,539	136,857 ± 11,682	0 ± 0	0 ± 0	0 ± 0	0.403 ± 0.011	0.391 ± 0.006
MOCHEG	GPT-4o	T	Yes	16,883.5 ± 1420.0	33.77 ± 2.92	3595 ± 328	3,384,400 ± 284,276	242,601 ± 20,884	900 ± 162	0 ± 0	115 ± 19	0.543 ± 0.019	0.527 ± 0.027
MOCHEG	GPT-4o	T/I	No	9275.9 ± 798.4	18.55 ± 1.65	1487 ± 131	2,198,224 ± 191,583	116,231 ± 10,419	0 ± 0	0 ± 0	0 ± 0	0.415 ± 0.009	0.399 ± 0.013
MOCHEG	GPT-4o	T/I	Yes	24,667.0 ± 2150.0	49.33 ± 4.42	4321 ± 398	7,869,913 ± 684,701	285,953 ± 25,112	561 ± 108	475 ± 92	42 ± 9	0.579 ± 0.008	0.571 ± 0.015
MOCHEG	LLaVA	T	No	10,390.0 ± 896.5	20.78 ± 1.84	1832 ± 167	1,123,353 ± 98,264	247,763 ± 21,348	0 ± 0	0 ± 0	0 ± 0	0.312 ± 0.014	0.275 ± 0.018
MOCHEG	LLaVA	T	Yes	21,675.5 ± 1880.0	43.35 ± 3.88	3079 ± 285	3,241,486 ± 284,633	267,727 ± 23,519	1362 ± 215	0 ± 0	30 ± 7	0.367 ± 0.023	0.320 ± 0.011
MOCHEG	LLaVA	T/I	No	25,234.7 ± 2210.0	50.47 ± 4.56	2163 ± 198	2,484,524 ± 217,836	288,439 ± 25,672	0 ± 0	0 ± 0	0 ± 0	0.347 ± 0.004	0.284 ± 0.017
MOCHEG	LLaVA	T/I	Yes	37,451.3 ± 3280.0	74.90 ± 6.68	5030 ± 462	10,108,833 ± 891,524	298,541 ± 26,845	1719 ± 285	802 ± 146	58 ± 11	0.403 ± 0.020	0.379 ± 0.024
VERITE	GPT-4o	T	No	5575.4 ± 468.2	11.15 ± 0.96	1511 ± 136	928,483 ± 80,177	149,851 ± 13,026	0 ± 0	0 ± 0	0 ± 0	0.630 ± 0.012	0.556 ± 0.005
VERITE	GPT-4o	T	Yes	15,477.2 ± 1350.0	30.95 ± 2.78	3283 ± 296	2,715,290 ± 237,615	269,146 ± 23,784	1886 ± 275	0 ± 0	444 ± 76	0.651 ± 0.028	0.589 ± 0.019
VERITE	GPT-4o	T/I	No	7423.9 ± 642.5	14.85 ± 1.32	1533 ± 138	1,728,273 ± 151,832	190,531 ± 16,779	0 ± 0	0 ± 0	0 ± 0	0.653 ± 0.015	0.594 ± 0.010
VERITE	GPT-4o	T/I	Yes	28,305.2 ± 2480.0	56.61 ± 5.08	3343 ± 305	5,696,962 ± 501,883	235,444 ± 20,873	1369 ± 228	353 ± 68	32 ± 8	0.657 ± 0.017	0.630 ± 0.029
VERITE	LLaVA	T	No	6195.7 ± 532.6	12.39 ± 1.09	1240 ± 112	994,828 ± 87,449	195,309 ± 17,215	0 ± 0	0 ± 0	0 ± 0	0.483 ± 0.019	0.446 ± 0.011
VERITE	LLaVA	T	Yes	19,837.2 ± 1750.0	39.67 ± 3.62	2500 ± 232	2,004,864 ± 176,903	234,813 ± 20,895	2013 ± 312	0 ± 0	12 ± 4	0.580 ± 0.013	0.572 ± 0.026
VERITE	LLaVA	T/I	No	13,680.4 ± 1190.0	27.36 ± 2.44	1240 ± 112	1,394,828 ± 122,574	216,517 ± 19,202	0 ± 0	0 ± 0	0 ± 0	0.634 ± 0.006	0.580 ± 0.014
VERITE	LLaVA	T/I	Yes	33,414.8 ± 2950.0	66.83 ± 5.78	4182 ± 385	4,903,413 ± 431,682	255,313 ± 22,675	1572 ± 256	736 ± 128	71 ± 14	0.646 ± 0.024	0.597 ± 0.012

Note: All metrics in this table are derived from original empirical experiments conducted in this study. All metrics report the mean ± standard deviation over five repeated runs on the exact same 500 claims per benchmark. Model configurations use fixed snapshots of GPT-4o (17 December 2024) and local LLaVA-OneVision-Qwen2-7B-ov-hf on a single RTX 4090. Inference parameters enforce greedy decoding with temperature T = 0 and a fixed maximum generation length. Prompt settings use standardized templates with JSON schemas from Appendix A. Retrieval conditions and evidence availability use uniform DEFAME search settings (top 3 web pages per query via Serper API, Google Vision, and DuckDuckGo).

Results show that multimodal processing leads to higher latency than unimodal settings. Without real-time retrieval enabled, the average processing time per claim in the multimodal configuration increases by 26.2% to 33.2% for GPT-4o and by 120.8% to 142.8% for the LLaVA backbone, compared to the unimodal configuration. This suggests that multimodal reasoning can improve verification quality when image context is necessary, but it is significantly more expensive than text-only processing. Retrieval further amplifies this cost by increasing latency, model calls, token usage, and search operations, although it generally improves accuracy and macro-F1. When real-time retrieval is enabled, latency increases by 48.4% to 365.7% compared to the non-retrieval setting. Furthermore, longer text inputs incur greater latency owing to the computational complexity of Transformer-based architectures in handling long sequences [385]. These findings highlight the inherent trade-offs in multimodal fact-checking systems, where improvements in retrieval completeness or accuracy often come at the expense of increased computational overhead. The RTX 4090 setup provides a controlled single-node environment for comparison and bottleneck analysis, while production deployment would require distributed serving, workload-aware scheduling, and explicit service-level objectives.

To evaluate classification differences across identical claims, McNemar's test [386] is calculated per run and averaged across the five repeated runs. This test confirms that real-time retrieval yields statistically significant accuracy gains over non-retrieval across all four benchmarks with $p < 0.001$. Differences in macro-F1 across the five repeated runs are evaluated using a paired two-tailed t-test. On VERITE under GPT-4o with retrieval, accuracy, class balance, and computational cost do not scale equally. Adding image inputs yields no statistically significant difference in accuracy from 0.651 to 0.657 with McNemar's test $p = 0.42$. In contrast, it produces a statistically significant improvement in macro-F1 from 0.589 to 0.630 with paired t-test $p = 0.018$ by detecting out-of-context miscaptioning. Yet, this balanced gain incurs an 82.9% latency increase and doubles token usage. Thus, multimodal retrieval is justified only when out-of-context detection is critical, while latency-constrained deployment favors text-only retrieval.

The empirical findings in Table 9, alongside literature synthesis, illustrate how five core factors govern relative performance and operational trade-offs. First, as established in previous research [51,184], claim complexity dictates retrieval dependence, as simple claims in FEVER [51] achieve high baseline accuracy without search, whereas complex claims in AVeriTeC [184] demand extensive retrieval to resolve. Second, directly evaluated in our experiments via multimodal ablation, visual information relevance determines multimodal utility, where decisive imagery in VERITE [176] yields meaningful macro-F1 gains that justify higher latency, whereas weakly relevant images merely inflate token expenditure. Third, as supported by the prior literature [32,55], retrieval quality governs search efficiency, as noisy open-web sources trigger repeated tool invocations and rapid token inflation. Fourth, directly substantiated in our experiments through retrieval toggling, evidence completeness bounds verification accuracy, where completing the evidence chain via retrieval yields the largest performance gains across all benchmarks. Finally, as highlighted in prior deployment studies [246,387], computational resources dictate trade-offs between the inference latency of local deployment and the cumulative token costs of commercial APIs.

6.1.2. Deployment Considerations

At the model-selection level, proprietary frontier models such as GPT-5.5, Claude Opus 4.8, and Gemini 3.5 Flash are often stronger because they are trained on larger and more diverse data and receive more extensive post-training for reasoning and multimodal alignment. However, they also introduce usage-based cost, provider dependency, rate

limits, and data-governance constraints [246]. By contrast, self-hosted open-weight models such as Qwen2-VL-7B [260], InternVL2.5-8B [388], and LLaVA-OneVision-7B [389] provide local control over serving, auditing, caching, and fine-tuning, but shift the burden to GPU resources, energy consumption, capacity planning, and maintenance [387].

In our experiments, GPT-4o and LLaVA-OneVision-Qwen2-7B instantiate these two settings, and GPT-4o performs better overall. Model selection should therefore balance accuracy, cost per claim, privacy, reproducibility, update control, and workload characteristics.

These trade-offs suggest that multimodal fact-checking should be evaluated and deployed with real-time constraints in mind. Beyond accuracy and macro-F1, deployment-oriented benchmarks should report latency, throughput under concurrent load, model-call counts, token usage, retrieval-call counts, modality-specific tool usage, GPU memory consumption, evidence freshness, monetary or energy cost per claim, and explicit latency and freshness requirements [390,391]. In production, systems should use modular pipelines that separate retrieval, evidence filtering, visual or video processing, multimodal reasoning, and verdict generation into independently scalable services [55]. Incremental index updates, selective visual invocation, dynamic batching, autoscaling, and budget-aware routing can control latency and cost, but batching should be scheduled against tail-latency objectives rather than average throughput alone [392]. Video inputs further require frame or clip sampling, ASR/OCR, temporal localization, event-order reasoning, and clip-budget control.

Reliable deployment also requires cache control, provenance, and uncertainty handling. Claim-level, retrieval-level, and embedding-level mechanisms reduce repeated computation, but require versioned evidence sources, retrieval indices, model identifiers, and access policies to avoid stale evidence, inconsistent cache hits, and privacy leakage. Each verdict should be traceable to its evidence, retrieval trace, model version, prompt, decoding parameters, timestamps, and source URLs or evidence hashes, enabling review after model or corpus updates.

When retrieval returns no reliable evidence, the system should abstain or output Not Enough Info. When evidence conflicts, it should preserve the contradiction and return a conflicting or uncertainty-aware verdict if source reliability, timestamp, relevance, and modality consistency cannot resolve it.

These costs motivate a Green AI perspective for multimodal fact-checking, where efficiency is evaluated alongside accuracy [387]. Energy reduction can first be pursued through model compression [393] and distillation [394,395], which transfer large-model behavior to smaller models. Low-precision computation, such as quantization [396], reduces memory use and inference energy. Efficient architectures [397] and parameter-efficient tuning [398,399] lower training and adaptation overhead by reducing unnecessary computation or trainable parameters. At the system level, accelerator-optimized serving, batching, caching, and scheduling improve hardware utilization under real workloads [400,401].

Future research should address the following questions to reduce the computational cost of multimodal fact-checking systems: How can we design staged or adaptive retrieval pipelines that invoke visual modules only when textual evidence is insufficient? Phased retrieval has shown potential in reducing unnecessary visual processing [402]. How can we develop more efficient multimodal architectures and tuning strategies that lower inference overhead without sacrificing performance? Lightweight encoders [48] and parameter-efficient tuning methods such as low-rank adaptation (LoRA) and Adapters [398,403] represent promising directions.

6.2. Adversarial Robustness in LLM-Based Fact-Checking

Research indicates that LLMs are highly susceptible to subtle perturbations, adversarial text rewriting, biased search results, manipulative prompting strategies, prompt

injection, poisoned retrieval corpora, and misleading outputs [404,405]. Owing to its reliance on precise semantic interpretation, reliable evidence retrieval, and strict logical entailment, fact-checking exhibits inherent vulnerability to adversarial perturbations. This vulnerability renders robustness a foundational requirement for the reliability of LLM-based fact-checking systems [406,407].

Following the OWASP GenAI Top 10 [390], these risks can be organized along the fact-checking pipeline. At the claim level, adversarial rewritings may preserve human-perceived meaning. For example, consider a claim rewritten from "The President refused to sign a bill" to "The President did not approve a bill". Such modifications can nonetheless alter the semantic interpretation of LLMs, disrupting alignment and degrading evidence retrieval [408]. Direct prompt injection in the claim further corresponds to LLM01: Prompt Injection. At the evidentiary level, attackers can inject forged or misleading content into retrieved documents [409], corresponding to LLM04: Data and Model Poisoning, while malicious instructions hidden in retrieved documents correspond to indirect prompt injection under LLM01. Retrieval-side leakage and embedding manipulation are related to LLM02: Sensitive Information Disclosure and LLM08: Vector and Embedding Weaknesses. At the image level, manipulated or AI-generated visuals may further distort verification. At the reasoning level, adversarial prompts can interfere with the model's inference process [410], leading to jailbreaks, output manipulation, unsupported verdicts, or unfaithful explanations under LLM05: Improper Output Handling and LLM09: Misinformation. Thus, the main threat classes include claim perturbation, direct prompt injection, indirect prompt injection, evidence poisoning, retrieval manipulation, and verifier/output manipulation. These adversarial pressures collectively underscore the need to strengthen the robustness of LLM-based fact-checking pipelines.

Future research should address the following questions to improve the robustness of fact-checking systems: How can we detect and filter adversarial perturbations during the input and retrieval stages? For claim perturbation and direct prompt injection, input-side defenses may include perplexity-based filters, semantic consistency checks, structured-input constraints, prompt isolation, and LLM-judge filters [411]. For indirect prompt injection, evidence poisoning, and retrieval manipulation, retrieval-side defenses may include source-trust scoring, provenance, and content-attribution checks, retrieval diversification, duplicate detection, and cross-retrieval verification [412]. How can we enhance the model-level resilience of LLMs against manipulated or corrupted evidence? For verifier/output manipulation, verifier-side defenses may include adversarial training, robustness-oriented instruction tuning, structured-output enforcement, uncertainty calibration, abstention under suspicious evidence, and explanation faithfulness checking [410]. In deployment, these defenses should be composed in order: input screening before retrieval, retrieval validation before verdict prediction, and output validation before user-facing explanations.

Finally, red-teaming should complement clean benchmark evaluation before deployment. Fact-checking systems should be stress-tested against the above OWASP-aligned threats, including prompt injection, adversarial rewriting, poisoned evidence, crafted retrievals, jailbreaks, verdict manipulation, unfaithful justifications, and privacy leakage [390]. Robustness evaluation should be separate from clean-accuracy benchmarks and report attack success rate, retrieval robustness, evidence provenance, explanation faithfulness, calibration, abstention behavior, latency, and computational cost.

6.3. Temporal Drift in Fact-Checking

Most fact-checking systems rely on existing datasets whose truth values were fixed at the time of dataset construction. However, when there is a substantial time gap between dataset creation and the moment of verification, "temporal drift" may occur between the

real world and the evidence corpus [413]. This issue has become more prominent with the adoption of LLMs, whose factual predictions are highly sensitive to outdated information. Studies show that LLMs may generate temporally incorrect statements, often referred to as temporal hallucinations. These arise when the model's internal knowledge was acquired at a different point in time than the evidence available at inference [414,415]. For example, a dataset assembled in November 2024 labeled the claim "The current US president is Joe Biden" as True. In October 2025, a real-time retrieval system returns evidence stating that "The current US president is Donald Trump". An LLM-based verifier would then classify this claim as False, contradicting the dataset label, not due to model error but because its outputs reflect the temporal state of its parametric knowledge.

Future research should address the following questions to mitigate temporal drift in fact-checking systems: How can we construct datasets that explicitly encode temporal information? This includes developing standardized time-annotation protocols for claims [416] and building versioned benchmarks that distinguish between volatile and constant categories [417], thereby enabling consistent annotation and comparable evaluation. Time-aware QA benchmarks such as TimeQA [416] and FreshQA [418] further show the value of timestamped questions, temporally constrained evidence, and periodically refreshed answers for evaluating time-sensitive knowledge. How can we model temporal dynamics within fact-checking pipelines? Potential directions include explicit extraction and normalization of event times, publication dates, and temporal relations [413], as well as incorporating the model's own knowledge timestamp and potential obsolescence into temporal reasoning.

6.4. Multilinguality in Automated Fact-Checking

Misinformation circulates rapidly across languages on social media platforms and news outlets, giving rise to challenges beyond those observed in monolingual environments. However, existing fact-checking resources remain predominantly English-centric: most datasets [51,193], evidence corpora, and LLM training data are primarily available in English.

To further investigate this imbalance, we select representative English, non-English, and multilingual fact-checking datasets for comparative analysis, as shown in Table 10. English datasets typically provide fixed or well-defined evidence sources, structured annotations, and support for multiple pipeline stages. This advantage is partly historical and infrastructural [51,193]. Many early benchmarks were based on large and easily standardized sources such as Wikipedia, scientific abstracts, and fact-checking organizations [193]. By contrast, non-English datasets are valuable for regional fact-checking, but many focus on verdict prediction rather than complete end-to-end pipelines. Multilingual datasets often prioritize language coverage. They therefore rely on heterogeneous fact-checking websites and social media, making unified retrieval corpora and consistent evidence granularity harder to obtain [185].

Lower reported scores on multilingual benchmarks often stem from dataset constraints rather than weaker reasoning capabilities [185]. Most English benchmarks provide clear sentence-level evidence annotations, fixed evaluation splits, and multi-stage pipeline support [51,184]. In contrast, non-English and multilingual benchmarks typically provide only raw, article-level web documents without standardized evaluation partitions [113,185]. This coarse granularity forces verifiers to sift through long, noisy pages filled with irrelevant text, which naturally dilutes retrieval precision and depresses classification accuracy. Furthermore, multilingual models face uneven tokenizer allocation, limited representation coverage, and sparser local web sources for non-English queries [419,420]. There-

fore, lower performance in non-English settings does not necessarily reflect an inferior verification methodology.

In addition to data scarcity, cross-language retrieval often fails because of weak language alignment, translation errors, and limited non-English sources [185]. Existing strategies include translation-mediated pipelines, multilingual encoders, and direct prompting of multilingual LLMs [420]. Translation-mediated pipelines can reuse English-centric retrievers and verifiers, but may introduce entity loss, ambiguity amplification, and cultural-context shift, causing errors to propagate from query generation to retrieval and verdict prediction. Multilingual encoders reduce translation dependence, but remain weak for low-resource languages, dialectal variants, and region-specific entities [419]. Direct multilingual LLM prompting avoids explicit translation, but still depends on uneven multilingual pretraining coverage and sparse local evidence [420].

Even when evidence is correctly retrieved, LLMs may still misinterpret its content due to cross-cultural and domain-specific variation [421]. Their parametric knowledge often omits region-specific concepts, thereby limiting their capacity for accurate reasoning across diverse linguistic and cultural contexts [422]. CulturalBench [423] provides a useful lens because it uses human-written and human-verified questions across diverse global regions to test whether models possess region-specific cultural knowledge [423]. For fact-checking, this lens suggests evaluating whether models can recognize local entities and institutions, preserve culturally sensitive meanings, reason within the correct regional or legal context, and avoid imposing dominant culture assumptions on local claims. This is crucial for political, religious, public health, and regional misinformation, where truth conditions may depend on jurisdiction, historical context, local terminology, or community norms.

Table 10. Comparison of representative English, non-English, and multilingual fact-checking datasets. Mod. denotes data modality, where T, I, V, and A denote text, image, video, and audio. Split denotes whether the dataset provides fixed data splits. Evi. Granularity denotes the granularity level of evidence. CE, ER, VP, and JP denote whether the dataset is associated with claim detection and extraction, evidence retrieval, verdict prediction, and justification production, respectively. In the Split column, ✓ indicates that fixed data splits are provided; in the CE, ER, VP, and JP columns, it indicates that the corresponding task is supported.

