
Proceedings Paper:
Deng, X., Wang, X. and Stevenson, R. orcid.org/0000-0002-9483-6006 (2025) The next phase of scientific fact-checking: advanced evidence retrieval from complex structured academic papers. In: Zamani, H., (ed.) ICTIR '25: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR). ICTIR '25: International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, 18 Jul 2025, Padua, Italy. ACM ISBN 9798400718618

https://doi.org/10.1145/3731120.3744614

The Next Phase of Scientific Fact-Checking: Advanced Evidence Retrieval from Complex Structured Academic Papers

Xingyu Deng
xdeng37@sheffield.ac.uk
University of Sheffield
Sheffield, UK

Xi Wang
xi.wang@sheffield.ac.uk
University of Sheffield
Sheffield, UK

Mark Stevenson
mark.stevenson@sheffield.ac.uk
University of Sheffield
Sheffield, UK

Abstract

Scientific fact-checking aims to determine the veracity of scientific claims by retrieving and analysing evidence from research literature. The problem is inherently more complex than general fact-checking since it must accommodate the evolving nature of scientific knowledge, the structural complexity of academic literature and the challenges posed by long-form, multimodal scientific expression. However, existing approaches focus on simplified versions of the problem based on small-scale datasets consisting of abstracts rather than full papers, thereby avoiding the distinct challenges associated with processing complete documents. This paper examines the limitations of current scientific fact-checking systems and reveals the many potential features and resources that could be exploited to advance their performance. It identifies key research challenges within evidence retrieval, including (1) evidence-driven retrieval that addresses semantic limitations and topic imbalance (2) time-aware evidence retrieval with citation tracking to mitigate outdated information, (3) structured document parsing to leverage long-range context, (4) handling complex scientific expressions, including tables, figures, and domain-specific terminology and (5) assessing the credibility of scientific literature. Preliminary experiments were conducted to substantiate these challenges and identify potential solutions. This perspective paper aims to advance scientific fact-checking with a specialised IR system tailored for real-world applications.

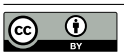
CCS Concepts

• **Information systems** → **Specialized information retrieval**.

Keywords

Evidence retrieval, Scientific fact-checking

ACM Reference Format:
Xingyu Deng, Xi Wang, and Mark Stevenson. 2025. The Next Phase of Scientific Fact-Checking: Advanced Evidence Retrieval from Complex Structured Academic Papers. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '25), July 18, 2025, Padua, Italy*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3731120.3744614



ICTIR '25, Padua, Italy

ACM ISBN 979-8-4007-1861-8/2025/07
https://doi.org/10.1145/3731120.3744614

1 Introduction

Fact-checking aims to assess the veracity of factual claims based on credible evidence [37, 116] and serves as a crucial safeguard for mitigating misinformation. Scientific fact-checking is a specialised variant of this task, grounded in scientific knowledge, with the objective of combating misinformation that affects the public, helping researchers in knowledge discovery and assisting individuals in understanding scientific advancements [97]. This is particularly important given the rapid emergence of new scientific findings, where both professionals and the public must assess the credibility of information. A prominent case occurred during the COVID-19 pandemic, in which politically motivated misinformation—ranging from inflated infection statistics to unsupported treatments—circulated extensively, eroding public trust and endangering health communication [58]. However, existing approaches to scientific fact-checking remain limited, primarily relying on the retrieval of evidence from relatively simple and small-scale sources [16, 49, 62, 72, 75, 100, 101, 104]. For example, SciFact-Open [101], the largest available dataset for scientific fact-checking, contains 500,000 documents – substantially smaller than PubMed, which contains over 37 million biomedical publications. In addition, SciFact-Open consists only of abstracts, rather than full-text papers, thereby excluding critical structural and citation information, ignoring long-range context and scientific expression conveyed through tables and figures. These design simplifications may hinder the applicability of current approaches in real-world settings, where scientific evidence is embedded in long and structurally complex documents with multimodal content.