Dataset	Lang	Mod.	Source	Split	Evi. Granularity	CE	ER	VP	JP
English									
FEVER [51]	En	T	Wikipedia	✓	Sentence level	✓	✓	✓	–
BEIR [121]	En	T	Multiple text corpora	✓	Article level	–	✓	–	–
MS MARCO [122]	En	T	Web	✓	Article level	–	✓	–	–
AVeriTeC [184]	En	T	Fact-checking websites	✓	Sentence level	✓	✓	✓	✓
MOCHEG [74]	En	T/I	PolitiFact/Snopes	✓	Sentence level	–	–	✓	✓
Non-English									
DanFEVER [190]	Da	T	Danish Wikipedia	✓	Sentence level	–	–	✓	–
DuReader [126]	Zh	T	Web	✓	Article level	–	✓	–	–
Weibo [95]	Zh	T/I	Weibo	–	Article level	✓	–	–	–
CheckMate [116]	Hi	T/I/V	ShareChat	–	Article level	✓	–	✓	–
MisDissem [117]	Pt	T/I/V/A	WhatsApp	–	Article level	✓	–	–	–
Multilingual									
X-FACT [185]	25 langs	T	Fact-checking websites	–	Article level	–	–	✓	–
MultiClaim [200]	39 langs	T	Fact-checking websites	✓	Article level	–	✓	✓	–
X-CLAIM [89]	6 langs	T	Fact-checking websites	–	Sentence level	✓	–	–	–
mCLAN [93]	20 langs	T	Fact-checking websites	–	Sentence level	✓	–	–	–
MuMiN [113]	41 langs	T/I	Twitter	–	Article level	✓	–	–	–

Future research should address the following questions to advance multilingual fact-checking systems: How can we construct larger-scale and linguistically balanced fact-checking resources? Such resources are necessary for training and evaluating multilingual verifiers with more consistent linguistic coverage. How can we enhance the multilingual reasoning capabilities of LLMs? Potential directions include multilingual CoT generation [424] and improved cross-lingual inference over claim–evidence pairs [425], enabling models to reason directly across languages rather than relying solely on translation. How can we improve fact-checking performance in culturally diverse settings? Culture-aware reasoning modules may help better capture region-specific semantics and domain knowledge [423].

6.5. LLM Hallucinations and Their Impact on Fact-Checking Tasks

In automated fact-checking, hallucinations generated by LLMs significantly undermine system reliability [426,427]. Following Maynez et al. [428], hallucinations can be divided into intrinsic hallucinations, where the generated verdict or explanation contradicts the provided evidence, and extrinsic hallucinations, where the output adds unverifiable information such as unsupported reasons, motivations, dates, locations, or causal relations.

During the evidence retrieval stage, LLMs may fabricate unsupported content, weakening downstream verification accuracy [428]. For RAG-based fact-checkers, factuality and faithfulness should be distinguished: faithfulness measures consistency with retrieved evidence and can be assessed through attribution and citation support, whereas factuality measures consistency with the real world and usually requires gold labels, external knowledge, or human verification. A faithful response may still be factually wrong if the evidence is outdated, biased, or poisoned, while a factually correct response may be unsupported by the cited evidence. In the judgment stage, exposure to such false evidence often leads to overconfident yet incorrect predictions, impairing uncertainty calibration [22]. Therefore, hallucination mitigation should not only suppress unsupported generation but also model uncertainty in verdicts [429]. Token-level and self-verbalized methods estimate confidence from model outputs, using either output distributions or directly reported confidence [430]. Inconsistent conclusions or contradictory rationales indicate unstable reasoning [431]. Mechanistic methods use abnormal internal representations to signal uncertainty, but remain limited in fact-checking. Since these methods mainly capture model uncertainty, fact-checking also needs evidence-level calibration for weak or conflicting sources. Probabilistic evidence aggregation [432] can trigger further retrieval, deferral [433], or "insufficient evidence" decisions [434]. Evaluation should therefore complement accuracy and faithfulness with calibration metrics, including expected calibration error, Brier score, risk–coverage curves, selective risk, and abstention behavior.

Multimodal fact-checking further amplifies hallucination risks because claims may involve images, videos, audio, metadata, OCR text, and retrieved textual evidence. Multimodal hallucinations include object hallucination, attribute or relation hallucination, visual illusion, fabricated locations or events, and cross-modal inconsistency [435,436]. Thus, evaluation should examine visual grounding, evidence provenance, and rationale-evidence alignment.

Hallucination-related methods should be separated into hallucination detection and prevention. Detection methods include SelfCheckGPT [437], FActScore [239], ALCE-style attribution evaluation [438], reference-free hallucination metrics [439], and RefChecker [363], which verify outputs through resampling, atomic-fact decomposition, citation support, or entailment checking against references or retrieved evidence. Prevention methods include retrieval-time filtering, reranking, diversification, and cross-source verification [241]; training-time grounding-oriented instruction tuning, faithful-citation optimization, absten-

tion learning, and robustness training [58]; and decoding-time constrained or evidence-conditioned generation, generate-then-verify pipelines, self-consistency, and uncertainty-aware abstention [360]. However, integrating these solutions into end-to-end pipelines remains a key challenge for future work.

In the short term, progress is likely to come from stronger RAG, attribution checking, hallucination detection, and uncertainty estimation [440]. In the long term, fact-checking systems should move toward grounded multimodal verification with faithful explanations, together with pipeline-level uncertainty propagation and reliable abstention [441].

6.6. Factual Error Correction

Beyond verdict assignment and justification generation, practical fact-checking systems may also need to correct false or partially false claims. Factual error correction (FEC) aims to revise inaccurate claims or generated statements into versions that are better supported by reliable evidence. Thorne and Vlachos [442] introduced evidence-based FEC, where a system edits a false claim so that the revised claim is supported by given evidence. Later work extended this direction to LLM-generated text: RARR [360] retrieves external evidence and revises unsupported parts of model outputs to improve attribution and factual consistency. Other studies explore constrained evidence-guided editing [443], synthetic factual-error injection for data construction [444], and few-shot LLM prompting for low-resource correction [445].

This area is emerging and increasingly active, but it remains less mature than verdict prediction and evidence retrieval. There is still limited consensus on task formulation, evaluation criteria, and the acceptable degree of rewriting. The central challenge is that correction requires more than detecting an error: a system must localize the unsupported content, identify sufficient evidence, revise only what is necessary, and avoid introducing new hallucinations. Over-correction may distort the original meaning, while under-correction may leave factual errors unresolved [443].

FEC may advance quickly in the short term because it can build on mature components from evidence retrieval, claim verification, text editing, and LLM-based rewriting. However, reliable deployment requires longer-term work on evidence sufficiency, abstention, and rewriting scope. In the near term, benchmark design should separately evaluate error localization, evidence grounding, minimal editing, and correction faithfulness, since a correction may be fluent but unsupported, or it may be factually correct but unnecessarily rewrite the original claim. Longer-term work should develop uncertainty-aware correction pipelines that combine retrieval-augmented revision with abstention when evidence is insufficient. Future evaluation should also include real-time and multimodal settings, together with human-centered protocols for judging whether corrections are accurate, useful, non-misleading, and appropriately conservative.

6.7. Privacy and Federated Fact-Checking

Fact-checking systems increasingly process sensitive user-generated content, such as private messages, political opinions, health-related claims, multimodal posts, and platform metadata. This creates privacy risks across the pipeline, especially when LLM-based systems send claims, retrieved evidence, embeddings, or justifications to external APIs [246]. Although privacy-preserving NLP is an active research area, privacy-preserving fact-checking remains underexplored. Existing work mainly offers general techniques, such as federated learning [446], differential privacy [447], secure aggregation [448], and secure retrieval [449], rather than fact-checking-specific datasets, benchmarks, or deployment protocols.

The core difficulty is the tension between privacy protection and evidence-based reasoning. Removing sensitive details may protect users but can weaken claim interpretation, evidence retrieval, and verdict prediction. Privacy risks also arise at multiple stages: claim detection may expose user content, retrieval may leak queries or embeddings, LLM prompts may reveal private context, and generated justifications may disclose sensitive information [246].

Short-term progress may come by adapting mature methods from federated NLP, privacy-preserving machine learning, and secure retrieval. For example, federated learning can support shared model training without centralizing raw user data [446], while differential privacy and secure aggregation can reduce leakage from model updates [447,448]. However, reliable deployment requires longer-term work because current privacy-preserving methods are not yet tailored to evidence-grounded reasoning, multimodal inputs, or LLM-based fact-checking [449]. Future work should develop privacy-aware benchmarks and reporting standards. Beyond accuracy and evidence quality, evaluation should consider privacy leakage, query or embedding exposure, memorization, and privacy-utility trade-offs. Promising directions include federated cross-platform benchmarks [450], privacy-aware evidence access protocols, and human-centered evaluation of justifications that remain informative without revealing sensitive user information [451].

6.8. Calibration and Reliable Uncertainty Estimation

Calibration remains underexplored in fact-checking research. Existing benchmarks and surveys mainly emphasize verdict accuracy, evidence–verdict consistency, and explanation quality, but rarely treat confidence reliability as a core evaluation target [24,25,51]. This gap is critical because uncertain and NOT_ENOUGH_INFO labels are meaningful only when confidence is aligned with correctness. Fact-checking uncertainty is multi-source, arising from insufficient evidence, retrieval failure, ambiguous or outdated claims, cross-modal grounding errors, or flawed reasoning [11]. For LLMs, confidence estimation is further complicated by prompt-sensitive signals such as verbalized uncertainty and self-consistency [452], reinforcement learning from human feedback (RLHF)-induced overconfidence toward helpful and decisive responses [12,453], and verbose but potentially unfaithful CoT rationales [454].

A verifier is calibrated when its stated confidence matches empirical correctness [455]. Calibration can be evaluated using reliability diagrams [456], expected calibration error (ECE), maximum calibration error (MCE) [457], and the Brier score [458], which expose overconfident errors hidden by accuracy or F1. Post-hoc methods such as temperature scaling, Platt scaling, and isotonic regression can adjust confidence estimates without redesigning the fact-checking system [455,459,460]. Conformal prediction enables set-valued verdicts with target coverage, allowing ambiguous cases to trigger abstention, further retrieval, or human review [461].

Future work should treat calibration as a first-class objective. In the short term, benchmarks should report ECE, MCE, Brier score, reliability diagrams, risk–coverage curves, abstention behavior, and class-wise analyses across verdict categories. In the longer term, multimodal fact-checking requires pipeline-level calibration over final verdicts, retrieval reliability, evidence quality, cross-modal grounding, OCR reliability, and explanation faithfulness.

6.9. Human–AI Collaboration and Ethical Risks

Automated fact-checking should support rather than replace human judgment. LLMs and MLLMs can improve system scalability, but high-stakes, uncertain, or controversial claims still require human oversight. The key difficulty is over-reliance: automated verdicts

may propagate model errors and obscure responsibility when evidence is incomplete or uncertainty is high [462]. In the near term, uncertainty reporting and provenance logging are practical priorities. In the long term, accountable human–AI workflows should combine automated verification, expert review, and appeal mechanisms.

Ethical concerns in fact-checking mainly center on bias and governance. Biases in training data and retrieval corpora can affect system verdicts [463]. Censorship concerns arise when fact-checking results are used to label or remove content without providing users with supporting evidence or transparent reasoning [464]. In the near term, standardized bias evaluation and transparent evidence reporting are practical priorities. In the long term, systems require continuous auditing across languages, regions, and user groups.

7. Conclusions

First, we review LLM/MLLM-based fact-checking across task formulations, datasets, methods, evaluation practices, system architectures, and persistent challenges. We organize the literature around the full verification pipeline, including claim detection and extraction, evidence retrieval, verdict prediction, and justification production, rather than treating fact-checking as a standalone classification task. This pipeline-oriented perspective enables unified comparison of unimodal, multimodal, and LLM/MLLM-based approaches and shows how claim understanding, source retrieval, cross-modal grounding, verdict reasoning, and explanation faithfulness are interdependent. Second, derived from previous research, our synthesis shows that methods have evolved from feature-based classifiers and task-specific neural models toward retrieval-augmented, instruction-following, and multimodal reasoning systems. LLMs improve semantic understanding, query reformulation, evidence filtering, zero-shot prediction, and justification generation, while MLLMs extend these capabilities to image–text and video verification. Third, our empirical findings in Section 6.1.1 substantiate a fundamental tension in this evolution. As established in previous research, while LLMs and MLLMs enhance cross-modal understanding and verification flexibility, fine-grained visual–textual consistency and evidential sufficiency remain critical challenges. Directly demonstrated in our experiments, these systems introduce heavy computational burdens through repeated search calls and visual token expansion. Crucially, as directly verified in our benchmark evaluations, verification gains do not scale equally with system complexity, where on simpler claims, full multimodal retrieval yields diminishing accuracy returns while sharply increasing latency and token costs. Therefore, as a broader interpretation supported by prior studies, future systems should not simply pursue static benchmark dominance. Instead, the practical goal is building adaptive architectures that align reasoning depth with claim complexity, evidential uncertainty, and deployment budgets.

A critical view of the data and evaluation landscape reveals uneven and sometimes hard-to-interpret progress. Large English text benchmarks have driven much of the field, whereas multimodal, multilingual, temporal, and explanation-oriented resources remain limited. As a result, SOTA comparisons are often shaped by differences in label spaces, evidence settings, split stability, model versions, data access, and possible contamination by pretraining corpora. At the metric level, accuracy and macro-F1 also miss asymmetric error costs, calibration, evidence grounding, hallucination, explanation faithfulness, and human-centered explanation quality. Future benchmarking should therefore become more contamination-aware, evidence-grounded, and deployment-sensitive, especially as fact-checking increasingly depends on dynamic information sources and evolving model infrastructures.

Looking forward, the central bottleneck lies in bridging fluent generation with verifiable provenance. Future work should prioritize grounding-first architectures that connect

verdicts and explanations to retrievable, timestamped, and modality-specific sources. It should also develop dynamic benchmarks that test multilingual, multimodal, temporal, and adversarial robustness under realistic deployment constraints. Finally, fact-checking systems should be designed as human–AI workflows, combining automated verification with uncertainty reporting, provenance logging, expert review, appeal mechanisms, and continuous bias auditing. These directions can help move fact-checking from isolated prediction models toward reliable, interpretable, scalable, and socially accountable verification systems.

Author Contributions: Conceptualization, T.Z. and Z.Z.; methodology, T.Z. and Z.W.; software, Z.W. and X.J.; validation, T.Z., Z.W. and X.J.; formal analysis, T.Z.; investigation, T.Z., Z.W., X.J. and Q.C.; resources, Q.C., J.Z. and Y.L.; data curation, Z.W. and X.J.; writing—original draft preparation, T.Z.; writing—review and editing, T.Z., Z.W., X.J., Q.C., J.Z., Y.L., Z.Z. and J.L.; visualization, Z.W.; supervision, Z.Z. and J.L.; project administration, J.Z.; funding acquisition, Z.Z. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 62406199, 62471310, 62302311, 62302316, and 62306287), the Scientific Foundation for Youth Scholars of Shenzhen University (No. 868-000001032413), the Internal Fund of the National Engineering Laboratory for Big Data System Computing Technology (No. SZU-BDSC-IF2024-10), the Shenzhen Medical Research Fund (No. D2601008), the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011681), the Shenzhen Science and Technology Program (No. QNXMC20250701094312017), the Independent Scientific Research Program of the State Key Laboratory of Radio Frequency Heterogeneous Integration (No. 2025006), and the Intelligent Computing Center of Shenzhen University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The empirical case study used four publicly available benchmark datasets. AVeriTeC is available at https://github.com/MichSchli/AVeriTeC (accessed on 3 November 2026), FEVER is available at https://fever.ai/resources.html (accessed on 3 November 2026), MOCHEG is available at https://github.com/VT-NLP/Mocheg (accessed on 3 November 2026), and VERITE is available at https://github.com/stevejpapad/image-text-verification (accessed on 3 November 2026). For the datasets surveyed elsewhere in the manuscript, the corresponding dataset-release papers, official download or repository links, and availability status are compiled in the authors' companion repository at https://github.com/zhicong992-droid/LLMs-for-Multimodal-Fact-Checking (accessed on 3 November 2026). Datasets for which raw data are restricted or no stable public link was identified are explicitly marked in that repository. Access and reuse of all third-party datasets are subject to the licenses and terms specified by their original providers.

Acknowledgments: During the preparation of this manuscript, the authors used AI-assisted tools exclusively for English-language editing, including grammar, wording, spelling, and punctuation, and for formatting checks. These tools were not used for literature selection, study design, data collection, data analysis, result interpretation, image generation, or the formulation of scientific conclusions. The authors reviewed and edited all suggested changes and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
CoT	Chain-of-Thought
LLM	Large Language Model
MLLM	Multimodal Large Language Model

NLP	Natural Language Processing
OCR	Optical Character Recognition
PLM	Pre-trained Language Model
QA	Question Answering
RAG	Retrieval-Augmented Generation
RLHF	Reinforcement Learning from Human Feedback

Appendix A. Prompt Templates

Listing A1. Prompt template for verdict prediction.

```
You are a fact-checking assistant. Evaluate the claim using only the
provided evidence and the dataset-specific label definitions.

Claim:
{claim}

Evidence:
{evidence}

Label Definitions:
{label_definitions}

Select exactly one valid label and provide a concise, evidence-grounded
justification. Do not introduce external or unsupported information.

Return only the following JSON object:
{
  "label": "<valid label>",
  "reason": "<concise evidence-based justification>"
}
```

Appendix B. Additional Comparisons and Dataset Details

The following tables provide additional methodological comparisons and dataset details referenced in the main text.

Table A1. Comparison of traditional and LLM/MLLM-based methods across the four-stage fact-checking pipeline. Rob., Hal., Grd., Lat., and Dep. denote robustness, hallucination risk, evidence grounding, latency, and deployment readiness, respectively.

Dim.	Traditional Pipeline	LLM-Based	MLLM-Based
Claim Detection and Extraction			
Rob.	Medium: stable for explicit claims, weak for implicit.	Medium: flexible for noisy text but prompt-sensitive.	Medium: uses visual or audio cues but alignment-sensitive.
Hal.	Low: mainly classification or span extraction errors.	Medium: may omit, merge, or rewrite claim details.	High: textualization may introduce unsupported multimodal details.
Grd.	Medium: grounded mainly in textual spans.	Medium: generated claims require retrieval verification.	High: can link claims to regions, frames, or audio cues.
Lat.	Low: lightweight classifiers or extractors.	Medium: decoding and claim decomposition add cost.	High: multimodal encoding and preprocessing are costly.
Dep.	Medium: mature for text triage but weak for rich media.	Medium: useful for open-domain triage with validation.	Low: costly and harder to audit in practice.

Table A1. *Cont.*

Dim.	Traditional Pipeline	LLM-Based	MLLM-Based
Evidence Retrieval			
Rob.	High: strong with well-formed queries and indexed sources.	Medium: RAG broadens coverage but causes query drift.	Medium: cross-media retrieval is sensitive to noisy features.
Hal.	Low: retrieval returns existing evidence without generation.	Medium: rewritten queries may add unsupported content.	High: captions/OCR/descriptions may hallucinate cues.
Grd.	High: documents, URLs, or knowledge entries are explicit.	Medium: grounding depends on retrievers and citations.	High: can retrieve text, image, video, or audio evidence.
Lat.	Medium: web search or reverse search adds overhead.	High: iterative search and generation increase runtime.	High: multimodal indexing and search are expensive.
Dep.	High: mature, auditable, and easy to integrate.	Medium: practical with caching and fallback strategies.	Low: infrastructure and evaluation remain complex.
Verdict Prediction			
Rob.	Medium: strong in-domain, weaker multi-hop reasoning.	Medium: flexible reasoning but prompt- and evidence-sensitive.	Medium: handles cross-media, suffers alignment errors.
Hal.	Low: usually predicts labels rather than generated facts.	High: may produce unsupported conclusions.	High: may produce unsupported visual or multimodal rationales.
Grd.	Medium: evidence-conditioned but may rely on shallow cues.	Medium: improves with RAG, self-checking, or verifier models.	High: can link decisions to regions, frames, or segments.
Lat.	Low: efficient once evidence is available.	High: chain-of-thought, self-checking, or agents add cost.	High: multimodal reasoning and long contexts are costly.
Dep.	High: mature for fixed-label prediction.	Medium: requires uncertainty handling and human oversight.	Low: useful for rich media but less mature.
Justification Production			
Rob.	Medium: stable but often shallow rationales.	Medium: fluent explanations but faithfulness varies.	Medium: explains cross-modal cues.
Hal.	Low: usually reuses source spans or structured evidence.	High: free-form prose may add unsupported details.	High: may misinterpret objects, events, or relations.
Grd.	High: auditable through spans, evidence IDs, or graphs.	Medium: improved with quoted evidence and evidence IDs.	High: can be tied to regions, frames, or audio intervals.
Lat.	Low: efficient after evidence and labels are fixed.	Medium: generation and factuality checking add cost.	High: multimodal localization and generation are costly.
Dep.	Medium: suitable for expert audit but less natural for users.	Medium: user-friendly but requires grounding checks.	Low: promising but evaluation remains immature.

Table A2. Language coverage of multilingual datasets in the survey.

Dataset	Language
CLEF2023 [43]	3 languages: Arabic, English, and Spanish.
CheckThat21-T1A [101]	5 languages: Arabic, Bulgarian, English, Spanish, and Turkish.
IndiaWApp [107]	4 Indian languages. The original dataset is reported as multilingual Indian WhatsApp data; the exact language list should be taken from the official release.

Table A2. *Cont.*

Dataset	Language
MuMiN [113]	41 languages: Arabic, Bengali, Burmese, Croatian, Czech, Danish, Dutch, English, Filipino, French, German, Greek, Gujarati, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Kazakh, Korean, Macedonian, Malay, Malayalam, Marathi, Norwegian, Oriya, Panjabi, Persian, Polish, Portuguese, Russian, Serbian, Simplified Chinese, Slovak, Spanish, Tamil, Telugu, Thai, Traditional Chinese, Turkish, Urdu.
X-CLAIM [89]	6 languages: English, Bengali, Hindi, Malayalam, Tamil, and Telugu.
mCLAN [93]	20 languages: English, Arabic, Bengali, Czech, German, Greek, French, Hindi, Korean, Marathi, Indonesian, Dutch, Punjabi, Romanian, Spanish, Portuguese, Tamil, Telugu, Thai, and Urdu.
mMARCO [123]	14 languages: Arabic, Chinese, Dutch, English, French, German, Hindi, Indonesian, Italian, Japanese, Portuguese, Russian, Spanish, and Vietnamese.
MIRACL [124]	18 languages: Arabic, Bengali, Chinese, English, Spanish, Persian, Finnish, French, Hindi, Indonesian, Japanese, Korean, Russian, Swahili, Telugu, Thai, German, and Yoruba.
Mr. TyDi [125]	11 languages: Arabic, Bengali, English, Finnish, Indonesian, Japanese, Korean, Russian, Swahili, Telugu, and Thai.
WIT [127]	108 languages: Afrikaans, Aragonese, Arabic, Egyptian Arabic, Asturian, Azerbaijani (Latin), South Azerbaijani, Bashkir, Bavarian, Belarusian, Belarusian (Taraskievica), Bulgarian, Bengali, Breton, Bosnian, Catalan, Chechen, Cebuano, Central Kurdish (Sorani), Czech, Chuvash, Welsh, Danish, German, Greek, English, Esperanto, Spanish, Estonian, Basque, Persian, Finnish, Filipino, French, Western Frisian, Irish, Galician, Hindi, Croatian, Upper Sorbian, Haitian Creole, Hungarian, Armenian, Interlingua, Indonesian, Ido, Icelandic, Italian, Hebrew, Japanese, Javanese, Georgian, Kazakh, Kannada, Korean, Latin, Western Punjabi (Lahnda), Luxembourgish, Lombard, Lithuanian, Latvian, Malagasy, Macedonian, Malayalam, Mongolian, Marathi, Malay, Burmese, Min Nan Chinese, Low German, Nepali, Dutch, Norwegian Nynorsk, Norwegian, Navajo, Occitan, Punjabi, Polish, Portuguese, Quechua, Romanian, Russian, Scots, Sinhala, Slovak, Slovenian, Albanian, Serbian (Cyrillic), Serbian (Latin), Swedish, Swahili, Tamil, Telugu, Tajik, Thai, Turkish, Tatar, Ukrainian, Urdu, Uzbek, Venetian, Vietnamese, Volapük, Waray, Mingrelian, Cantonese, Chinese (Simplified), Chinese (Traditional, Taiwan).
Deepfake-Eval-2024 [171]	52 languages. The official release reports 52 languages across audio, video, and image data; the full language list is provided in the restricted release metadata.
X-FACT [185]	25 languages: Arabic, Azerbaijani, Bengali, German, Spanish, Persian, French, Gujarati, Hindi, Indonesian, Italian, Georgian, Marathi, Dutch, Norwegian, Punjabi, Polish, Portuguese, Romanian, Russian, Sinhala, Albanian, Serbian, Tamil, and Turkish.
MultiClaim [200]	39 languages for fact-checks: Arabic, Assamese, Azerbaijani, Bulgarian, Bengali, Catalan, Czech, Danish, German, Greek, English, Spanish, Farsi, Finnish, Tagalog, French, Serbo-Croatian, Hindi, Hungarian, Italian, Hebrew, Japanese, Khmer, Korean, Macedonian, Malayalam, Malay, Myanmar, Dutch, Norwegian, Polish, Portuguese, Romanian, Russian, Sinhala, Slovak, Albanian, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, and Chinese. The associated social-media posts cover 27 of these languages.

Table A3. General-purpose retrieval datasets. *Input* denotes the query modality, *Output* denotes the returned top-*k* candidate type, *Evidence* denotes the candidate corpus modality, and #Input denotes the number of queries, annotations, or input instances. T, I, V, and A denote text, image, video, and audio, respectively. Lang indicates the language of textual queries, captions, or annotations; "–" means not applicable or not reported.

Dataset	Input	Output	Evidence	Retrieval Source	#Input	Lang
MS COCO [465]	T	Top-k Img	I	Flickr	616,435	En
Flickr30k [466]	T	Top-k Img	I	Flickr	158,915	En
Multi30K [467]	T	Top-k Img	I	Flickr	155,070	En/De
MSR-VTT [468]	T	Top-k Vid	V	YouTube	200,000	En
ActivityNet [469]	T	Top-k Vid	V	YouTube	100,000	En
VATEX [470]	T	Top-k Vid	V	YouTube	825,000	En/Zh
AudioCaps [471]	T	Top-k Aud	A	AudioSet	45,513	En
Clotho [472]	T	Top-k Aud	A	Freesound	24,905	En
AudioSetCaps [473]	T	Top-k Aud	A	AudioSet/YouTube	6,117,099	En
FashionIQ [474]	T/I	Top-k Img	I	Web	60,272	En
CIRR [475]	T/I	Top-k Img	I	Web	36,554	En
CIRCO [476]	T/I	Top-k Img	I	Flickr	1020	En
ROxford/RPar [477]	I	Top-k Img	I	Flickr	70	–
Covers80 [478]	A	Top-k Aud	A	uspop2002 [479]	160	En
SHS100K [480]	A	Top-k Aud	A	SecondHandSongs	100,000	–
Da-TACOS [481]	A	Top-k Aud	A	SecondHandSongs	15,000	–
AVSET-10M [482]	A/V	Top-k Vid/Aud	A/V	AudioSet/YouTube	10,605,005	En

Table A4. Maturity assessment of major challenges in LLM-based fact-checking. "Emerging" denotes that a challenge has attracted initial research attention, but lacks mature methods, stable benchmarks, or widely accepted evaluation protocols.

Challenge	Section	Maturity	Near-Term Priorities	Long-Term Open Issues
Complexity trade-offs	Section 6.1	Medium	Reduce per-claim latency and cost through batching, caching, PEFT, and selective module invocation	Achieve scalable real-time fact-checking with stable latency, cost, and accuracy
Adversarial robustness	Section 6.2	Low	Improve input filtering, retrieval validation, prompt-injection defense, and red-teaming	Build robust defenses against multimodal attacks, evidence poisoning, and indirect prompt injection
Temporal drift	Section 6.3	Low	Add timestamps, refresh evidence, and evaluate time-sensitive claims	Develop dynamic benchmarks and reliable temporal reasoning over changing facts
Multilinguality	Section 6.4	Low	Expand multilingual datasets and improve cross-lingual evidence retrieval	Enable reliable low-resource and culture-aware fact-checking
Hallucination	Section 6.5	Medium	Improve RAG, attribution checking, hallucination detection, and uncertainty estimation	Achieve grounded multimodal verification with faithful explanations
Factual error correction	Section 6.6	Emerging	Develop evidence-guided editing and correction-quality metrics	Build faithful correction pipelines that avoid over-correction and new errors
Privacy and federated fact-checking	Section 6.7	Low	Adapt federated learning, differential privacy, and secure retrieval to fact-checking	Enable privacy-preserving evidence reasoning and fact-checking-specific benchmarks
Calibration and uncertainty	Section 6.8	Emerging	Add calibration metrics, abstention, and risk–coverage evaluation	Support pipeline-level uncertainty propagation and reliable abstention
Human–AI collaboration and ethics	Section 6.9	Low	Improve human review, provenance logging, uncertainty reporting, and bias evaluation	Establish accountable workflows with appeals and continuous governance audits

Table A5. Relationship between retrieval datasets and fact-checking. The "Relationship to Fact-Checking" column denotes whether a dataset is directly designed for fact-checking, indirectly useful for evidence retrieval, or used only as a general background benchmark. The "Importance/Rationale" column denotes why each dataset is included and how it relates to fact-checking.

Dataset	Relationship to Fact-Checking	Importance/Rationale
BEIR [121]	Indirect: Text Evidence Retrieval	Evaluates open-domain text retrieval, a key component of evidence-based fact-checking.
MS MARCO [122]	Indirect: Web Passage Retrieval	Provides web-scale passage retrieval data for training evidence retrievers.
mMARCO [123]	Indirect: Multilingual Passage Retrieval	Extends evidence retrieval evaluation to multilingual fact-checking scenarios.
MIRACL [124]	Indirect: Multilingual Wikipedia Retrieval	Supports multilingual retrieval from encyclopedic sources commonly used as evidence.
Mr. TyDi [125]	Indirect: Multilingual Dense Retrieval	Evaluates retrieval across diverse languages for multilingual evidence search.
DuReader [126]	Indirect: Chinese Web Retrieval	Supports Chinese-language web evidence retrieval.
BingCheck [106]	Direct: LLM Claim Fact-Checking	Directly targets evidence retrieval for verifying LLM-generated claims.
MS COCO [465]	General Background Only	Provides a general image-caption retrieval benchmark without fact-checking annotations.
Flickr30k [466]	General Background Only	Provides a general image-caption retrieval benchmark.
Multi30K [467]	General Background Only	Provides a multilingual image-caption benchmark not designed for fact-checking.
WIT [127]	Indirect: Wikipedia Image Retrieval	Connects images with Wikipedia text, useful for encyclopedic visual evidence search.
MSR-VTT [468]	General Background Only	Provides a general video-caption retrieval benchmark without fact-checking annotations.
ActivityNet [469]	General Background Only	Provides an activity video-caption dataset not designed for fact-checking.
VATEX [470]	General Background Only	Provides a bilingual video-caption benchmark without a direct fact-checking task.
AudioCaps [471]	General Background Only	Provides a general audio-caption retrieval dataset without fact-checking labels.
Clotho [472]	General Background Only	Provides a general audio-caption benchmark not designed for fact-checking.
AudioSetCaps [473]	General Background Only	Provides a large-scale audio-caption dataset without a fact-checking component.
FashionIQ [474]	General Background Only	Provides a fashion-domain retrieval dataset with limited relevance to fact-checking.
CIRR [475]	General Background Only	Provides a composed image retrieval benchmark without fact-checking annotations.
CIRCO [476]	General Background Only	Provides an open-domain image retrieval benchmark not designed for fact-checking.
TRECVID [128]	Indirect: Multimedia Video Retrieval	Evaluates video search capabilities relevant to locating video evidence.
ROxford/RPar [477]	General Background Only	Provides a landmark image retrieval benchmark without a claim-verification setting.
INRIA Holidays [129]	Indirect: Near-Duplicate Image Retrieval	Supports reused-image detection, relevant to out-of-context visual claims.
Stanford I2V [130]	Indirect: Image-To-Video Retrieval	Helps retrieve source videos from image queries.
NTU-VOI [131]	Indirect: Image-To-Video Instance Retrieval	Supports locating object instances in video evidence.
FIVR-200K [132]	Indirect: Incident Video Retrieval	Retrieves videos depicting the same real-world incident.
VCDB [133]	Indirect: Video Copy Detection	Detects reused or partially copied videos.
CC_WEB_VIDEO [134]	Indirect: Web Video Duplicate Retrieval	Supports tracing near-duplicate videos on the web.
EVVE [135]	Indirect: Event Video Retrieval	Retrieves videos associated with real-world events.
Covers80 [478]	General Background Only	Provides a cover-song retrieval dataset with limited relevance to fact-checking.
SHS100K [480]	General Background Only	Provides a cover-song retrieval benchmark with weak relevance to fact-checking.
Da-TACOS [481]	General Background Only	Provides a cover-song identification dataset not designed for fact-checking.
AVSET-10M [482]	General Background Only	Provides an audio-visual retrieval dataset without fact-checking annotations.

References

1. Vlachos, A.; Riedel, S. Fact Checking: Task Definition and Dataset Construction. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, MD, USA, 26 June 2014; pp. 18–22. [CrossRef]
2. Althabiti, S.; Alsalka, M.A.; Atwell, E. Generative AI for Explainable Automated Fact Checking on the FactEx: A New Benchmark Dataset. In *Disinformation in Open Online Media*; LNCS; Springer: Cham, Switzerland, 2023; Volume 14397, pp. 1–13. [CrossRef]
3. Hassan, N.; Arslan, F.; Li, C.; Tremayne, M. Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims by ClaimBuster. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 1803–1812. [CrossRef]
4. Shaar, S.; Babulkov, N.; Da San Martino, G.; Nakov, P. That Is a Known Lie: Detecting Previously Fact-Checked Claims. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 3607–3618. [CrossRef]
5. Vo, N.; Lee, K. Where Are the Facts? Searching for Fact-Checked Information to Alleviate the Spread of Fake News. *arXiv* **2020**, arXiv:2010.03159. [CrossRef]
6. Laban, P.; Hearst, M.A. newsLens: Building and Visualizing Long-Ranging News Stories. In Proceedings of the Events and Stories in the News Workshop, Vancouver, BC, Canada, 4 August 2017; pp. 1–9. [CrossRef]
7. Cohen, S.; Hamilton, J.T.; Turner, F. Computational Journalism. *Commun. ACM* **2011**, *54*, 66–71. [CrossRef]
8. Martín, A.; Huertas-Tato, J.; Huertas-García, Á.; Villar-Rodríguez, G.; Camacho, D. FacTeR-Check: Semi-Automated Fact-Checking Through Semantic Similarity and Natural Language Inference. *Knowl.-Based Syst.* **2022**, *251*, 109265. [CrossRef]
9. Soleimani, A.; Monz, C.; Worring, M. BERT for Evidence Retrieval and Claim Verification. In *Advances in Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 359–366. [CrossRef]
10. Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; Hajishirzi, H. UnifiedQA: Crossing Format Boundaries with a Single QA System. In *Proceedings of the Findings of EMNLP*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1896–1907. [CrossRef]
11. Guo, Z.; Schlichtkrull, M.S.; Vlachos, A. A Survey on Automated Fact-Checking. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 178–206. [CrossRef]
12. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 27730–27744.
13. Wang, G.; Harwood, K.; Chillrud, L.; Ananthram, A.; Subbiah, M.; McKeown, K.R. Check-COVID: Fact-Checking COVID-19 News Claims with Scientific Evidence. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 14114–14127. [CrossRef]
14. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837. [CrossRef]
15. Jain, S.; Wallace, B.C. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3543–3556. [CrossRef]
16. Abolghasemi, M.; Ganbold, O.; Rotaru, K. Humans vs. Large Language Models: Judgmental Forecasting in an Era of Advanced AI. *Int. J. Forecast.* **2025**, *41*, 631–648. [CrossRef]
17. Zhang, L.; Zhang, X.; Zhou, Z.; Zhang, X.; Yu, P.S.; Li, C. Knowledge-Aware Multimodal Pre-Training for Fake News Detection. *Inf. Fusion* **2025**, *114*, 102715. [CrossRef]
18. Metzger, M.J.; Flanagin, A.J.; Medders, R.B. Social and Heuristic Approaches to Credibility Evaluation Online. *J. Commun.* **2010**, *60*, 413–439. [CrossRef]
19. Petrou, N.; Christodoulou, C.; Anastasiou, A.; Pallis, G.; Dikaiakos, M.D. A Multiple Change-Point Detection Framework on Linguistic Characteristics of Real Versus Fake News Articles. *Sci. Rep.* **2023**, *13*, 6086. [CrossRef] [PubMed]
20. Cekinel, R.F.; Karagoz, P.; Çöltekin, Ç. Multimodal Fact-Checking with Vision Language Models: A Probing Classifier Based Solution with Embedding Strategies. In Proceedings of the 31st International Conference on Computational Linguistics, Dhabi, United Arab Emirates, 19–24 January 2025; pp. 4622–4633.
21. Yang, S.; Dai, Y.; Wang, G.; Zheng, X.; Xu, J.; Li, J.; Ying, Z.; Wang, W.; Ngai, E.C.H. RealFactBench: A Benchmark for Evaluating Large Language Models in Real-World Fact-Checking. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, 27–31 October, 2025*; Gurrin, C., Schoeffmann, K., Zhang, M., Rossetto, L., Rudinac, S., Dang-Nguyen, D., Cheng, W., Chen, P., Benois-Pineau, J., Eds.; ACM: New York, NY, USA, 2025; pp. 13435–13441. [CrossRef]
22. Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; Choi, Y. Defending Against Neural Fake News. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 9054–9065.

23. Lee, N.; Li, B.Z.; Wang, S.; Yih, W.t.; Ma, H.; Khabsa, M. Language Models as Fact Checkers? *arXiv* **2020**, arXiv:2006.04102. [CrossRef]
24. Saleh, A.O.; Karaoğlan, K.M.; Çakmak, M. A Comprehensive Survey on Automatic Detection of Fake News Using Natural Language Processing: Challenges and Limitations. In *Proceedings of the 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)*; IEEE: New York, NY, USA, 2024; pp. 1–7. [CrossRef]
25. Vykopal, I.; Pikuliak, M.; Ostermann, S.; Simko, M. Generative Large Language Models in Automated Fact-Checking: A Survey. *arXiv* **2024**, arXiv:2407.02351. [CrossRef]
26. Quelle, D.; Bovet, A. The Perils and Promises of Fact-Checking with Large Language Models. *Front. Artif. Intell.* **2024**, *7*, 1341697. [CrossRef] [PubMed]
27. Eldifrawi, I.; Wang, S.; Trabelsi, A. Automated Justification Production for Claim Veracity in Fact Checking: A Survey on Architectures and Approaches. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; ACL: Stroudsburg, PA, USA, 2024; pp. 6679–6692. [CrossRef]
28. Tufchi, S.; Yadav, A.; Ahmed, T. A Comprehensive Survey of Multimodal Fake News Detection Techniques: Advances, Challenges, and Opportunities. *Int. J. Multimed. Inf. Retr.* **2023**, *12*, 28. [CrossRef]
29. Akhtar, M.; Schlichtkrull, M.S.; Guo, Z.; Cocarascu, O.; Simperl, E.; Vlachos, A. Multimodal Automated Fact-Checking: A Survey. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023*; ACL: Stroudsburg, PA, USA, 2023; pp. 5430–5448. [CrossRef]
30. Alam, F.; Cresci, S.; Chakraborty, T.; Silvestri, F.; Dimitrov, D.; Da San Martino, G.; Shaar, S.; Firooz, H.; Nakov, P. A Survey on Multimodal Disinformation Detection. In *Proceedings of the 29th International Conference on Computational Linguistics*; ICCL: New York, NY, USA, 2022; pp. 6625–6643.
31. Huang, T.; Yi, J.; Yu, P.; Xu, X. Unmasking Digital Falsehoods: A Comparative Analysis of LLM-Based Misinformation Detection Strategies. *arXiv* **2025**, arXiv:2503.00724. [CrossRef]
32. Yang, S.; Yuan, X.; Gan, T.; Wu, Y. A Survey of Automatic Fact Verification Research. In *Proceedings of the 2024 7th World Conference on Computing and Communication Technologies (WCCCT)*; IEEE: New York, NY, USA, 2024; pp. 64–73. [CrossRef]
33. Geng, J.; Kementchedjhieva, Y.; Nakov, P.; Gurevych, I. Multimodal Large Language Models to Support Real-World Fact-Checking. *arXiv* **2024**, arXiv:2403.03627. [CrossRef]
34. Lin, L.; Gupta, N.; Zhang, Y.; Ren, H.; Liu, C.; Ding, F.; Wang, X.; Li, X.; Verdoliva, L.; Hu, S. Detecting Multimedia Generated by Large AI Models: A Survey. *arXiv* **2024**, arXiv:2402.00045. [CrossRef]
35. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; ACL: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [CrossRef]
36. Augenstein, I.; Lioma, C.; Wang, D.; Lima, L.C.; Hansen, C.; Hansen, C.; Simonsen, J.G. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; ACL: Stroudsburg, PA, USA, 2019; pp. 4684–4696. [CrossRef]
37. Hassan, N.; Zhang, G.; Arslan, F.; Caraballo, J.; Jimenez, D.; Gawsane, S.; Hasan, S.; Joseph, M.; Kulkarni, A.; Nayak, A.K.; et al. ClaimBuster: The First-Ever End-to-End Fact-Checking System. *Proc. VLDB Endow.* **2017**, *10*, 1945–1948. [CrossRef]
38. Zhang, F.; Liu, J.; Zhang, Q.; Sun, E.; Xie, J.; Zha, Z.J. ECENet: Explainable and Context-Enhanced Network for Multi-Modal Fact Verification. In *Proceedings of the 31st ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2023; pp. 1231–1240. [CrossRef]
39. Hassan, N.; Li, C.; Tremayne, M. Detecting Check-Worthy Factual Claims in Presidential Debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*; ACM: New York, NY, USA, 2015; pp. 1835–1838. [CrossRef]
40. Ullrich, H.; Mlynár, T.; Drchal, J. Claim Extraction for Fact-Checking: Data, Models, and Automated Metrics. *arXiv* **2025**, arXiv:2502.04955. [CrossRef]
41. Geng, J.; Tonglet, J.; Gurevych, I. M4FC: A Multimodal, Multilingual, Multicultural, Multitask Real-World Fact-Checking Dataset. *arXiv* **2025**, arXiv:2510.23508. [CrossRef]
42. Deng, Z.; Schlichtkrull, M.S.; Vlachos, A. Document-level Claim Extraction and Decontextualisation for Fact-Checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Ku, L., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 11943–11954. [CrossRef]
43. Barrón-Cedeño, A.; Alam, F.; Caselli, T.; Martino, G.D.S.; Elsayed, T.; Galassi, A.; Haouari, F.; Ruggeri, F.; Struß, J.M.; Nandi, R.N.; et al. The CLEF-2023 CheckThat! Lab: Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority. In *Advances in Information Retrieval*; LNCS; Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2023; Volume 13982, pp. 506–517. [CrossRef]
44. Song, Y.; Kim, Y.; Iyyer, M. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Proceedings of the Findings of EMNLP*; Al-Onaizan, Y., Bansal, M., Chen, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 9447–9474. [CrossRef]

45. Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; Luo, J. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*; Liu, Q., Lienhart, R., Wang, H., Chen, S.K., Boll, S., Chen, Y.P., Friedland, G., Li, J., Yan, S., Eds.; ACM: New York, NY, USA, 2017; pp. 795–816. [CrossRef]
46. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; ACM: New York, NY, USA, 2018; pp. 849–857. [CrossRef]
47. Xie, J.; Mao, W.; Bai, Z.; Zhang, D.J.; Wang, W.; Lin, K.Q.; Gu, Y.; Chen, Z.; Yang, Z.; Shou, M.Z. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation. In Proceedings of the Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.
48. Li, J.; Li, D.; Savarese, S.; Hoi, S.C.H. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the 40th International Conference on Machine Learning, Honolulu HA USA, 23–29 July 2023; Volume 202, pp. 19730–19742.
49. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
50. Henzinger, M.R.; Motwani, R.; Silverstein, C. Challenges in Web Search Engines. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003)*; ACM: New York, NY, USA, 2002; p. 1.
51. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. FEVER: A Large-scale Dataset for Fact Extraction and Verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*; ACL: Stroudsburg, PA, USA, 2018; pp. 809–819. [CrossRef]
52. Wanner, M.; Van Durme, B.; Dredze, M. Dndscore: Decontextualization and decomposition for factuality verification in long-form text generation. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 4–9 November 2025; pp. 23620–23637. [CrossRef]
53. Hu, Q.; Long, Q.; Wang, W. 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), Albuquerque, NM, USA, 29 April–4 May 2025; pp. 6313–6336. [CrossRef]
54. Chen, X.; Chen, X.; He, B.; Wen, T.; Sun, L. Analyze, Generate and Refine: Query Expansion with LLMs for Zero-Shot Open-Domain QA. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 12028–12041. [CrossRef]
55. Braun, T.; Rothermel, M.; Rohrbach, M.; Rohrbach, A. DEFAME: Dynamic Evidence-Based Fact-Checking with Multimodal Experts. *arXiv* **2024**, arXiv:2412.10510. [CrossRef]
56. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 8748–8763.
57. Lewis, P.S.H.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Volume 33, pp. 9459–9474.
58. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.
59. Ge, Z.; Wu, Y.; Chin, D.W.K.; Lee, R.K.W.; Cao, R. Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented llms. *arXiv* **2025**, arXiv:2505.17762. [CrossRef]
60. Wang, H.; Prasad, A.; Stengel-Eskin, E.; Bansal, M. Retrieval-augmented generation with conflicting evidence. *arXiv* **2025**, arXiv:2504.13079.
61. Nie, Y.; Chen, H.; Bansal, M. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 6859–6866. [CrossRef]
62. Tan, X.; Zou, B.; Aw, A.T. Evidence-Based Interpretable Open-Domain Fact-Checking with Large Language Models. *arXiv* **2023**, arXiv:2312.05834. [CrossRef]
63. Sahitaj, P.; Maab, I.; Yamagishi, J.; Kolanowski, J.; Möller, S.; Schmitt, V. Towards Automated Fact-Checking of Real-World Claims: Exploring Task Formulation and Assessment with LLMs. *arXiv* **2025**, arXiv:2502.08909. [CrossRef]
64. Liu, H.; Wang, W.; Li, H.; Li, H. TELLER: A Trustworthy Framework for Explainable, Generalizable and Controllable Fake News Detection. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 15556–15583. [CrossRef]
65. Zhao, X.; Wang, L.; Wang, Z.; Cheng, H.; Zhang, R.; Wong, K.F. 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), Turin, Italy, 20–25 May 2024; pp. 12564–12573. [CrossRef]

66. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 1–11. [CrossRef]
67. Aneja, S.; Bregler, C.; Nießner, M. COSMOS: Catching Out-of-Context Image Misuse Using Self-Supervised Learning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*; Williams, B., Chen, Y., Neville, J., Eds.; AAAI Press: Washington, DC, USA, 2023; pp. 14084–14092. [CrossRef]
68. Yoo, K.; Ahn, W.; Jang, J.; Kwak, N. Robust Multi-bit Natural Language Watermarking through Invariant Features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Rogers, A., Boyd-Graber, J.L., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 2092–2115. [CrossRef]
69. Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; Li, H. DIRE for Diffusion-Generated Image Detection. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 22388–22398. [CrossRef]
70. Ji, L.; Lin, Y.; Huang, Z.; Han, Y.; Xu, X.; Wu, J.; Wang, C.; Liu, Z. Distinguish Any Fake Videos: Unleashing the Power of Large-scale Data and Motion Features. *arXiv* **2024**, arXiv:2405.15343. [CrossRef]
71. Sun, C.; Jia, S.; Hou, S.; Lyu, S. AI-Synthesized Voice Detection Using Neural Vocoder Artifacts. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2023; pp. 904–912. [CrossRef]
72. Sun, Z.; Jiang, H.; Chen, H.; Cao, Y.; Qiu, X.; Wu, Z.; Jiang, Y.G. ForgerySleuth: Empowering Multimodal Large Language Models for Image Manipulation Detection. In *Advances in Neural Information Processing Systems 38*; Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Eds.; Curran Associates, Inc.: New York, NY, USA, 2025; Volume 38, pp. 153136–153175. [CrossRef]
73. Li, F.; Wu, J.; He, C.; Zhou, W. CMIE: Combining MLLM Insights with External Evidence for Explainable Out-of-Context Misinformation Detection. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025*; Che, W., Nabende, J., Shutova, E., Pilehvar, M.T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; Volume 2025, pp. 9342–9354. [CrossRef]
74. Yao, B.M.; Shah, A.; Sun, L.; Cho, J.H.; Huang, L. 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*; ACM: New York, NY, USA, 2023; pp. 2733–2743. [CrossRef]
75. Graves, L. *Understanding the Promise and Limits of Automated Fact-Checking*; Technical Report; Reuters Institute for the Study of Journalism, University of Oxford: Oxford, UK, 2018.
76. Kotonya, N.; Toni, F. Explainable Automated Fact-Checking: A Survey. In *Proceedings of the 28th International Conference on Computational Linguistics*; Scott, D., Bel, N., Zong, C., Eds.; International Committee on Computational Linguistics: New York, NY, USA, 2020; pp. 5430–5443. [CrossRef]
77. Popat, K.; Mukherjee, S.; Strötgen, J.; Weikum, G. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proceedings of the 26th International Conference on World Wide Web Companion*; Barrett, R., Cummings, R., Agichtein, E., Gabrilovich, E., Eds.; ACM: New York, NY, USA, 2017; pp. 1003–1012. [CrossRef]
78. Shang, L.; Kou, Z.; Zhang, Y.; Wang, D. A Duo-generative Approach to Explainable Multimodal COVID-19 Misinformation Detection. In *Proceedings of the ACM Web Conference 2022*; Laforest, F., Troncy, R., Simperl, E., Agarwal, D., Gionis, A., Herman, I., Médini, L., Eds.; ACM: New York, NY, USA, 2022; pp. 3623–3631. [CrossRef]
79. Wang, J.; Yang, H.; Wang, X.; Wang, W.; Li, Y. MAGNet: A multimodal knowledge-augmented graph network for early-stage misinformation detection. *Neurocomputing* **2025**, *658*, 131533. [CrossRef]
80. Xia, S.; Wang, X.; Liang, J.; Zhang, Y.; Zhou, W.; Deng, J.; Yu, F.; Xiao, Y. Ground Every Sentence: Improving Retrieval-Augmented LLMs with Interleaved Reference-Claim Generation. In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025*; Chiruzzo, L., Ritter, A., Wang, L., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; Volume 2025, pp. 969–988. [CrossRef]
81. Atanasova, P.; Simonsen, J.G.; Lioma, C.; Augenstein, I. Generating Fact Checking Explanations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 5–10 July 2020; pp. 7352–7364. [CrossRef]
82. Chen, T.C.; Tang, C.W.; Thomas, C. MetaSumPerceiver: Multimodal Multi-Document Evidence Summarization for Fact-Checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; ACL: Stroudsburg, PA, USA, 2024; pp. 8742–8757. [CrossRef]
83. Niu, K.; Xu, D.; Yang, B.; Liu, W.; Wang, Z. Pioneering Explainable Video Fact-Checking with a New Dataset and Multi-Role Multimodal Model Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2025; pp. 28276–28283. [CrossRef]
84. Cheema, G.S.; Hakimov, S.; Sittar, A.; Müller-Budack, E.; Otto, C.; Ewerth, R. MM-Claims: A Dataset for Multimodal Claim Detection in Social Media. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022, Washington, DC, USA, 10–15 July 2022; pp. 962–979. [CrossRef]

85. Giedemann, P.; von Däniken, P.; Deriu, J.; Rodrigo, Á.; Peñas, A.; Cieliebak, M. ViClaim: A Multilingual Multilabel Dataset for Automatic Claim Detection in Videos. *arXiv* **2025**, arXiv:2504.12882. [CrossRef]
86. Maros, A.; Almeida, J.M.; Vasconcelos, M. A Study of Misinformation in Audio Messages Shared in WhatsApp Groups. In *Disinformation in Open Online Media*; LNCS; Bright, J., Giachanou, A., Spaiser, V., Spezzano, F., George, A., Pavliuc, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2021; Volume 12887, pp. 85–100. [CrossRef]
87. Shang, L.; Kou, Z.; Zhang, Y.; Wang, D. A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. In *Proceedings of the 2021 IEEE International Conference on Big Data (Big Data)*; Chen, Y., Ludwig, H., Tu, Y., Fayyad, U.M., Zhu, X., Hu, X., Byna, S., Liu, X., Zhang, J., Pan, S., et al., Eds.; IEEE: New York, NY, USA, 2021; pp. 899–908. [CrossRef]
88. Sundriyal, M.; Kulkarni, A.; Pulastya, V.; Akhtar, M.S.; Chakraborty, T. Empowering the Fact-checkers! Automatic Identification of Claim Spans on Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 7701–7715. [CrossRef]
89. Mittal, S.; Sundriyal, M.; Nakov, P. Lost in Translation, Found in Spans: Identifying Claims in Multilingual Social Media. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 3887–3902. [CrossRef]
90. Poddar, S.; Paul, B.; Basu, M.; Ghosh, S. ICPR 2024 Competition on Multilingual Claim-Span Identification. In *Pattern Recognition. Competitions*; LNCS; Antonacopoulos, A., Chaudhuri, S., Chellappa, R., Liu, C., Bhattacharya, S., Pal, U., Eds.; Springer: Berlin/Heidelberg, Germany, 2024; Volume 15334, pp. 134–144. [CrossRef]
91. Roy, R.; Pal, P.; Das, D.; Ghosh, S.; Paul, B. Enhancing Textual Understanding: Automated Claim Span Identification in English, Hindi, Bengali, and CodeMix. In *Proceedings of Recent Advances in Natural Language Processing, Varna, Bulgaria, 8–10 September 2025*; Angelova, G., Kunilovskaya, M., Escribe, M., Mitkov, R., Eds.; ACL: Stroudsburg, PA, USA, 2025; Volume 2025, pp. 1030–1035. [CrossRef]
92. Sundriyal, M.; Chakraborty, T.; Nakov, P. From Chaos to Clarity: Claim Normalization to Empower Fact-Checking. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 6594–6609. [CrossRef]
93. Sundriyal, M.; Chakraborty, T.; Nakov, P.I. Overview of the CLEF-2025 CheckThat! lab task 2 on claim normalization. In Proceedings of the Experimental IR Meets Multilinguality, Multimodality, and Interaction: 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, 9–12 September 2025.
94. Mitra, T.; Gilbert, E. CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*; Cha, M., Mascolo, C., Sandvig, C., Eds.; AAAI Press: Washington, DC, USA, 2015; Volume 9, pp. 258–267. [CrossRef]
95. Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B.J.; Wong, K.; Cha, M. Detecting Rumors from Microblogs with Recurrent Neural Networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*; Kambhampati, S., Ed.; IJCAI/AAAI Press: Washington, DC, USA, 2016; pp. 3818–3824.
96. Zubiaga, A.; Liakata, M.; Procter, R.; Wong Sak Hoi, G.; Tolmie, P. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE* **2016**, *11*, e0150989. [CrossRef] [PubMed]
97. Gorrell, G.; Aker, A.; Bontcheva, K.; Derczynski, L.; Kochkina, E.; Liakata, M.; Zubiaga, A. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*; May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., Mohammad, S.M., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 845–854. [CrossRef]
98. Lillie, A.E.; Middelboe, E.R.; Derczynski, L. Joint Rumour Stance and Veracity Prediction. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*; Linköping University Electronic Press: Linköping, Sweden, 2019; pp. 208–221.
99. Volkova, S.; Shaffer, K.; Jang, J.Y.; Hodas, N.O. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*; Barzilay, R., Kan, M., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 647–653. [CrossRef]
100. Barrón-Cedeño, A.; Elsayed, T.; Nakov, P.; Martino, G.D.S.; Hasanain, M.; Suwaileh, R.; Haouari, F.; Babulkov, N.; Hamdan, B.; Nikolov, A.; et al. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*; LNCS; Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12260, pp. 215–236. [CrossRef]
101. Nakov, P.; Martino, G.D.S.; Elsayed, T.; Barrón-Cedeño, A.; Míguez, R.; Shaar, S.; Alam, F.; Haouari, F.; Hasanain, M.; Babulkov, N.; et al. The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In *Advances in Information Retrieval*; LNCS; Hiemstra, D., Moens, M., Mothe, J., Perego, R., Potthast, M., Sebastiani, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2021; Volume 12657, pp. 639–649. [CrossRef]

102. Gencheva, P.; Nakov, P.; Màrquez, L.; Barrón-Cedeño, A.; Koychev, I. A Context-Aware Approach for Detecting Worth-Checking Claims in Political Debates. In *Proceedings of Recent Advances in Natural Language Processing*; Mitkov, R., Angelova, G., Eds.; INCOMA Ltd.: Sofia, Bulgaria, 2017; pp. 267–276. [CrossRef] [PubMed]
103. Atanasova, P.; Màrquez, L.; Barrón-Cedeño, A.; Elsayed, T.; Suwaileh, R.; Zaghouani, W.; Kyuchukov, S.; Martino, G.D.S.; Nakov, P. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. Task 1: Check-Worthiness. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*; Cappellato, L., Ferro, N., Nie, J., Soulier, L., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; Volume 2125, pp. 1–12. [CrossRef]
104. Redi, M.; Fetahu, B.; Morgan, J.T.; Taraborelli, D. Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability. In *Proceedings of the World Wide Web Conference*; Liu, L., White, R.W., Mantrach, A., Silvestri, F., McAuley, J.J., Baeza-Yates, R., Zia, L., Eds.; ACM: New York, NY, USA, 2019; pp. 1567–1578. [CrossRef]
105. Konstantinovskiy, L.; Price, O.; Babakar, M.; Zubiaga, A. Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digit. Threat. Res. Pract.* **2021**, *2*, 14. [CrossRef]
106. Li, M.; Peng, B.; Galley, M.; Gao, J.; Zhang, Z. Self-Checker: Plug-and-Play Modules for Fact-Checking with Large Language Models. In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024*; Duh, K., Gómez-Adorno, H., Bethard, S., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 163–181. [CrossRef]
107. Garimella, K.; Eckles, D. Images and Misinformation in Political Groups: Evidence from WhatsApp in India. *arXiv* **2020**, arXiv:2005.09784. [CrossRef]
108. Qu, J.; Li, L.H.; Zhao, J.; Dev, S.; Chang, K. DisinfoMeme: A Multimodal Dataset for Detecting Meme Intentionally Spreading Out Disinformation. *arXiv* **2022**, arXiv:2205.12617. [CrossRef]
109. Liu, F.; Yacoob, Y.; Shrivastava, A. COVID-VTS: Fact Extraction and Verification on Short Video Platforms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*; Vlachos, A., Augenstein, I., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 178–188. [CrossRef]
110. Qi, P.; Bu, Y.; Cao, J.; Ji, W.; Shui, R.; Xiao, J.; Wang, D.; Chua, T.S. FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2023; pp. 14444–14452. [CrossRef]
111. Zhang, D.Y.; Shang, L.; Geng, B.; Lai, S.; Li, K.; Zhu, H.; Al Amin, M.T.; Wang, D. FauxBuster: A Content-free Fauxtography Detector Using Social Media Comments. In *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data)*; Abe, N., Liu, H., Pu, C., Hu, X., Ahmed, N.K., Qiao, M., Song, Y., Kossmann, D., Liu, B., Lee, K., et al., Eds.; IEEE: New York, NY, USA, 2018; pp. 891–900. [CrossRef]
112. Kou, Z.; Zhang, D.Y.; Shang, L.; Wang, D. ExFaux: A Weakly Supervised Approach to Explainable Fauxtography Detection. In *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*; Wu, X., Jermaine, C., Xiong, L., Hu, X., Kotevska, O., Lu, S., Xu, W., Aluru, S., Zhai, C., Al-Masri, E., et al., Eds.; IEEE: New York, NY, USA, 2020; pp. 631–636. [CrossRef]
113. Nielsen, D.S.; McConville, R. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 3141–3153.
114. Zong, R.; Zhang, Y.; Shang, L.; Wang, D. ContrastFaux: Sparse Semi-supervised Fauxtography Detection on the Web using Multi-view Contrastive Learning. In *Proceedings of the ACM Web Conference 2023*; Ding, Y., Tang, J., Sequeda, J.F., Aroyo, L., Castillo, C., Houben, G., Eds.; ACM: New York, NY, USA, 2023; pp. 3994–4003. [CrossRef]
115. Hu, X.; Guo, Z.; Chen, J.; Wen, L.; Yu, P.S. MR2: A Benchmark for Multimodal Retrieval-Augmented Rumor Detection in Social Media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Chen, H., Duh, W.E., Huang, H., Kato, M.P., Mothe, J., Poblete, B., Eds.; ACM: New York, NY, USA, 2023; pp. 2901–2912. [CrossRef]
116. Prabhakar, T.; Gupta, A.; Nadig, K.; George, D. Check Mate: Prioritizing User Generated Multi-Media Content for Fact-Checking. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*; AAAI Press: Washington, DC, USA, 2021; Volume 15, pp. 1025–1033. [CrossRef]
117. Resende, G.; Melo, P.F.; Sousa, H.; Messias, J.; Vasconcelos, M.; Almeida, J.M.; Benevenuto, F. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *Proceedings of the World Wide Web Conference*; Liu, L., White, R.W., Mantrach, A., Silvestri, F., McAuley, J.J., Baeza-Yates, R., Zia, L., Eds.; ACM: New York, NY, USA, 2019; pp. 818–828. [CrossRef]
118. Chen, J.; Sriram, A.; Choi, E.; Durrett, G. Generating Literal and Implied Subquestions to Fact-check Complex Claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 3495–3516. [CrossRef]
119. Hanselowski, A.; Zhang, H.; Li, Z.; Sorokin, D.; Schiller, B.; Schulz, C.; Gurevych, I. UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 103–108. [CrossRef]
120. Voorhees, E.M. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 355–370. [CrossRef] [PubMed]

121. Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; Gurevych, I. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv* **2021**, arXiv:2104.08663. [CrossRef]
122. Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; Deng, L. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv* **2016**, arXiv:1611.09268.
123. Bonifacio, L.; Jeronymo, V.; Abonizio, H.Q.; Campiotti, I.; Fadaee, M.; Lotufo, R.; Nogueira, R. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv* **2021**, arXiv:2108.13897.
124. Zhang, X.; Thakur, N.; Ogundepo, O.; Kamalloo, E.; Alfonso-Hermelo, D.; Li, X.; Liu, Q.; Rezagholizadeh, M.; Lin, J. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 1114–1131. [CrossRef]
125. Zhang, X.; Ma, X.; Shi, P.; Lin, J. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 127–137. [CrossRef]
126. Qiu, Y.; Li, H.; Qu, Y.; Chen, Y.; She, Q.; Liu, J.; Wu, H.; Wang, H. DuReader-Retrieval: A Large-scale Chinese Benchmark for Passage Retrieval from Web Search Engine. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 5326–5338. [CrossRef]
127. Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; Najork, M. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T., Eds.; ACM: New York, NY, USA, 2021; pp. 2443–2449. [CrossRef]
128. Awad, G.; Kraaij, W.; Over, P.; Satoh, S. Instance search retrospective with focus on TRECVID. *Int. J. Multimed. Inf. Retr.* **2017**, *6*, 1–29. [CrossRef] [PubMed]
129. Jégou, H.; Douze, M.; Schmid, C. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In *Computer Vision—ECCV 2008*; LNCS; Forsyth, D.A., Torr, P.H.S., Zisserman, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5302, pp. 304–317. [CrossRef]
130. Araújo, A.; Chaves, J.; Chen, D.M.; Angst, R.; Girod, B. Stanford I2V: A news video dataset for query-by-image experiments. In *Proceedings of the 6th ACM Multimedia Systems Conference*; Ooi, W.T., Feng, W., Liu, F., Eds.; ACM: New York, NY, USA, 2015; pp. 237–242. [CrossRef]
131. Meng, J.; Yuan, J.; Yang, J.; Wang, G.; Tan, Y. Object Instance Search in Videos via Spatio-Temporal Trajectory Discovery. *IEEE Trans. Multim.* **2016**, *18*, 116–127. [CrossRef]
132. Kordopatis-Zilos, G.; Papadopoulos, S.; Patras, I.; Kompatsiaris, I. FIVR: Fine-Grained Incident Video Retrieval. *IEEE Trans. Multim.* **2019**, *21*, 2638–2652. [CrossRef]
133. Jiang, Y.; Jiang, Y.; Wang, J. VCDB: A Large-Scale Database for Partial Copy Detection in Videos. In *Computer Vision—ECCV 2014*; LNCS; Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8692, pp. 357–371. [CrossRef]
134. Shang, L.; Yang, L.; Wang, F.; Chan, K.; Hua, X. Real-time large scale near-duplicate web video retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*; Bimbo, A.D., Chang, S., Smeulders, A.W.M., Eds.; ACM: New York, NY, USA, 2010; pp. 531–540. [CrossRef]
135. Revaud, J.; Douze, M.; Schmid, C.; Jégou, H. Event Retrieval in Large Video Collections with Circulant Temporal Encoding. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE Computer Society: New York, NY, USA, 2013; pp. 2459–2466. [CrossRef]
136. Thorne, J.; Vlachos, A. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*; Bender, E.M., Derczynski, L., Isabelle, P., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 3346–3359.
137. Aneja, S.; Bregler, C.; Nießner, M. Catching Out-of-Context Misinformation with Self-supervised Learning. *arXiv* **2021**, arXiv:2101.06278. [CrossRef]
138. Jaiswal, A.; Sabir, E.; Abd-Almageed, W.; Natarajan, P. Multimedia Semantic Integrity Assessment Using Joint Embedding Of Images And Text. In *Proceedings of the 25th ACM International Conference on Multimedia*; Liu, Q., Lienhart, R., Wang, H., Chen, S.K., Boll, S., Chen, Y.P., Friedland, G., Li, J., Yan, S., Eds.; ACM: New York, NY, USA, 2017; pp. 1465–1471. [CrossRef]
139. Luo, G.; Darrell, T.; Rohrbach, A. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv* **2021**, arXiv:2104.05893. [CrossRef]
140. Boididou, C.; Papadopoulos, S.; Kompatsiaris, Y.; Schifferes, S.; Newman, N. Challenges of computational verification in social multimedia. In *Proceedings of the 23rd International Conference on World Wide Web*; Chung, C., Broder, A.Z., Shim, K., Suel, T., Eds.; ACM: New York, NY, USA, 2014; pp. 743–748. [CrossRef]
141. Tralic, D.; Zupancic, I.; Grgic, S.; Grgic, M. CoMoFoD–New database for copy-move forgery detection. In *Proceedings of the Proceedings ELMAR-2013*; IEEE: New York, NY, USA, 2013; pp. 49–54.

142. Yan, H.; Hong, Y.; Zhan, J.; Ji, Y.; Lan, J.; Zhu, H.; Wang, W.; Zhang, J. COCO-Inpaint: A Benchmark for Image Inpainting Detection and Manipulation Localization. *arXiv* **2025**, arXiv:2504.18361. [CrossRef]
143. Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; Wang, Y. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. In *Advances in Neural Information Processing Systems 36*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2023; pp. 77771–77782. [CrossRef]
144. Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Canton-Ferrer, C. The Deepfake Detection Challenge (DFDC) Preview Dataset. *arXiv* **2019**, arXiv:1910.08854. [CrossRef]
145. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. *arXiv* **2018**, arXiv:1803.09179. [CrossRef]
146. Reimao, R.; Tzerpos, V. FoR: A Dataset for Synthetic Speech Detection. In *Proceedings of the 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*; Burileanu, C., Teodorescu, H., Eds.; IEEE: New York, NY, USA, 2019; pp. 1–10. [CrossRef]
147. Müller-Budack, E.; Theiner, J.; Diering, S.; Idahl, M.; Ewerth, R. Multimodal Analytics for Real-world News using Measures of Cross-modal Entity Consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*; Gurrin, C., Jónsson, B.Þ., Kando, N., Schöffmann, K., Chen, Y.P., O'Connor, N.E., Eds.; ACM: New York, NY, USA, 2020; pp. 16–25. [CrossRef]
148. Pham, K.; Nguyen-Nhat, M.; Dinh, A.; Le, Q.; Nguyen, M.; Tran, A.; Tran, M.; Dang-Nguyen, D. Ookpik- A Collection of Out-of-Context Image-Caption Pairs. In *MultiMedia Modeling*; LNCS; Rudinac, S., Hanjalic, A., Liem, C.C.S., Worring, M., Jónsson, B.Þ., Liu, B., Yamakata, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2024; Volume 14557, pp. 132–144. [CrossRef]
149. Fung, Y.R.; Thomas, C.; Reddy, R.G.; Polisetty, S.; Ji, H.; Chang, S.; McKeown, K.R.; Bansal, M.; Sil, A. InfoSurgeon: Cross-Media Fine-grained Information Consistency Checking for Fake News Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: New York, NY, USA, 2021; pp. 1683–1698. [CrossRef]
150. Tan, R.; Plummer, B.A.; Saenko, K. Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Webber, B., Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: New York, NY, USA, 2020; pp. 2081–2106. [CrossRef]
151. Sabir, E.; AbdAlmageed, W.; Wu, Y.; Natarajan, P. Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2018; pp. 1337–1345. [CrossRef]
152. Biamby, G.; Luo, G.; Darrell, T.; Rohrbach, A. Twitter-COMMs: Detecting Climate, COVID, and Military Multimodal Misinformation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies; Carpuat, M., de Marneffe, M., Ruíz, I.V.M., Eds.; Association for Computational Linguistics: New York, NY, USA, 2022; pp. 1530–1549. [CrossRef]
153. Xu, Q.; Du, H.; Chen, H.; Liu, B.; Yu, X. MMOOC: A Multimodal Misinformation Dataset for Out-of-Context News Analysis. In *Information Security and Privacy*; LNCS; Zhu, T., Li, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2024; Volume 14897, pp. 444–459. [CrossRef]
154. Huh, M.; Liu, A.; Owens, A.; Efros, A.A. Fighting Fake News: Image Splice Detection via Learned Self-Consistency. In *Computer Vision—ECCV 2018*; LNCS; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11215, pp. 106–124. [CrossRef]
155. Gupta, A.; Lamba, H.; Kumaraguru, P.; Joshi, A. Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. In *Proceedings of the 22nd International Conference on World Wide Web*; Carr, L., Laender, A.H.F., Lóscio, B.F., King, I., Fontoura, M., Vrandecic, D., Aroyo, L., de Oliveira, J.P.M., Lima, F., Wilde, E., Eds.; International World Wide Web Conferences Steering Committee/ACM: New York, NY, USA, 2013; pp. 729–736. [CrossRef]
156. Heller, S.; Rossetto, L.; Schuldt, H. The PS-Battles Dataset—An Image Collection for Image Manipulation Detection. *arXiv* **2018**, arXiv:1804.04866. [CrossRef]
157. Shao, R.; Wu, T.; Liu, Z. Detecting and Grounding Multi-Modal Media Manipulation. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 6904–6913. [CrossRef]
158. Huang, Z.; Hu, J.; Li, X.; He, Y.; Zhao, X.; Peng, B.; Wu, B.; Huang, X.; Cheng, G. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 28831–28841. [CrossRef]
159. Al-Sanjary, O.I.; Ahmed, A.A.; Sulong, G. Development of a video tampering dataset for forensic investigation. *Forensic Sci. Int.* **2016**, *266*, 565–572. [CrossRef] [PubMed]
160. Guera, D.; Delp, E.J. Deepfake Video Detection Using Recurrent Neural Networks. In *Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*; IEEE: New York, NY, USA, 2018; pp. 1–6. [CrossRef]

161. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 3204–3213. [CrossRef]
162. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 2886–2895. [CrossRef]
163. Narayan, K.; Agarwal, H.; Thakral, K.; Mittal, S.; Vatsa, M.; Singh, R. DF-Platter: Multi-Face Heterogeneous Deepfake Dataset. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 9039–9048. [CrossRef]
164. Kwon, P.; You, J.; Nam, G.; Park, S.; Chae, G. Kodf: A large-scale korean deepfake detection dataset. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10744–10753. [CrossRef]
165. Wu, Z.; Kinnunen, T.; Evans, N.W.D.; Yamagishi, J.; Hanilçi, C.; Sahidullah, M.; Sizov, A. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015), Dresden, Germany, 6–10 September 2015; pp. 2037–2041. [CrossRef]
166. Wang, R.; Juefei-Xu, F.; Huang, Y.; Guo, Q.; Xie, X.; Ma, L.; Liu, Y. DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices. In *Proceedings of the 28th ACM International Conference on Multimedia*; Chen, C.W., Cucchiara, R., Hua, X., Qi, G., Ricci, E., Zhang, Z., Zimmermann, R., Eds.; ACM: New York, NY, USA, 2020; pp. 1207–1216. [CrossRef]
167. Zhao, B.; Kang, Z.; He, Y.; Qu, X.; Peng, J.; Xiao, J.; Wang, J. Generalized Audio Deepfake Detection Using Frame-level Latent Information Entropy. In *Proceedings of the 2025 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: New York, NY, USA, 2025; pp. 1–6. [CrossRef]
168. Lavrentyeva, G.; Novoselov, S.; Volkova, M.; Matveev, Y.; De Marsico, M. Phonespoof: A New Dataset for Spoofing Attack Detection in Telephone Channel. In *Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2019; pp. 2572–2576. [CrossRef]
169. Yi, J.; Bai, Y.; Tao, J.; Ma, H.; Tian, Z.; Wang, C.; Wang, T.; Fu, R. Half-Truth: A Partially Fake Audio Detection Dataset. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021)*; Hermansky, H., Cernocký, H., Burget, L., Lamel, L., Scharenborg, O., Motlícek, P., Eds.; ISCA: Grenoble, France, 2021; pp. 1654–1658. [CrossRef]
170. Wang, X.; Delgado, H.; Tak, H.; Jung, J.; Shim, H.; Todisco, M.; Kukanov, I.; Liu, X.; Sahidullah, M.; Kinnunen, T.; et al. ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale. *arXiv* **2024**, arXiv:2408.08739. [CrossRef]
171. Chandra, N.A.; Murtfeldt, R.; Qiu, L.; Karmakar, A.; Lee, H.; Tanumihardja, E.; Farhat, K.; Caffee, B.; Paik, S.; Lee, C.; et al. Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. *arXiv* **2025**, arXiv:2503.02857. [CrossRef]
172. Khalid, H.; Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*; Vanschoren, J., Yeung, S., Eds.; NeurIPS: San Diego, CA, USA, 2021; Volume 34, pp. 1018–1030.
173. Cai, Z.; Ghosh, S.; Adatia, A.P.; Hayat, M.; Dhall, A.; Gedeon, T.; Stefanov, K. AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset. In *Proceedings of the 32nd ACM International Conference on Multimedia*; Cai, J., Kankanhalli, M.S., Prabhakaran, B., Boll, S., Subramanian, R., Zheng, L., Singh, V.K., César, P., Xie, L., Xu, D., Eds.; New York, NY, USA, 2024; pp. 7414–7423. [CrossRef]
174. Singhal, S.; Shah, R.R.; Kumaraguru, P. FactDrill: A Data Repository of Fact-Checked Social Media Content to Study Fake News Incidents in India. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media*; Budak, C., Cha, M., Quercia, D., Eds.; AAAI Press: Washington, DC, USA, 2022; Volume 16, pp. 1322–1331. [CrossRef]
175. Wang, W.Y. "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)*; ACL: Stroudsburg, PA, USA, 2017; pp. 422–426. [CrossRef]
176. Papadopoulos, S.I.; Koutlis, C.; Papadopoulos, S.; Petrantonakis, P.C. VERITE: A Robust Benchmark for Multimodal Misinformation Detection Accounting for Unimodal Bias. *Int. J. Multimed. Inf. Retr.* **2024**, *13*, 4. [CrossRef]
177. Mishra, S.; Suryavardan, S.; Bhaskar, A.; Chopra, P.; Reganti, A.N.; Patwa, P.; Das, A.; Chakraborty, T.; Sheth, A.P.; Ekbal, A. FACTIFY: A Multi-Modal Fact Verification Dataset. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, BC, Canada, 30 July–4 August 2022; Volume 3199, pp. 1–11.
178. Suryavardan, S.; Mishra, S.; Patwa, P.; Chakraborty, M.; Rani, A.; Reganti, A.N.; Chadha, A.; Das, A.; Sheth, A.P.; Chinnakotla, M.; et al. Factify 2: A Multimodal Fake News and Satire News Dataset. In Proceedings of the 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, Co-Located with AAAI 2023, Washington, DC, USA, 14 February 2023; Volume 3439, pp. 11–20.

179. Upravitelev, M.; Solopova, V.; Sahitaj, P.; Sahitaj, A.; Jakob, C.; Möller, S.; Schmitt, V. Take It All: Ensemble Retrieval for Multimodal Evidence Aggregation. In *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*; ACL: Stroudsburg, PA, USA, 2026; pp. 91–107. [CrossRef]
180. Bu, Y.; Sheng, Q.; Cao, J.; Qi, P.; Wang, D.; Li, J. FakingRecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process. In *Proceedings of the 32nd ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2024; pp. 1351–1360. [CrossRef]
181. Aly, R.; Guo, Z.; Schlichtkrull, M.; Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A. The FEVEROUS Shared Task: Fact Extraction and Verification over Unstructured and Structured Information. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*; ACL: Stroudsburg, PA, USA, 2021; pp. 1–14. [CrossRef]
182. Schuster, T.; Fisch, A.; Barzilay, R. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; ACL: Stroudsburg, PA, USA, 2021; pp. 624–643. [CrossRef]
183. Jiang, Y.; Bordia, S.; Zhong, Z.; Dognin, C.; Singh, M.K.; Bansal, M. HoVer: A Dataset for Many-Hop Fact Extraction and Claim Verification. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020*; ACL: Stroudsburg, PA, USA, 2020; pp. 3441–3460. [CrossRef]
184. Schlichtkrull, M.S.; Guo, Z.; Vlachos, A. AVeriTeC: A Dataset for Real-World Claim Verification with Evidence from the Web. In *Proceedings of the Advances in Neural Information Processing Systems 36*; NeurIPS: San Diego, CA, USA, 2023; Volume 36, pp. 49301–49321. [CrossRef]
185. Gupta, A.; Srikumar, V. X-Fact: A New Benchmark Dataset for Multilingual Fact Checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*; ACL: Stroudsburg, PA, USA, 2021; pp. 675–682. [CrossRef]
186. Xu, C.; Kechadi, M.T. An Enhanced Fake News Detection System with Fuzzy Deep Learning. *IEEE Access* **2024**, *12*, 88006–88021. [CrossRef]
187. Lin, Y.J.; Lin, C.Y.; Yeh, C.J.; Li, Y.T.; Hu, Y.Y.; Hsu, C.H.; Lee, M.F.; Kao, H.Y. CFEVER: A Chinese Fact Extraction and Verification Dataset. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 18626–18634. [CrossRef]
188. Onoe, Y.; Zhang, M.J.Q.; Choi, E.; Durrett, G. CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*; NeurIPS: San Diego, CA, USA, 2021; Volume 34, pp. 15166–15178.
189. Park, J.; Min, S.; Kang, J.; Zettlemoyer, L.; Hajishirzi, H. FaVIQ: Fact Verification from Information-Seeking Questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; ACL: Stroudsburg, PA, USA, 2022; pp. 5154–5166. [CrossRef]
190. Nørregaard, J.; Derczynski, L. DanFEVER: Claim Verification Dataset for Danish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*; ACL: Stroudsburg, PA, USA, 2021; pp. 422–428.
191. Mohr, I.; Wührl, A.; Klinger, R. CoVERT: A Corpus of Fact-Checked Biomedical COVID-19 Tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*; ELRA: Paris, France, 2022; pp. 244–257. [CrossRef]
192. Zheng, J.; Baheti, A.; Naous, T.; Xu, W.; Ritter, A. Stanceosaurus: Classifying Stance Towards Multicultural Misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; ACL: Stroudsburg, PA, USA, 2022; pp. 2132–2151. [CrossRef]
193. Wadden, D.; Lin, S.; Lo, K.; Wang, L.L.; van Zuylen, M.; Cohan, A.; Hajishirzi, H. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; ACL: Stroudsburg, PA, USA, 2020; pp. 7534–7550. [CrossRef]
194. Chen, W.; Wang, H.; Chen, J.; Zhang, Y.; Wang, H.; Li, S.; Zhou, X.; Wang, W.Y. TabFact: A Large-Scale Dataset for Table-based Fact Verification. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 26–30 April 2020.
195. Khan, K.; Wang, R.; Poupart, P. WatClaimCheck: A New Dataset for Claim Entailment and Inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; ACL: Stroudsburg, PA, USA, 2022; pp. 1293–1304. [CrossRef]
196. Sathe, A.; Ather, S.; Le, T.M.; Perry, N.; Park, J. Automated Fact-Checking of Claims from Wikipedia. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), Marseille, France, 11–16 May 2020; pp. 6874–6882.
197. Hu, X.; Guo, Z.; Wu, G.; Liu, A.; Wen, L.; Yu, P.S. CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 3362–3376. [CrossRef]
198. Li, Z.; Zhang, H.; Zhang, J. A Revisit of Fake News Dataset with Augmented Fact-Checking by ChatGPT. *arXiv* **2023**, arXiv:2312.11870. [CrossRef]
199. Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data* **2020**, *8*, 171–188. [CrossRef] [PubMed]