Fact-checking is a knowledge-intensive task, where the verification process relies on sourcing evidence from a reliable upstream Information Retrieval (IR) system. Emerging findings indicate the value of effective retrieval in improving fact-checking systems. For example, introducing even a small amount of noise into evidence can significantly degrade fact-checking performance [76]. Recent Retrieval-Augmented-Generation (RAG) techniques have been widely used for fact-checking [28, 47, 63, 70, 80, 84, 94], where retrieval models are fine-tuned to identify high-quality evidence for claim verification. These observations underscore the critical role of robust evidence retrieval, as an ideal IR system for fact-checking should rank all relevant evidence at the top while filtering out non-evidential noise. Ensuring retrieval robustness is crucial to maintaining sufficient yet relevant evidence, which is essential for improving scientific fact-checking accuracy.

A major trend in fact-checking research is to consider realistic settings that employ rich, diverse and timely evidence sources, as seen in FEVER (using Wikipedia) [90] and AVeriTeC (using web-wide resources) [78]. In **document level evidence retrieval**, current

general fact-checking systems over-rely on commercial search APIs, which do not consider the specific requirements of fact-checking [77, 91, 116]. Such reliance on commercial search APIs – with limited adaptability – has left document retrieval methodologies underexplored in fact-checking, especially for domain-specific corpora such as scientific fact-checking. Current scientific fact-checking systems primarily employ off-the-shelf IR methods [97], such as lexical matching and semantic relevance ranking, which do not scale effectively for large-scale scientific corpora. In addition, the distribution of relevant evidence across scientific topics is highly imbalanced, which degrades both retrieval effectiveness and efficiency, especially for claims with scarce supporting literature. SciFact-Open [101], which extends the original SciFact dataset [100] for large-scale evaluation, illustrates this issue: verification performance on SciFact-Open drops by at least 15 F1 points for all well-performed fact-checking systems developed in SciFact [101]. While increasing corpus size enhances evidence diversity, it also amplifies retrieval noise, reducing efficiency in both retrieval and verification. Beyond that, high semantic relevance does not guarantee high evidential relevance, and irrelevant yet semantically similar documents can introduce noise into downstream verification [117]. These challenges underscore the necessity of developing tailored document retrieval systems specifically designed for scientific fact-checking, as effective retrieval is a prerequisite for accurate claim verification.

Beyond document-level evidence retrieval, **within-document evidence retrieval** is also essential for processing complex scientific literature. Scientific papers, unlike general fact-checking documents, are long, structured, domain-specific and involve additional metadata. As scientific fact-checking evolves from abstract-based to full-paper retrieval, retrieval models must account for metadata (e.g., publish date, citations) and complex structured data format (e.g., charts, tables, figures). This necessitates the adaptation of verification models such as SciBERT [11] for domain-specific terminologies [100], Longformer [12] for long-range dependencies [102] and TAPAS [38] for tabular data verification [3]. Scientific expressions in academic papers are highly structured and contextually interdependent, where textual content, tabular data, and figures mutually reinforce the conveyed information. However, existing scientific fact-checking systems primarily operate at the abstract level, adopting methodologies similar to general fact-checking [49, 50, 68, 72, 75, 97, 100, 102, 111, 119], albeit incorporating domain-specific models such as BioSentVec [21] and SciBERT [11]. The development of public full-paper datasets aligns with the requirement of real-world scientific fact-checking systems for effective verification. This highlights the urgent need for retrieval and verification methodologies that can leverage entire scientific documents. Accordingly, within-document evidence retrieval and its integration into verification pipelines should be explored to fully unlock the potential of scientific literature for fact-checking.

This perspective paper presents a comprehensive examination of the challenges associated with evidence retrieval in scientific fact-checking, **highlighting challenges that are not typically faced within general fact-checking**, leading to the need for specialised retrieval and verification strategies. We advocate for proactive research efforts to develop scalable methodologies while addressing the limitations of current datasets. We structure our discussion into two parts following a typical fact-checking pipeline: Sections 2–3 explore document-level evidence retrieval while Sections 4–8 explore within-document evidence retrieval in scientific publication for scientific fact-checking. Each of the following sections identifies a research challenge followed by a tentative and illustrative research direction (**RD**).