200. Pikuliak, M.; Srba, I.; Móro, R.; Hromadka, T.; Smolen, T.; Melisek, M.; Vykopal, I.; Simko, J.; Podrouzek, J.; Bieliková, M. Multilingual Previously Fact-Checked Claim Retrieval. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 16477–16500. [CrossRef]
201. Rayar, F.; Delalandre, M.; Le, V.H. A Large-Scale TV Video and Metadata Database for French Political Content Analysis and Fact-Checking. In *Proceedings of the CBMI 2022: International Conference on Content-Based Multimedia Indexing*; ACM: New York, NY, USA, 2022; pp. 181–185. [CrossRef]
202. Zhang, F.; Li, D.; Zhang, Q.; Chen, J.; Liu, G.; Lin, J.; Yan, J.; Liu, J.; Zha, Z.J. Fact-R1: Towards Explainable Video Misinformation Detection with Deep Reasoning. *arXiv* **2025**, arXiv:2505.16836. [CrossRef]
203. Chun, C.; Terrisse, L.; Zhang, D.C.; Lee, D. MAD: A Benchmark for Multi-Turn Audio Dialogue Fact-Checking. *arXiv* **2025**, arXiv:2508.12186. [CrossRef]
204. Chun, C.; Zhang, D.C.; Lee, D. Context-Aware Multimodal Claim Verification in Spoken Dialogues. *arXiv* **2026**, arXiv:2606.11420. [CrossRef]
205. Tan, X.; Zou, B.; Aw, A.T. Improving Explainable Fact-Checking with Claim-Evidence Correlations. In *Proceedings of the 31st International Conference on Computational Linguistics*; ACL: Stroudsburg, PA, USA, 2025; pp. 1600–1612.
206. Aslam, N.; Khan, I.U.; Alotaibi, F.S.; Aldaej, L.A.; Aldubaikil, A.K. Fake Detect: A Deep Learning Ensemble Model for Fake News Detection. *Complexity* **2021**, *2021*, 5557784. [CrossRef]
207. Varnosfaderani, S.D.; Kruengkrai, C.; Yahyapour, R.; Yamagishi, J. Bridging Textual and Tabular Worlds for Fact Verification: A Lightweight, Attention-Based Model. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 20–25 May 2024; pp. 2515–2519. [CrossRef]
208. Zheng, L.; Li, C.; Zhang, L.; Jia, H.; Wang, S.; Liu, Z.; Zhang, X. MRR-FV: Unlocking Complex Fact Verification with Multi-Hop Retrieval and Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2025; pp. 26066–26074. [CrossRef]
209. Yoon, Y.; Jung, J.; Yoon, S.; Park, K. HerO at AVeriTeC: The Herd of Open Large Language Models for Verifying Real-World Claims. *arXiv* **2024**, arXiv:2410.12377. [CrossRef]
210. Singhal, A.; Law, T.; Kassner, C.; Gupta, A.; Duan, E.; Damle, A.; Li, R.L. Multilingual Fact-Checking Using LLMs. In *Proceedings of the Third Workshop on NLP for Positive Impact*; ACL: Stroudsburg, PA, USA, 2024; pp. 13–31. [CrossRef]
211. Wang, L.; Zhang, P.; Lu, X.; Zhang, L.; Yan, C.; Zhang, C. QaDialMoE: Question-Answering Dialogue Based Fact Verification with Mixture of Experts. In *Proceedings of the Findings of EMNLP*; ACL: Stroudsburg, PA, USA, 2022; pp. 3146–3159. [CrossRef]
212. Liu, H.; Soroush, A.; Nestor, J.G.; Park, E.; Idnay, B.; Fang, Y.; Pan, J.; Liao, S.; Bernard, M.; Peng, Y.; et al. Retrieval Augmented Scientific Claim Verification. *JAMIA Open* **2024**, *7*, ooae021. [CrossRef] [PubMed]
213. Kosprdic, M.; Ljajic, A.; Medvecki, D.; Basaragin, B.; Milosevic, N. Scientific Claim Verification with Fine-Tuned NLI Models. In Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Porto, Portugal, 17–19 November 2024; Volume 3, pp. 15–25. [CrossRef]
214. Hussain, A.S. ARTEMIS-DA: An Advanced Reasoning and Transformation Engine for Multi-Step Insight Synthesis in Data Analytics. *arXiv* **2025**, arXiv:2412.14146. [CrossRef]
215. Cong, M.; Liu, L.; Han, J.; Zhu, B.; Yuan, X.; Peng, T. DRIVE: An Adjustable Parallel Architecture Based on Evidence Awareness for Fake News Detection. *Expert Syst. Appl.* **2025**, *266*, 126043. [CrossRef]
216. Al-obaidi, S.A.; Çağlıkantar, T. Automated Fake News Detection System. *Iraqi J. Comput. Sci. Math.* **2024**, *5*, 14–24. [CrossRef]
217. Dhankar, A.; Zaïane, O.; Bolduc, F. UofA-Truth at Factify 2022: A Simple Approach to Multi-Modal Fact-Checking. In Proceedings of the Workshop on Multi-Modal Fake News and Hate-Speech Detection (DE-FACTIFY 2022), Virtual, 27 February 2022; Volume 3199, pp. 58–64.
218. Suryavardan, S.; Mishra, S.; Chakraborty, M.; Patwa, P.; Rani, A.; Chadha, A.; Reganti, A.N.; Das, A.; Sheth, A.P.; Chinnakotla, M.; et al. Findings of Factify 2: Multimodal Fake News Detection. In Proceedings of the De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, Co-Located with AAAI 2023 (DE-FACTIFY 2023), Washington, DC, USA, 13–14 February 2023; Volume 3439, pp. 1–10.
219. Fontana, N.; Corso, F.; Zuccolotto, E.; Pierri, F. Evaluating open-source Large Language Models for automated fact-checking. *arXiv* **2025**, arXiv:2503.05565. [CrossRef]
220. DeYoung, J.; Jain, S.; Rajani, N.F.; Lehman, E.; Xiong, C.; Socher, R.; Wallace, B.C. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; ACL: Stroudsburg, PA, USA, 2020; pp. 4443–4458. [CrossRef]
221. Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; Jain, A.K. On the Detection of Digital Face Manipulation. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: Piscataway, NJ, USA, 2020; pp. 5780–5789. [CrossRef]
222. Bonettini, N.; Cannas, E.D.; Mandelli, S.; Bondi, L.; Bestagini, P.; Tubaro, S. Video face manipulation detection through ensemble of cnns. In *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 5012–5019. [CrossRef]

223. Li, M.; Zhang, X.P. Interpretable temporal class activation representation for audio spoofing detection. *arXiv* **2024**, arXiv:2406.08825. [CrossRef]
224. Zeng, F.; Gao, W. Justilm: Few-Shot Justification Generation for Explainable Fact-Checking of Real-World Claims. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 334–354. [CrossRef]
225. Alhindi, T.; Petridis, S.; Muresan, S. Where is Your Evidence: Improving Fact-checking by Justification Modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*; Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., Mittal, A., Eds.; ACL: Brussels, Belgium, 2018; pp. 85–90. [CrossRef]
226. Russo, D.; Tekiroğlu, S.S.; Guerini, M. Benchmarking the generation of fact checking explanations. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 1250–1264. [CrossRef]
227. Kotonya, N.; Toni, F. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; ACL: Stroudsburg, PA, USA, 2020; pp. 7740–7759. [CrossRef]
228. Ma, H.; Xu, W.; Wei, Y.; Chen, L.; Wang, L.; Liu, Q.; Wu, S.; Wang, L. EX-FEVER: A Dataset for Multi-hop Explainable Fact Verification. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*; ACL: Stroudsburg, PA, USA, 2024; pp. 6378–6393. [CrossRef]
229. Stammbach, D.; Ash, E. e-fever: Explanations and summaries for automated fact checking. In *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*; Hacks Hackers: New York, NY, USA, 2020; pp. 32–43.
230. Kao, W.Y.; Yen, A.Z. How We Refute Claims: Automatic Fact-Checking Through Flaw Identification and Explanation. In *Proceedings of the WWW '24: Companion Proceedings of the ACM Web Conference 2024*; ACM: New York, NY, USA, 2024; pp. 758–761. [CrossRef]
231. Akhtar, M.; Subedi, N.; Gupta, V.; Tahmasebi, S.; Cocarascu, O.; Simperl, E. ChartCheck: Explainable Fact-Checking over Real-World Chart Images. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*; ACL: Stroudsburg, PA, USA, 2024; pp. 13876–13893. [CrossRef]
232. Xu, D.; Fan, S.; Cheng, H.; Kankanhalli, M. Multimodal Fact-Checking: An Agent-based Approach. *arXiv* **2025**, arXiv:2512.22933v1.
233. Xu, W.; Xiang, D.; Ding, T.; Lu, W. MMM-Fact: A Multimodal, Multi-Domain Fact-Checking Dataset with Multi-Level Retrieval Difficulty. *arXiv* **2025**, arXiv:2510.25120. [CrossRef]
234. Elkan, C. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01), Seattle, WA, USA, 4–10 August2001*; IJCAI: San Francisco, CA, USA, 2001; pp. 973–978.
235. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In *Proceedings of the ICML '06: Proceedings of the 23rd International Conference on Machine Learning*; ICML: New York, NY, USA, 2006; pp. 233–240. [CrossRef]
236. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
237. Geng, J.; Cai, F.; Wang, Y.; Koeppl, H.; Nakov, P.; Gurevych, I. A Survey of Confidence Estimation and Calibration in Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*; Duh, K., Gomez, H., Bethard, S., Eds.; ACL: Stroudsburg, PA, USA, 2024; pp. 6577–6595. [CrossRef]
238. Golovneva, O.; Chen, M.; Poff, S.; Corredor, M.; Zettlemoyer, L.; Fazel-Zarandi, M.; Celikyilmaz, A. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. *arXiv* **2023**, arXiv:2212.07919.
239. Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.t.; Koh, P.; Iyyer, M.; Zettlemoyer, L.; Hajishirzi, H. 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*; Bouamor, H., Pino, J., Bali, K., Eds.; ACL: Stroudsburg, PA, USA, 2023; pp. 12076–12100. [CrossRef]
240. Jacovi, A.; Goldberg, Y. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; ACL: Stroudsburg, PA, USA, 2020; pp. 4198–4205. [CrossRef]
241. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 248. [CrossRef]
242. Deng, C.; Zhao, Y.; Tang, X.; Gerstein, M.; Cohan, A. Investigating Data Contamination in Modern Benchmarks for Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*; Duh, K., Gomez, H., Bethard, S., Eds.; ACL: Stroudsburg, PA, USA, 2024; pp. 8706–8719. [CrossRef]
243. Stepanova, N.; Ross, B. Temporal Generalizability in Multimodal Misinformation Detection. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*; Hupkes, D., Dankers, V., Batsuren, K., Sinha, K., Kazemnejad, A., Christodoulopoulos, C., Cotterell, R., Bruni, E., Eds.; ACL: Stroudsburg, PA, USA, 2023; pp. 76–88. [CrossRef]
244. Rothermel, M.; Kornmann, M.; Rohrbach, M.; Rohrbach, A. VeriTaS: The First Dynamic Benchmark for Multimodal Automated Fact-Checking. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; ACL: Stroudsburg, PA, USA, 2026; pp. 42066–42100. [CrossRef]

245. Tsoneva, Y.; Feig, P.C.; Li, J.; Solopova, V.; Foroutan, N.; Hilbert, A.; Schmitt, V. Selective Multimodal Retrieval for Automated Verification of Image–Text Claims. In *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*; ACL: Stroudsburg, PA, USA, 2026; pp. 127–135. [CrossRef]
246. Das, B.C.; Amini, M.H.; Wu, Y. Security and Privacy Challenges of Large Language Models: A Survey. *ACM Comput. Surv.* **2025**, *57*, 152. [CrossRef]
247. Majer, L.; Snajder, J. Claim Check-Worthiness Detection: How Well Do LLMs Grasp Annotation Guidelines? *arXiv* **2024**, arXiv:2404.12174. [CrossRef]
248. Li, Y.; Panchendrarajan, R.; Zubiaga, A. FactFinders at CheckThat! 2024: Refining Check-worthy Statement Detection with LLMs through Data Pruning. In *Proceedings of the CLEF Working Notes*; Faggioli, G., Ferro, N., Galuscáková, P., de Herrera, A.G.S., Eds.; CEUR-WS.org: Aachen, Germany, 2024; Volume 3740, pp. 520–537.
249. Qian, S.; Wang, J.; Hu, J.; Fang, Q.; Xu, C. Hierarchical Multi-modal Contextual Attention Network for Fake News Detection. In *Proceedings of the SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T., Eds.; ACM: New York, NY, USA, 2021; pp. 153–162. [CrossRef]
250. Liu, H.; Wang, W.; Li, H. Interpretable Multimodal Misinformation Detection with Logic Reasoning. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023*; Rogers, A., Boyd-Graber, J.L., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 9781–9796. [CrossRef]
251. Qi, P.; Zhao, Y.; Shen, Y.; Ji, W.; Cao, J.; Chua, T. Two Heads Are Better Than One: Improving Fake News Video Detection by Correlating with Neighbors. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023*; Rogers, A., Boyd-Graber, J.L., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 11947–11959. [CrossRef]
252. Lee, M.; Pavlovic, V. Private-Shared Disentangled Multimodal VAE for Learning of Latent Representations. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2021; pp. 1692–1700. [CrossRef]
253. Lei, F.; Liu, X.; Jiang, J.; Liao, L.; Cai, J.; Zhao, H. Graph convolutional networks with higher-order pooling for semisupervised node classification. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e5695. [CrossRef]
254. Li, J.; Bin, Y.; Zou, J.; Wei, J.; Wang, G.; Yang, Y. Cross-Modal Consistency Learning with Fine-Grained Fusion Network for Multimodal Fake News Detection. In *MMAsia '23: Proceedings of the 5th ACM International Conference on Multimedia in Asia*; ACM: New York, NY, USA, 2023; pp. 1–7. [CrossRef]
255. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*; Su, J., Duh, K., Carreras, X., Eds.; ACM: New York, NY, USA, 2016; pp. 457–468. [CrossRef]
256. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked Cross Attention for Image-Text Matching. In *Proceedings of the Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 212–228. [CrossRef]
257. Wang, J.; Jiang, H.; Liu, Y.; Ma, C.; Zhang, X.; Pan, Y.; Liu, M.; Gu, P.; Xia, S.; Li, W.; et al. A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks. *arXiv* **2024**, arXiv:2408.01319. [CrossRef]
258. Team, C. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *arXiv* **2024**, arXiv:2405.09818. [CrossRef]
259. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Laptev, I.; Sivic, J.; Zisserman, A. Flamingo: A Visual Language Model for Few-Shot Learning. In *Proceedings of the NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing System*; Curran Associates Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 23716–23736.
260. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv* **2024**, arXiv:2409.12191. [CrossRef]
261. Alvar, S.R.; Singh, G.; Akbari, M.; Zhang, Y. DivPrune: Diversity-based Visual Token Pruning for Large Multimodal Models. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2025; pp. 9392–9401. [CrossRef]
262. Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; Li, C. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv* **2024**, arXiv:2407.07895. [CrossRef]
263. van der Meer, M.; Korshunov, P.; Marcel, S.; van der Plas, L. HintsOfTruth: A Multimodal Checkworthiness Detection Dataset with Real and Synthetic Claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Che, W., Nabende, J., Shutova, E., Pilehvar, M.T., Eds.; Association for Computational Linguistics: New York, NY, USA, 2025; pp. 31274–31291. [CrossRef]
264. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
265. Iashin, V.; Rahtu, E. Multi-modal Dense Video Captioning. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2020; pp. 4117–4126. [CrossRef]

266. Avila, A.R.; Rezagholizadeh, M.; Xing, C. Multimodal Audio-textual Architecture for Robust Spoken Language Understanding. *arXiv* **2023**, arXiv:2306.06819. [CrossRef]
267. Yang, S.; Yu, Z.; Ying, Z.; Dai, Y.; Wang, G.; Lan, J.; Xu, J.; Li, J.; Ngai, E.C.H. RAMA: Retrieval-Augmented Multi-Agent Framework for Misinformation Detection in Multimodal Fact-Checking. *arXiv* **2025**, arXiv:2507.09174. [CrossRef]
268. Teo, J.; Cao, R.; Deng, Z.; Ding, Z.; Schlichtkrull, M.S.; Vlachos, A. Multimodal Claim Extraction for Fact-Checking. In *Proceedings of the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2026; pp. 289–304. [CrossRef]
269. Kakizaki, K.; Matsunaga, Y.; Furukawa, R. MAFT: Multimodal Automated Fact-Checking via Textualization. In *Proceedings of the AAAI'25/IAAI'25/EAAI'25: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*; Walsh, T., Shah, J., Kolter, Z., Eds.; AAAI Press: Washington, DC, USA, 2025; pp. 29646–29648. [CrossRef]
270. Vatndal, H.; Setty, V. ShortCheck: Checkworthiness Detection of Multilingual Short-Form Videos. *arXiv* **2025**, arXiv:2509.20467. [CrossRef]
271. Barone, M.; Romano, A.; Riccio, G.; Postiglione, M.; Moscato, V. Combating Biomedical Misinformation through Multi-modal Claim Detection and Evidence-based Verification. In *Proceedings of the SIGIR '25: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Ferro, N., Maistro, M., Pasi, G., Alonso, O., Trotman, A., Verberne, S., Eds.; ACM: Stroudsburg, PA, USA, 2025; pp. 4025–4029. [CrossRef]
272. Qi, P.; Cao, J.; Yang, T.; Guo, J.; Li, J. Exploiting multi-domain visual information for fake news detection. In *Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM)*; IEEE: Piscataway, NJ, USA, 2019; pp. 518–527. [CrossRef]
273. Sheng, Q.; Cao, J.; Zhang, X.; Li, X.; Zhong, L. Article Reranking by Memory-Enhanced Key Sentence Matching for Detecting Previously Fact-Checked Claims. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 5468–5481. [CrossRef]
274. Abdelnabi, S.; Hasan, R.; Fritz, M. Open-Domain, Content-Based, Multi-Modal Fact-Checking of Out-of-Context Images via Online Resources. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 14940–14949. [CrossRef]
275. Ullrich, H.; Drchal, J. AIC CTUAVerImaTeC: Dual-retriever RAG for image-text fact checking. In *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*; Association for Computational Linguistics: New York, NY, USA, 2026; pp. 136–142. [CrossRef]
276. Samarinas, C.; Hsu, W.; Lee, M. Improving Evidence Retrieval for Automated Explainable Fact-Checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*; Sil, A., Lin, X.V., Eds.; Association for Computational Linguistics: New York, NY, USA, 2021; pp. 84–91. [CrossRef]
277. Zheng, L.; Li, C.; Zhang, X.; Shang, Y.; Huang, F.; Jia, H. Evidence Retrieval is almost All You Need for Fact Verification. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*; Ku, L., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: New York, NY, USA, 2024; pp. 9274–9281. [CrossRef]
278. Xie, Z.; Xing, R.; Wang, Y.; Geng, J.; Iqbal, H.; Sahnan, D.; Gurevych, I.; Nakov, P. FIRE: Fact-checking with iterative retrieval and verification. In *Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2025*; Association for Computational Linguistics: New York, NY, USA, 2025; pp. 2901–2914. [CrossRef]
279. Liu, H.; Zhao, Y.; Cohan, A.; Zhao, C. SUCEA: Reasoning-intensive retrieval for adversarial fact-checking through claim decomposition and editing. *arXiv* **2025**, arXiv:2506.04583.
280. Wei, J.; Yang, C.; Song, X.; Lu, Y.; Hu, N.; Huang, J.; Tran, D.; Peng, D.; Liu, R.; Huang, D.; et al. Long-Form Factuality in Large Language Models. In *Proceedings of the Advances in Neural Information Processing Systems 37*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2024; Volume 37, pp. 28682–28731. [CrossRef]
281. Tariq, A.; Kementchedjhieva, Y. REVEAL: Retrieval-Enhanced Verification for Multimodal Fact-Checking. In *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*; Association for Computational Linguistics: New York, NY, USA, 2026; pp. 108–113. [CrossRef]
282. Rasiwasia, N.; Pereira, J.C.; Coviello, E.; Doyle, G.; Lanckriet, G.R.G.; Levy, R.; Vasconcelos, N. A new approach to cross-modal multimedia retrieval. In *MM '10: Proceedings of the 18th ACM International Conference on Multimedia*; Bimbo, A.D., Chang, S., Smeulders, A.W.M., Eds.; ACM: New York, NY, USA, 2010; pp. 251–260. [CrossRef]
283. Tran, D.; Bourdev, L.D.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2015; pp. 4489–4497. [CrossRef]
284. Kordopatis-Zilos, G.; Papadopoulos, S.; Patras, I.; Kompatsiaris, Y. ViSiL: Fine-Grained Spatio-Temporal Video Similarity Learning. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2019; pp. 6350–6359. [CrossRef]

285. Chechik, G.; Ie, E.; Rehn, M.; Bengio, S.; Lyon, D. Large-scale content-based audio retrieval from text queries. In *MIR '08: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*; Lew, M.S., Bimbo, A.D., Bakker, E.M., Eds.; ACM: New York, NY, USA, 2008; pp. 105–112. [CrossRef]
286. Xin, Y.; Yang, D.; Zou, Y. Improving Text-Audio Retrieval by Text-Aware Attention Pooling and Prior Matrix Revised Loss. In Proceedings of the *ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5. [CrossRef]
287. Araújo, A.; Girod, B. Large-Scale Video Retrieval Using Image Queries. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1406–1420. [CrossRef]
288. Xu, R.; Niu, L.; Zhang, J.; Zhang, L. A Proposal-Based Approach for Activity Image-to-Video Retrieval. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020*; AAAI: Palo Alto, CA, USA, 2020; pp. 12524–12531. [CrossRef]
289. Wang, T.; Li, F.; Zhu, L.; Li, J.; Zhang, Z.; Shen, H.T. Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions. *Proc. IEEE* **2024**, *112*, 1716–1754. [CrossRef]
290. Hao, X.; Zhang, W.; Wu, D.; Zhu, F.; Li, B. Listen and Look: Multi-Modal Aggregation and Co-Attention Network for Video-Audio Retrieval. In *Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [CrossRef]
291. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Guo, Q.; Wang, M.; et al. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2023**, arXiv:2312.10997. [CrossRef]
292. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Webber, B., Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6769–6781. [CrossRef]
293. Yan, S.; Gu, J.; Zhu, Y.; Ling, Z. Corrective Retrieval Augmented Generation. *arXiv* **2024**, arXiv:2401.15884. [CrossRef]
294. Gao, L.; Ma, X.; Lin, J.; Callan, J. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; ACL: Stroudsburg, PA, USA, 2023; pp. 1762–1777. [CrossRef]
295. Kangur, U.; Agrawal, K.; Chakraborty, R.; Sharma, R. Evigenerate: Generative Evidence in Automated Fact-Checking. In *Proceedings of the AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLM)*; AAAI: Washington, DC, USA, 2025.
296. Sarthi, P.; Abdullah, S.; Tuli, A.; Khanna, S.; Goldie, A.; Manning, C.D. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In Proceedings of the ICLR 2024, Vienna, Austria, 7–11 May 2024.
297. Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2024**, arXiv:2404.16130. [CrossRef]
298. Emonet, V.; Bolleman, J.T.; Duvaud, S.; de Farias, T.M.; Sima, A.C. LLM-based SPARQL Query Generation from Natural Language over Federated Knowledge Graphs. *arXiv* **2024**, arXiv:2410.06062. [CrossRef]
299. Li, D.; Li, F.; Song, B.; Tang, L.; Zhou, W. Imrrf: Integrating multi-source retrieval and redundancy filtering for llm-based fake news detection. 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: Stroudsburg, PA, USA, 2025; pp. 9127–9142. [CrossRef]
300. Kangur, U.; Agrawal, K.; Singh, Y.; Sabir, A.; Sharma, R. MultiReflect: Multimodal Self-Reflective RAG-based Automated Fact-Checking. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 1–17. [CrossRef]
301. Wang, S.; Lin, H.; Luo, Z.; Ye, Z.; Chen, G.; Ma, J. MFC-Bench: Benchmarking Multimodal Fact-Checking with Large Vision-Language Models. *arXiv* **2024**, arXiv:2406.11288. [CrossRef]
302. Zhang, K.; Kong, C.; Liu, H.; Ding, B.; Jiang, X.; Li, H. Propose and Rectify: A Forensics-Driven MLLM Framework for Image Manipulation Localization. *arXiv* **2025**. [CrossRef]
303. Zhao, X.; Li, L.; Wang, Y. Distillation-Resistant Watermarking for Model Protection in NLP. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022*; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 5044–5055. [CrossRef]
304. Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C.D.; Finn, C. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning*; Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., Eds.; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 24950–24962.
305. Fernandez, P.; Chaffin, A.; Tit, K.; Chappelier, V.; Furon, T. Three Bricks to Consolidate Watermarks for Large Language Models. In *Proceedings of the 2023 IEEE International Workshop on Information Forensics and Security (WIFS)*; IEEE: New York, NY, USA, 2023; pp. 1–6. [CrossRef]

306. Zhu, B.; Yuan, L.; Cui, G.; Chen, Y.; Fu, C.; He, B.; Deng, Y.; Liu, Z.; Sun, M.; Gu, M. Beat LLMs at Their Own Game: Zero-Shot LLM-Generated Text Detection via Querying ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 7470–7483. [CrossRef]
307. Wu, J.; Zhan, R.; Wong, D.F.; Yang, S.; Liu, X.; Chao, L.S.; Zhang, M. Who Wrote This? The Key to Zero-Shot LLM-Generated Text Detection Is GECScore. In *Proceedings of the 31st International Conference on Computational Linguistics*; Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 10275–10292.
308. Hans, A.; Schwarzschild, A.; Cherepanova, V.; Kazemi, H.; Saha, A.; Goldblum, M.; Geiping, J.; Goldstein, T. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. In *Proceedings of the 41st International Conference on Machine Learning*; Salakhutdinov, R., Kolter, Z., Heller, K.A., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F., Eds.; PMLR: Cambridge, MA, USA, 2024; Volume 235, pp. 17519–17537.
309. Min, R.; Li, S.; Chen, H.; Cheng, M. A Watermark-Conditioned Diffusion Model for IP Protection. In *Computer Vision–ECCV 2024*; LNCS; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2024; Volume 15127, pp. 104–120. [CrossRef]
310. Borji, A. Qualitative failures of image generation models and their application in detecting deepfakes. *Image Vis. Comput.* **2023**, *137*, 104771. [CrossRef]
311. Zhong, N.; Xu, Y.; Li, S.; Qian, Z.; Zhang, X. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv* **2023**, arXiv:2311.12397.
312. Fernandez, P.; Couairon, G.; Jégou, H.; Douze, M.; Furon, T. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 22409–22420. [CrossRef]
313. Farid, H. Perspective (In)consistency of Paint by Text. *arXiv* **2022**, arXiv:2206.14617. [CrossRef]
314. Farid, H. Lighting (In)consistency of Paint by Text. *arXiv* **2022**, arXiv:2207.13744. [CrossRef]
315. Sinitsa, S.; Fried, O. Deep Image Fingerprint: Accurate And Low Budget Synthetic Image Detector. *arXiv* **2023**, arXiv:2303.10762. [CrossRef]
316. Rajan, A.S.; Ojha, U.; Schloesser, J.; Lee, Y.J. Aligned Datasets Improve Detection of Latent Diffusion-Generated Images. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.
317. Nguyen, M.; Ho, K.; Nguyen, H.; Tu, C.; Tran, M.; Do, T. Unmasking The Artist: Discriminating Human-Drawn And AI-Generated Human Face Art Through Facial Feature Analysis. In *Proceedings of the 2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*; IEEE: New York, NY, USA, 2023; pp. 1–6. [CrossRef]
318. Poredi, N.; Nagothu, D.; Chen, Y. Ausome: Authenticating social media images using frequency analysis. In *Disruptive Technologies in Information Sciences VII*; SPIE: Bellingham, WA, USA, 2023; Volume 12542, pp. 44–56. [CrossRef]
319. He, P.; Zhu, L.; Li, J.; Wang, S.; Li, H. Exposing AI-generated Videos: A Benchmark Dataset and a Local-and-Global Temporal Defect Based Detection Method. *arXiv* **2024**, arXiv:2405.04133. [CrossRef]
320. Wen, Y.; Kirchenbauer, J.; Geiping, J.; Goldstein, T. Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. In *Advances in Neural Information Processing Systems 36*; Curran Associates, Inc.: New York, NY, USA, 2023; Volume 36, pp. 63688–63704. [CrossRef]
321. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781. [CrossRef]
322. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE Computer Society: New York, NY, USA, 2017; pp. 2980–2988. [CrossRef]
323. La, T.; Tran, Q.; Tran, T.; Tran, A.; Dang-Nguyen, D.; Dao, M. Multimodal Cheapfakes Detection by Utilizing Image Captioning for Global Context. In *Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*; Dao, M., Dang-Nguyen, D., Riegler, M., Eds.; ACM: New York, NY, USA, 2022; pp. 9–16. [CrossRef]
324. Papadopoulos, S.; Koutlis, C.; Papadopoulos, S.; Petrantonakis, P.C. Similarity Over Factuality: Are we Making Progress on Multimodal Out-of-Context Misinformation Detection? In *Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2025; pp. 5041–5050. [CrossRef]
325. Shalabi, F.; Felouat, H.; Nguyen, H.H.; Echizen, I. Leveraging Chat-Based Large Vision Language Models for Multimodal Out-of-Context Detection. In *Advanced Information Networking and Applications*; LNDECT; Barolli, L., Ed.; Springer: Berlin/Heidelberg, Germany, 2024; Volume 202, pp. 86–98. [CrossRef]
326. Gu, Y.; Tong, Z.; Castro, I.; Wu, S.; Tyson, G. Multi-MLLM Knowledge Distillation for Out-of-Context News Detection. *arXiv* **2025**, arXiv:2505.22517. [CrossRef]
327. Lakara, K.; Sock, J.; Rupprecht, C.; Torr, P.; Collomosse, J.P.; de Witt, C.S. MAD-Sherlock: Multi-Agent Debates for Out-of-Context Misinformation Detection. *arXiv* **2024**, arXiv:2410.20140. [CrossRef]

328. Maab, I.; Yamagishi, J. Pushing the Frontiers of Scientific Fact-Checking: The SCINLP Dataset. In *Proceedings of the Findings of the Association for Computational Linguistics: EACL 2026*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2026; pp. 1590–1617. [CrossRef]
329. Alvarez, C.; Bennett, M.; Wang, L.L. Zero-shot scientific claim verification using LLMs and citation text. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 269–276. [CrossRef]
330. Sevgili, Ö.; Nikishina, I.; Yimam, S.M.; Semmann, M.; Biemann, C. UHH at AVeriTeC: RAG for fact-checking with real-world claims. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 55–63. [CrossRef]
331. Ullrich, H.; Mlynár, T.; Drchal, J. AIC CTU system at AVeriTeC: Re-framing automated fact-checking as a simple RAG task. *arXiv* **2024**, arXiv:2410.11446. [CrossRef]
332. Zheng, H.S.; Mishra, S.; Chen, X.; Cheng, H.T.; Chi, E.H.; Le, Q.V.; Zhou, D. Take a step back: Evoking reasoning via abstraction in large language models. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
333. Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; Weston, J. Chain-of-verification reduces hallucination in large language models. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 3563–3578. [CrossRef]
334. Weng, Y.; Zhu, M.; Xia, F.; Li, B.; He, S.; Liu, S.; Sun, B.; Liu, K.; Zhao, J. Large language models are better reasoners with self-verification. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 2550–2575. [CrossRef]
335. Wu, Z.; Zeng, Q.; Zhang, Z.; Tan, Z.; Shen, C.; Jiang, M. Large language models can self-correct with key condition verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Al-Onaizan, Y., Bansal, M., Chen, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 12846–12867. [CrossRef]
336. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. Self-refine: Iterative refinement with self-feedback. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46534–46594. [CrossRef]
337. Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 11809–11822. [CrossRef]
338. Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. Graph of thoughts: Solving elaborate problems with large language models. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 17682–17690. [CrossRef]
339. Yang, B.; Xu, D.; Niu, K.; Liu, W.; Wang, Z.; Kankanhalli, M. A New Dataset and Benchmark for Grounding Multimodal Misinformation. In *Proceedings of the 33rd ACM International Conference on Multimedia*; ACM: New York, NY, USA, 2025; pp. 12571–12577. [CrossRef]
340. Ma, J.; Hu, L.; Li, R.; Fu, W. LoCal: Logical and Causal Fact-Checking with LLM-Based Multi-Agents. In *Proceedings of the WWW '25: Proceedings of the ACM on Web Conference 2025*; ACM: New York, NY, USA, 2025; pp. 1614–1625. [CrossRef]
341. Leippold, M.; Ashraf Vaghefi, S.; Stammbach, D.; Muccione, V.; Bingler, J.A.; Ni, J.; Colesanti Senni, C.; Wekhof, T.; Schimanski, T.; Gostlow, G.; et al. Automated Fact-Checking of Climate Change Claims with Large Language Models. *arXiv* **2024**, arXiv:2401.12566. [CrossRef]
342. DeVerna, M.R.; Yang, K.C.; Yan, H.Y.; Menczer, F. Large Language Models Require Curated Context for Reliable Political Fact-Checking—Even with Reasoning and Web Search. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2026*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2026; pp. 29338–29360. [CrossRef]
343. Sun, J.; Warren, G.; Shklovski, I.; Augenstein, I. Explaining Sources of Uncertainty in Automated Fact-Checking. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2026; pp. 45510–45534. [CrossRef]
344. Chun, C.; Zhang, D.C.; Lee, D. When Misinformation Speaks and Converses: Rethinking Fact-Checking in Audio Platforms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2026; pp. 2060–2075. [CrossRef]
345. Yang, F.; Pentyala, S.K.; Mohseni, S.; Du, M.; Yuan, H.; Linder, R.; Ragan, E.D.; Ji, S.; Hu, X.B. XFake: Explainable Fake News Detector with Visualizations. In *Proceedings of the World Wide Web Conference*; Liu, L., White, R.W., Mantrach, A., Silvestri, F., McAuley, J.J., Baeza-Yates, R., Zia, L., Eds.; ACM: New York, NY, USA, 2019; pp. 3600–3604. [CrossRef]
346. Shu, K.; Cui, L.; Wang, S.; Lee, D.; Liu, H. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Teredesai, A., Kumar, V., Li, Y., Rosales, R., Terzi, E., Karypis, G., Eds.; ACM: New York, NY, USA, 2019; pp. 395–405. [CrossRef]
347. Lu, Y.; Li, C. 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*; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 505–514. [CrossRef]

348. Lourenço, V.; Paes, A. A Modality-level Explainable Framework for Misinformation Checking in Social Networks. *arXiv* **2022**, arXiv:2212.04272. [CrossRef]
349. Wu, L.; Rao, Y.; Zhao, Y.; Liang, H.; Nazir, A. DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1024–1035. [CrossRef]
350. Gad-Elrab, M.H.; Stepanova, D.; Urbani, J.; Weikum, G. 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*; Culpepper, J.S., Moffat, A., Bennett, P.N., Lerman, K., Eds.; ACM: New York, NY, USA, 2019; pp. 87–95. [CrossRef]
351. Ahmadi, N.; Lee, J.; Papotti, P.; Saeed, M. Explainable Fact Checking with Probabilistic Answer Set Programming. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*; Liakata, M., Vlachos, A., Eds.; ACM: New York, NY, USA, 2019; pp. 1–8. [CrossRef]
352. Dziri, N.; Madotto, A.; Zaïane, O.; Bose, A.J. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*; Moens, M., Huang, X., Specia, L., Yih, S.W., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2197–2214. [CrossRef]
353. Cao, H.; Wei, L.; Zhou, W.; Hu, S. Multi-source Knowledge Enhanced Graph Attention Networks for Multimodal Fact Verification. In *Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: New York, NY, USA, 2024; pp. 1–6. [CrossRef]
354. Duwal, S.; Shopnil, M.N.S.; Tyagi, A.; Proma, A.M. Evidence-Grounded Multimodal Misinformation Detection with Attention-Based GNNs. *arXiv* **2025**, arXiv:2505.18221. [CrossRef]
355. Serrano, S.; Smith, N.A. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Korhonen, A., Traum, D.R., Màrquez, L., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 2931–2951. [CrossRef]
356. Das, A.; Agrawal, H.; Zitnick, L.; Parikh, D.; Batra, D. Human Attention in Visual Question Answering: Do Humans and Deep Networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*; Su, J., Carreras, X., Duh, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 932–937. [CrossRef]
357. Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; Rohrbach, M. Towards VQA Models That Can Read. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 8317–8326. [CrossRef]
358. Chaudhary, A.; Rajabi, E.; Kafaie, S.; Milios, E.E. Fact retrieval from knowledge graphs through semantic and contextual attention. *Expert Syst. Appl.* **2025**, *282*, 127612. [CrossRef]
359. Slobodkin, A.; Hirsch, E.; Cattan, A.; Schuster, T.; Dagan, I. Attribute First, then Generate: Locally-attributable Grounded Text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Ku, L., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 3309–3344. [CrossRef]
360. Gao, L.; Dai, Z.; Pasupat, P.; Chen, A.; Chaganty, A.T.; Fan, Y.; Zhao, V.Y.; Lao, N.; Lee, H.; Juan, D.; et al. RARR: Researching and Revising What Language Models Say, Using Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Rogers, A., Boyd-Graber, J.L., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 16477–16508. [CrossRef]
361. Wang, Y.; Reddy, R.G.; Mujahid, Z.M.; Arora, A.; Rubashevskii, A.; Geng, J.; Afzal, O.M.; Pan, L.; Borenstein, N.; Pillai, A.; et al. Factcheck-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output. *arXiv* **2023**, arXiv:2311.09000. [CrossRef]
362. Li, H.; Han, X.; Wang, H.; Wang, Y.; Wang, M.; Xing, R.; Geng, Y.; Zhai, Z.; Nakov, P.; Baldwin, T. Loki: An Open-Source Tool for Fact Verification. In *Proceedings of the 31st International Conference on Computational Linguistics*; Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S., Mather, B., Dras, M., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 28–36.
363. Hu, X.; Ru, D.; Qiu, L.; Guo, Q.; Zhang, T.; Xu, Y.; Luo, Y.; Liu, P.; Zhang, Y.; Zhang, Z. RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models. *arXiv* **2024**, arXiv:2405.14486. [CrossRef]
364. Iqbal, H.; Wang, Y.; Wang, M.; Georgiev, G.N.; Geng, J.; Gurevych, I.; Nakov, P. OpenFactCheck: A Unified Framework for Factuality Evaluation of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*; Farías, D.I.H., Hope, T., Li, M., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 219–229. [CrossRef]
365. Zheng, L.; Li, C.; Liu, Z.; Huang, F.; Jia, H.; Ye, Z.; Zhang, X. Fact in fragments: Deconstructing complex claims via LLM-based atomic fact extraction and verification. *Expert Syst. Appl.* **2026**, *302*, 130572. [CrossRef]
366. Tang, L.; Laban, P.; Durrett, G. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Al-Onaizan, Y., Bansal, M., Chen, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 8818–8847. [CrossRef]

367. Zhang, X.; Gao, W. 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: Stroudsburg, PA, USA, 2023; pp. 996–1011. [CrossRef]
368. He, Y.; Li, X.; Huang, Z.; Dong, Y.; Fei, H.; Zhang, J.; Wu, B.; Cheng, G. Towards Explainable Bilingual Multimodal Misinformation Detection and Localization. *arXiv* **2025**, arXiv:2506.22930. [CrossRef]
369. Chakraborty, M.; Pahwa, K.; Rani, A.; Chatterjee, S.; Dalal, D.; Dave, H.; G, R.; Gurumurthy, P.; Mahor, A.; Mukherjee, S.; et al. FACTIFY3M: A benchmark for multimodal fact verification with explainability through 5W Question-Answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 15282–15322. [CrossRef]
370. Jolly, S.; Atanasova, P.; Augenstein, I. Generating Fluent Fact Checking Explanations with Unsupervised Post-Editing. *Information* **2022**, *13*, 500. [CrossRef]
371. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 7871–7880. [CrossRef]
372. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 140.
373. de Oliveira, L.L.; Vargas, D.S.; Alexandre, A.M.A.; Cordeiro, F.C.; Gomes, D.d.S.M.; Rodrigues, M.d.C.; Romeu, R.K.; Moreira, V.P. Evaluating and mitigating the impact of OCR errors on information retrieval. *Int. J. Digit. Libr.* **2023**, *24*, 45–62. [CrossRef]
374. Szymanski, P.; Augustyniak, L.; Morzy, M.; Szymczak, A.; Surdyk, K.; Zelasko, P. Why Aren't We NER Yet? Artifacts of ASR Errors in Named Entity Recognition in Spontaneous Speech Transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Rogers, A., Boyd-Graber, J.L., Okazaki, N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 1746–1761. [CrossRef]
375. Rohrbach, A.; Hendricks, L.A.; Burns, K.; Darrell, T.; Saenko, K. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*; Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 4035–4045. [CrossRef]
376. Jayaweera, C.; Youm, S.; Dorr, B.J. AMREx: AMR for Explainable Fact Verification. *arXiv* **2024**, arXiv:2411.01343. [CrossRef]
377. Althabiti, S.; Alsalka, M.A.; Atwell, E. Ta'keed: The First Generative Fact-Checking System for Arabic Claims. *arXiv* **2024**, arXiv:2401.14067. [CrossRef]
378. Xie, A.; Zhu, F.; Han, J.; Hu, S. Integrating Open-Domain Knowledge via Large Language Model for Multimodal Fake News Detection. In *Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*; IEEE: New York, NY, USA, 2024; pp. 1917–1922. [CrossRef]
379. Khaliq, M.A.; Chang, P.; Ma, M.; Pflugfelder, B.; Miletic, F. 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: Stroudsburg, PA, USA, 2024; pp. 280–296. [CrossRef]
380. Pan, L.; Wu, X.; Lu, X.; Luu, A.T.; Wang, W.Y.; Kan, M.Y.; Nakov, P. 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: Stroudsburg, PA, USA, 2023; pp. 6981–7004. [CrossRef]
381. Cekinel, R.F.; Karagoz, P. Explaining Veracity Predictions with Evidence Summarization: A Multi-Task Model Approach. *arXiv* **2024**, arXiv:2402.06443. [CrossRef]
382. Kim, K.; Lee, S.; Huang, K.H.; Chan, H.P.; Li, M.; Ji, H. Can LLMs Produce Faithful Explanations for Fact-Checking? Towards Faithful Explainable Fact-Checking via Multi-Agent Debate. *arXiv* **2024**, arXiv:2402.07401. [CrossRef]
383. Sahnan, D.; Corney, D.; Larraz, I.; Zagni, G.; Miguez, R.; Xie, Z.; Gurevych, I.; Churchill, E.; Chakraborty, T.; Nakov, P. Can LLMs Automate Fact-Checking Article Writing? *Trans. Assoc. Comput. Linguist.* **2026**, *14*, 489–509. [CrossRef]
384. Comito, C.; Caroprese, L.; Zumpano, E. Multimodal Fake News Detection on Social Media: A Survey of Deep Learning Techniques. *Soc. Netw. Anal. Min.* **2023**, *13*, 101. [CrossRef]
385. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The Long-Document Transformer. *arXiv* **2020**, arXiv:2004.05150. [CrossRef]
386. Dietterich, T.G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [CrossRef] [PubMed]
387. Khan, S.; Naz, N.S.; Mazhar, T.; Tariq, M.U.; Shahzad, T.; Guizani, S.; Hamam, H. Green AI techniques for reducing energy consumption in AI systems. *Array* **2026**, *29*, 100652. [CrossRef]
388. Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv* **2024**, arXiv:2412.05271. [CrossRef]
389. Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. LLaVA-OneVision: Easy Visual Task Transfer. *Trans. Mach. Learn. Res.* **2025**, *2025*, 1–44.