2 Beyond Semantics

Evidence retrieval for fact checking is closely related to traditional document retrieval techniques, which typically focus on retrieving documents that are semantically similar to a query or contain matching keywords. While this approach is effective in many scenarios, it often fails to address the ultimate objective of fact-checking – successful claim verification. Evidence retrieval that relies solely on semantic similarity may prioritise irrelevant or low-context information, introducing noise into the subsequent verification process. Recent IR studies [67, 92] show that semantic relevance alone may not ensure utility in knowledge-intensive NLP tasks under the RAG framework, suggesting the importance of utility-aware retrieval strategies. To improve the verification utility of evidence retrieval systems, techniques such as fine-tuning, joint optimisation, and learning from verification feedback have been developed. These approaches leverage relevance labels derived from annotated gold evidence [40, 68, 117, 119, 120]. Although graded relevance has been extensively explored in general IR, current evidence retrieval systems for fact-checking often oversimplify relevance as binary, failing to differentiate between fully non-evidential and partially relevant evidence. This coarse-grained labelling scheme fails to differentiate between completely non-evidential documents and partially relevant (plausible) evidence. Negative examples and randomly retrieved examples are equally treated as 0, despite exhibiting varying degrees of evidential support. We argue that evidence retrieval should distinguish between non-evidential information and plausible evidence, enhancing the model's ability to identify previously unobserved but potentially useful evidence within large-scale corpora.

Developing an IR system that can effectively differentiate between evidential and non-evidential information requires access to fine-grained relevance labels during training. However, manually constructing negative samples is both complex and resource-intensive due to the vast number of unlabelled documents and sentences that lack explicit pairing with given claims. Furthermore, assessing the degree of evidential support for a claim within unlabelled documents is inherently challenging. To validate the impact of fine-grained evidential relevance, beyond semantic relevance, we carry out preliminary experiments which explore the use of downstream verification feedback to capture different levels of evidential values.

Experiment Overview. The experiment investigates whether combining verification feedback with semantic relevance improves the performance of document evidence retrieval. Figure 1 presents the pipeline, where probabilities from downstream verification serve as feedback. Specifically, we integrate two components: 1) the semantic relevance score, computed using an off-the-shelf reranker model, and (2) the verification success feedback score, derived from a fine-tuned verifier model, indicating the degree to which a document is evidential for a given claim.

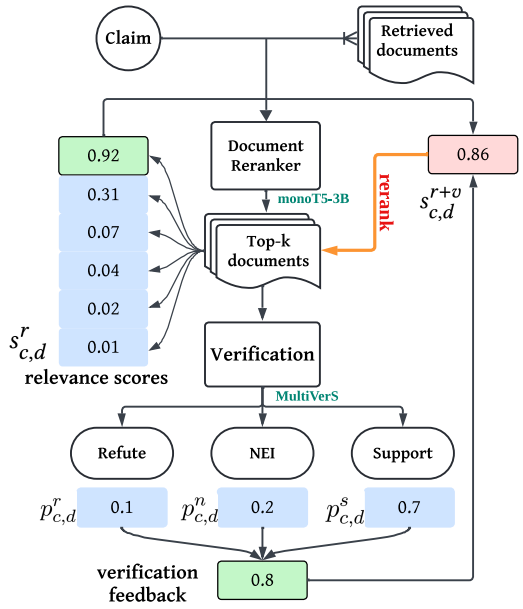


Figure 1: Pipeline of experiment

Approaches. ***monoT5-3B*** [64] has demonstrated strong performance as a reranker for SciFact, as evidenced by its widespread use as a strong baseline in studies [55, 86] on the BEIR benchmark [89]. It also demonstrated state-of-the-art performance in evidence retrieval for verification [68, 97, 102]. The model assigns a predicted score, $s^r_{c,d}$, representing the semantic relevance for a document d to a claim c as defined in Equation 1.

$$f(c,d) \rightarrow s^r_{c,d},\ s^r_{c,d} \in (0,1) \tag{1}$$

MultiVerS is the best-performing verifier model on SciFact [97, 102]. We reproduced this model using the official implementation[1] while adjusting the negative sampling parameter from 20 to 5 to avoid over-fitted verification feedback. The model predicts the verification outcome as per the calculated probabilities for a document d either supporting $p^r_{c,d}$, refuting $p^s_{c,d}$ or providing insufficient information $p^n_{c,d}$, relative to a claim c, as follows:

$$V(c,d) = argmax(p^r_{c,d}, p^n_{c,d}, p^s_{c,d}) \tag{2}$$

+Verification *(Ideal Reranker Model)* combines semantic relevance and verification feedback to refine evidential retrieval. The final retrieval score is calculated by summing the semantic score ($s^r_{c,d}$) and the verification probabilities ($p^r_{c,d}$ and $p^n_{c,d}$), followed by a normalisation step to ensure the score remains within (0,1), formulated as:

$$s^{r+v}_{c,d} = 1/2 * (s^r_{c,d} + p^r_{c,d} + p^s_{c,d}) \in (0,1) \tag{3}$$

This formulation ensures that documents contributing to support or refute labels receive higher retrieval scores, enhancing the evidential quality of retrieved documents.