390. John, S.; Del, R.R.F.; Evgeniy, K.; Helen, O.; Idan, H.; Kayla, U.; Ken, H.; Peter, S.; Rakshith, A.; Ron, B.; et al. *Owasp Top 10 for LLM Apps & Gen AI Agentic Security Initiative*; Technical Report; OWASP Foundation: Wilmington, DE, USA, 2025.
391. Ali, R. Scalable and Secure AI Inference in Healthcare: A Comparative Benchmarking of FastAPI and Triton Inference Server on Kubernetes. *arXiv* **2026**, arXiv:2602.00053. [CrossRef]
392. Li, B.; Jiang, Y.; Gadepally, V.; Tiwari, D. LLM Inference Serving: Survey of Recent Advances and Opportunities. In *Proceedings of the 2024 IEEE High Performance Extreme Computing Conference (HPEC)*; IEEE: New York, NY, USA, 2024; pp. 1–8. [CrossRef]
393. Ashkboos, S.; Croci, M.L.; Nascimento, M.G.D.; Hoefler, T.; Hensman, J. SliceGPT: Compress Large Language Models by Deleting Rows and Columns. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
394. Bao, Z.; Ding, S.; Chen, B.; Zhang, P. Difficulty-aware and relational decoupled knowledge distillation. *Neurocomputing* **2026**, *692*, 133850. [CrossRef]
395. Saadi, K.; Wang, D. TASKD-LLM: Task-Aware Selective Knowledge Distillation for LLMs. In Proceedings of the ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy, Singapore, 28 April 2025.
396. Huang, W.; Qin, H.; Liu, Y.; Li, Y.; Liu, Q.; Liu, X.; Benini, L.; Magno, M.; Zhang, S.; Qi, X. SliM-LLM: Salience-Driven Mixed-Precision Quantization for Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*; Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., Zhu, J., Eds.; PMLR: Cambridge, MA, USA, 2025; Volume 267, pp. 25672–25692.
397. Li, Y.; Liao, B.; Liu, W.; Wang, X. MaTVLM: Hybrid Mamba-Transformer for Efficient Vision-Language Modeling. In *Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2025; pp. 20878–20888. [CrossRef]
398. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
399. Albert, P.; Zhang, F.Z.; Saratchandran, H.; Opazo, C.R.; Hengel, A.v.d.; Abbasnejad, E. RandLoRA: Full rank parameter-efficient fine-tuning of large models. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.
400. Shah, J.; Bikshandi, G.; Zhang, Y.; Thakkar, V.; Ramani, P.; Dao, T. FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision. In *Advances in Neural Information Processing Systems 37*; Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C., Eds.; Curran Associates, Inc.: New York, NY, USA, 2024; Volume 37, pp. 68658–68685. [CrossRef]
401. Agrawal, A.; Kedia, N.; Panwar, A.; Mohan, J.; Kwatra, N.; Gulavani, B.S.; Tumanov, A.; Ramjee, R. Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve. In *Proceedings of the 18th USENIX Conference on Operating Systems Design and Implementation*; Gavrilovska, A., Terry, D.B., Eds.; ACM: New York, NY, USA, 2024; pp. 117–134.
402. Cho, J.; Lei, J.; Tan, H.; Bansal, M. Unifying Vision-and-Language Tasks via Text Generation. In *Proceedings of the 38th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2021; Volume 139, pp. 1931–1942.
403. Pfeiffer, J.; Kamath, A.; Rücklé, A.; Cho, K.; Gurevych, I. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 487–503. [CrossRef]
404. Shayegani, E.; Al Mamun, M.A.; Fu, Y.; Zaree, P.; Dong, Y.; Abu-Ghazaleh, N.B. Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks. *arXiv* **2023**, arXiv:2310.10844. [CrossRef]
405. Zhu, K.; Wang, J.; Zhou, J.; Wang, Z.; Chen, H.; Wang, Y.; Yang, L.; Ye, W.; Zhang, Y.; Gong, N.; et al. PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*; ACM: Stroudsburg, PA, USA, 2024; pp. 57–68. [CrossRef]
406. Yang, Z.; Meng, Z.; Zheng, X.; Wattenhofer, R. Assessing Adversarial Robustness of Large Language Models: An Empirical Study. *arXiv* **2024**, arXiv:2405.02764. [CrossRef]
407. Zou, J.; Zhang, S.; Qiu, M. Adversarial Attacks on Large Language Models. In *Knowledge Science, Engineering and Management*; LNCS; Springer: Berlin/Heidelberg, Germany, 2024; Volume 14887, pp. 85–96. [CrossRef]
408. Jin, D.; Jin, Z.; Zhou, J.T.; Szolovits, P. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 8018–8025. [CrossRef]
409. Golchin, S.; Surdeanu, M. Time Travel in LLMs: Tracing Data Contamination in Large Language Models. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.
410. Zou, A.; Wang, Z.; Kolter, J.Z.; Fredrikson, M. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv* **2024**, arXiv:2307.15043. [CrossRef]
411. Liu, F.; Abuadbba, S.; Moore, K.; Nepal, S.; Paris, C.; Wu, J.; Yang, J.; Sheng, Q.Z. Adversarial Attacks Against Automated Fact-Checking: A Survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 22968–22990. [CrossRef]

412. Wang, Z.; Fan, W.; Zong, Q.; Zhang, H.; Choi, S.; Fang, T.; Liu, X.; Song, Y.; Wong, G.Y.; See, S. AbsInstruct: Eliciting Abstraction Ability from LLMs through Explanation Tuning with Plausibility Estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 973–994. [CrossRef]
413. Allein, L.; Saelens, M.; Cartuyvels, R.; Moens, M.F. Implicit Temporal Reasoning for Evidence-Based Fact-Checking. In *Findings of the Association for Computational Linguistics: EACL 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 176–189. [CrossRef]
414. Jiang, X.; Chang, D.; McAuley, J.; Xu, X. When Benchmarks Age: Temporal Misalignment through Large Language Model Factuality Evaluation. *arXiv* **2025**, arXiv:2510.07238. [CrossRef]
415. Herel, D.; Bartek, V.; Jirak, J.; Mikolov, T. Time Awareness in Large Language Models: Benchmarking Fact Recall Across Time. *arXiv* **2024**, arXiv:2409.13338. [CrossRef]
416. Chen, W.; Wang, X.; Wang, W.Y. A Dataset for Answering Time-Sensitive Questions. *Proc. Proc. NeurIPS* **2021**, *34*, 8305–8317.
417. Allein, L.; Augenstein, I.; Moens, M.F. Time-Aware Evidence Ranking for Fact-Checking. *J. Web Semant.* **2021**, *71*, 100663. [CrossRef]
418. Vu, T.; Iyyer, M.; Wang, X.; Constant, N.; Wei, J.W.; Wei, J.; Tar, C.; Sung, Y.; Zhou, D.; Le, Q.V.; et al. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024*; Ku, L., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 13697–13720. [CrossRef]
419. Wu, S.; Dredze, M. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*; Gella, S., Welbl, J., Rei, M., Petroni, F., Lewis, P., Strubell, E., Seo, M.J., Hajishirzi, H., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 120–130. [CrossRef]
420. Qin, L.; Chen, Q.; Zhou, Y.; Chen, Z.; Li, Y.; Liao, L.; Li, M.; Che, W.; Yu, P.S. A survey of multilingual large language models. *Patterns* **2025**, *6*, 101118. [CrossRef] [PubMed]
421. Tao, Y.; Viberg, O.; Baker, R.S.; Kizilcec, R.F. Cultural Bias and Cultural Alignment of Large Language Models. *PNAS Nexus* **2024**, *3*, gae346. [CrossRef] [PubMed]
422. Basmova, V.; Goldberg, Y.; Tsarfaty, R. LLMs' Reading Comprehension Is Affected by Parametric Knowledge and Struggles with Hypothetical Statements. *arXiv* **2024**, arXiv:2404.06283. [CrossRef]
423. Chiu, Y.Y.; Jiang, L.; Lin, B.Y.; Park, C.Y.; Li, S.S.; Ravi, S.; Bhatia, M.; Antoniak, M.; Tsvetkov, Y.; Shwartz, V.; et al. CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs' Cultural Knowledge through Human-AI Red-Teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 25663–25701. [CrossRef]
424. Lai, H.; Nissim, M. MCoT: Multilingual Instruction Tuning for Reasoning Consistency in Language Models. In *Proceedings of the MCoT: Multilingual Instruction Tuning for Reasoning Consistency in Language Models*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 12012–12026. [CrossRef]
425. Conneau, A.; Rinott, R.; Lample, G.; Williams, A.; Bowman, S.R.; Schwenk, H.; Stoyanov, V. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2475–2485. [CrossRef]
426. Zhang, J.; Xu, C.; Gai, Y.; Lécué, F.; Song, D.; Li, B. KnowHalu: Hallucination Detection via Multi-Form Knowledge Based Factual Checking. *arXiv* **2024**, arXiv:2404.02935. [CrossRef]
427. Augenstein, I.; Baldwin, T.; Cha, M.; Chakraborty, T.; Ciampaglia, G.L.; Corney, D.P.A.; DiResta, R.; Ferrara, E.; Hale, S.; Halevy, A.Y.; et al. Factuality Challenges in the Era of Large Language Models and Opportunities for Fact-Checking. *Nat. Mach. Intell.* **2024**, *6*, 852–863. [CrossRef]
428. Maynez, J.; Narayan, S.; Bohnet, B.; McDonald, R.T. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1906–1919. [CrossRef]
429. Shorinwa, O.; Mei, Z.; Lidard, J.; Ren, A.Z.; Majumdar, A. A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions. *ACM Comput. Surv.* **2026**, *58*, 63. [CrossRef]
430. Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; Hooi, B. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *Proceedings of the International Conference on Learning Representations*; OpenReview.net: Amherst, MA, USA.
431. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.V.; Chi, E.H.; Narang, S.; Chowdhery, A.; Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*; OpenReview.net: Amherst, MA, USA, 2023.
432. Bach, S.H.; Broecheler, M.; Huang, B.; Getoor, L. Hinge-Loss Markov Random Fields and Probabilistic Soft Logic. *J. Mach. Learn. Res.* **2017**, *18*, 109.

433. Geifman, Y.; El-Yaniv, R. Selective Classification for Deep Neural Networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4878–4887.
434. Bates, S.; Angelopoulos, A.; Lei, L.; Malik, J.; Jordan, M.I. Distribution-free, Risk-controlling Prediction Sets. *J. ACM* **2021**, *68*, 43. [CrossRef]
435. Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. Hallusionbench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 14375–14385. [CrossRef]
436. Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Wang, J.; Xu, H.; Yan, M.; Zhang, J.; et al. AMBER: An LLM-Free Multi-Dimensional Benchmark for MLLMs Hallucination Evaluation. *arXiv* **2023**, arXiv:2311.07397. [CrossRef]
437. Manakul, P.; Liusie, A.; Gales, M.J.F. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 9004–9017. [CrossRef]
438. Gao, T.; Yen, H.; Yu, J.; Chen, D. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 6465–6488. [CrossRef]
439. Li, Q.; Geng, J.; Lyu, C.; Zhu, D.; Panov, M.; Karray, F. Reference-free Hallucination Detection for Large Vision-Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*; Al-Onaizan, Y., Bansal, M., Chen, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 4542–4551. [CrossRef]
440. Chen, X.; Wang, C.; Xue, Y.; Zhang, N.; Yang, X.; Li, Q.; Shen, Y.; Liang, L.; Gu, J.; Chen, H. Unified Hallucination Detection for Multimodal Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 3235–3252. [CrossRef]
441. Bouchard, D.; Chauhan, M.S. Uncertainty Quantification for Language Models: A Suite of Black-Box, White-Box, LLM Judge, and Ensemble Scorers. *arXiv* **2025**, arXiv:2504.19254. [CrossRef]
442. Thorne, J.; Vlachos, A. Evidence-based Factual Error Correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 3298–3309. [CrossRef]
443. Chen, J.; Xu, R.; Zeng, W.; Sun, C.; Li, L.; Xiao, Y. Converge to the Truth: Factual Error Correction via Iterative Constrained Editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Williams, B., Chen, Y., Neville, J., Eds.; AAAI Press: Washington, DC, USA, 2023; pp. 12616–12625. [CrossRef]
444. He, X.; Zhang, Q.; Jin, A.; Ma, J.; Yuan, Y.; Yiu, S.M. Improving Factual Error Correction by Learning to Inject Factual Errors. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Wooldridge, M.J., Dy, J.G., Natarajan, S., Eds.; AAAI Press: Washington, DC, USA, 2024; pp. 18197–18205. [CrossRef]
445. He, X.; Jin, A.; Ma, J.; Yuan, Y.; Yiu, S.M. PivotFEC: Enhancing Few-shot Factual Error Correction with a Pivot Task Approach using Large Language Models. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 9960–9976. [CrossRef]
446. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A.Y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*; Singh, A., Zhu, X.J., Eds.; Proceedings of Machine Learning Research (PMLR): Cambridge, MA, USA, 2017; pp. 1273–1282.
447. Dwork, C. Differential Privacy. In *Automata, Languages and Programming*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4052, pp. 1–12. [CrossRef]
448. Bonawitz, K.A.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H.B.; Patel, S.; Ramage, D.; Segal, A.; Seth, K. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*; Thuraisingham, B., Evans, D., Malkin, T., Xu, D., Eds.; ACM: New York, NY, USA, 2017; pp. 1175–1191. [CrossRef]
449. Edemacu, K.; Wu, X. Privacy Preserving Prompt Engineering: A Survey. *ACM Comput. Surv.* **2025**, *57*, 255. [CrossRef]
450. Zhao, H.; Peng, P.; Chen, C.; Huang, L.; Liu, G. FedRS-Bench: Realistic Federated Learning Datasets and Benchmarks in Remote Sensing. *arXiv* **2025**, arXiv:2505.08325. [CrossRef]
451. Shokri, R.; Strobel, M.; Zick, Y. On the Privacy Risks of Model Explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*; Fourcade, M., Kuipers, B., Lazar, S., Mulligan, D.K., Eds.; ACM: New York, NY, USA, 2021; pp. 231–241. [CrossRef]
452. Kadavath, S.; Conerly, T.; Askell, A. Language Models (Mostly) Know What They Know. *arXiv* **2022**, arXiv:2207.05221. [CrossRef]
453. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]

454. Turpin, M.; Michael, J.; Perez, E.; Bowman, S. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 74952–74965. [CrossRef]
455. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70, Sydney, NSW, Australia, 6–11 August 2017*; Precup, D., Teh, Y.W., Eds.; Proceedings of Machine Learning Research (PMLR): Cambridge, MA, USA, 2017; pp. 1321–1330.
456. DeGroot, M.H.; Fienberg, S.E. The Comparison and Evaluation of Forecasters. *J. R. Stat. Soc. Ser. D* **1983**, *32*, 12–22. [CrossRef]
457. Naeini, M.P.; Cooper, G.F.; Hauskrecht, M. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Bonet, B., Koenig, S., Eds.; AAAI Press: Washington, DC, USA, 2015; pp. 2901–2907. [CrossRef]
458. Zhu, K.; Zheng, Y.; Chan, K.C.G. Weighted Brier Score–An Overall Summary Measure for Risk Prediction Models with Clinical Utility Consideration. *Stat. Biosci.* **2025**. Correction in *Stat. Biosci.* **2025**. https://doi.org/10.1007/s12561-025-09507-3. [CrossRef] [PubMed]
459. Platt, J.C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*; MIT Press: Cambridge, MA, USA, 1999; pp. 61–74.
460. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2002; pp. 694–699. [CrossRef]
461. Angelopoulos, A.N.; Bates, S. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv* **2021**, arXiv:2107.07511. [CrossRef]
462. Nakov, P.; Corney, D.P.A.; Hasanain, M.; Alam, F.; Elsayed, T.; Barrón-Cedeño, A.; Papotti, P.; Shaar, S.; Martino, G.D.S. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*; Zhou, Z., Ed.; IJCAI: Oregon City, OR, USA, 2021; pp. 4551–4558. [CrossRef] [PubMed]
463. Dai, S.; Xu, C.; Xu, S.; Pang, L.; Dong, Z.; Xu, J. Unifying Bias and Unfairness in Information Retrieval: A Survey of Challenges and Opportunities with Large Language Models. *arXiv* **2024**, arXiv:2404.11457. [CrossRef]
464. Gorwa, R.; Binns, R.; Katzenbach, C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Soc.* **2020**, *7*, 205395171989794. [CrossRef]
465. Chen, X.; Fang, H.; Lin, T.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv* **2015**, arXiv:1504.00325. [CrossRef]
466. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
467. Elliott, D.; Frank, S.; Sima'an, K.; Specia, L. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*; The Association for Computer Linguistics: Stroudsburg, PA, USA, 2016; pp. 84–90. [CrossRef]
468. Xu, J.; Mei, T.; Yao, T.; Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5288–5296. [CrossRef]
469. Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; Niebles, J.C. Dense-Captioning Events in Videos. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE Computer Society: Washington, DC, USA, 2017; pp. 706–715. [CrossRef]
470. Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.; Wang, W.Y. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4580–4590. [CrossRef]
471. Kim, C.D.; Kim, B.; Lee, H.; Kim, G. AudioCaps: Generating Captions for Audios in the Wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 119–132. [CrossRef]
472. Drossos, K.; Lipping, S.; Virtanen, T. Clotho: An Audio Captioning Dataset. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 736–740. [CrossRef]
473. Bai, J.; Liu, H.; Wang, M.; Shi, D.; Wang, W.; Plumbley, M.D.; Gan, W.S.; Chen, J. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2025**, *33*, 2817–2829. [CrossRef]
474. Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; Feris, R. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11307–11317. [CrossRef]
475. Liu, Z.; Opazo, C.R.; Teney, D.; Gould, S. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 2105–2114. [CrossRef]

476. Baldrati, A.; Agnolucci, L.; Bertini, M.; Del Bimbo, A. Zero-Shot Composed Image Retrieval with Textual Inversion. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 15292–15301. [CrossRef]
477. Radenovic, F.; Iscen, A.; Tolias, G.; Avrithis, Y.; Chum, O. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition/IEEE Computer Society, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5706–5715. [CrossRef]
478. Ellis, D.P.W.; Cotton, C.V. *The 2007 LabROSA Cover Song Detection System*; MIREX 2007 Extended Abstract; Music Information Retrieval Evaluation EXchange: Champaign, IL, USA, 2007.
479. Berenzweig, A.; Logan, B.; Ellis, D.P.W.; Whitman, B. A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measures. *Comput. Music J.* **2004**, *28*, 63–76. [CrossRef]
480. Xu, X.; Chen, X.; Yang, D. Key-Invariant Convolutional Neural Network Toward Efficient Cover Song Identification. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6. [CrossRef]
481. Yesiler, F.; Tralie, C.; Correya, A.A.; Silva, D.F.; Tovstogan, P.; Gómez, E.; Serra, X. Da-TACOS: A Dataset for Cover Song Identification and Understanding. In *Proceedings of the 20th Conference of the International Society for Music Information Retrieval (ISMIR 2019), Delft, The Netherlands, 4–8 November 2019*; Flexer, A., Peeters, G., Urbano, J., Volk, A., Eds.; International Society for Music Information Retrieval (ISMIR): Montreal, QC, Canada, 2019; pp. 327–334.
482. Cheng, X.; Zhang, Z.; Wang, Z.; Fang, M.; Huang, R.; Zheng, S.; Hu, R.; Jionghao, B.; Jin, T.; Zhao, Z. AVSET-10M: An Open Large-Scale Audio-Visual Dataset with High Correspondence. GitHub Repository. [Dataset]. 2024. Available online: https://github.com/AVSET-10M/avset-10m (accessed on 3 November 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.