To evaluate retrieval effectiveness, we use *Recall@k* (R@k), a common evaluation approach that measures the proportion of relevant evidence successfully retrieved within the top k results.

[1]https://github.com/dwadden/multivers

Table 1: Retrieval result on SciFact-Open and Check-COVID

SciFact-Open	R@50	R@20	R@10	R@5	R@3	R@1
BM25	66.09	54.78	45.22	38.04	30.87	20.22
monoT5-3B	88.91	79.13	71.09	57.17	48.26	31.09
+Verification	**91.96**	**81.95**	**71.30**	**62.61**	**52.83**	**32.17**
Check-COVID	R@50	R@20	R@10	R@5	R@3	R@1
BM25	87.91	81.96	75.02	67.59	61.35	46.18
monoT5-3B	95.84	93.16	89.49	82.06	74.93	58.28
+Verification	**96.13**	**94.55**	**91.48**	**84.04**	**77.80**	**61.84**

Table 2: Evidence positions in the retrieved list. We select an example from the SciFact-Open dataset. Claim: Female carriers of the Apolipoprotein E4 (APOE4) allele have a reduced risk for Alzheimer's disease. Gold Evidence: [E1,E2,E3,E4,E5]

reranker model	E1	E2	E3	E4	**E5**
BM25	838th	141th	7th	163th	**67th**
monot5-3B	1st	8th	16th	2nd	**302nd**
+Verification	1st	3rd	5th	2nd	**27th**

Datasets. We conduct our evaluation using datasets including: **(1) SciFact** [100]: A corpus of 5,183 abstracts from scientific articles, with 809/300/300 samples for train, validation and test sets. The test set is not publicly accessible. **(2) SciFact-Open** [101]: An extended version of SciFact with 500K abstracts, re-annotating evidence documents for 279 claims from the original SciFact test set. **(3) Check-COVID** [104]: A COVID-19-specific fact-checking dataset, containing 347 abstracts from CORD-19 journal articles and 1,504 expert annotated news-related claims.

MultiVerS is trained on the SciFact train set to create a verifier model that provides verification feedback. Since the SciFact test set is inaccessible, we evaluate document evidence retrieval on SciFact-Open and full Check-COVID.

Results. The integration of verification feedback consistently enhanced document evidence retrieval, as +Verification outperformed monoT5-3B across nearly all cut-off thresholds in both SciFact-Open and Check-COVID (Table 1). The improvements are particularly evident at lower cut-offs, where retrieving the most relevant evidence is crucial. In SciFact-Open, Recall@5 and Recall@3 increased from 57.17% to 62.61% and from 48.26% to 52.83%, respectively. Similarly, in Check-COVID, these metrics improved from 82.06% to 84.04% and from 74.93% to 77.80%. These improvements are particularly meaningful given that the average number of gold evidence documents per claim is 1 in Check-COVID and approximately 2 in SciFact-Open. To illustrate this improvement, we conducted a case study examining how different retrieval methods ranked specific evidence documents (Table 2). Compared to monoT5-3B, +Verification successfully elevated the ranks of E2, E3 and E5, retrieving three additional pieces of evidence within the top 5 results. Notably, E5, ranked 302nd by monoT5-3B, was effectively rescued by adding verification feedback in +Verification, demonstrating the value of integrating verification-informed retrieval signals.

These findings highlight a promising research direction: shifting from semantic-only retrieval towards evidence-aware retrieval, where retrieval models explicitly account for evidential value. Our results suggest that leveraging well-performing verification models can help refine retrieval systems by distinguishing between purely

semantic relevance and plausible evidential relevance among unannotated documents. Furthermore, to continually improve evidence-aware retrieval, we propose the development of tailored IR systems capable of identifying evidential information, thereby enhancing evidence retrieval for scientific fact-checking.

RD.1. Benchmark tailored IR system for fact-checking

The preliminary study presented in this work outlined a framework to enhance evidence retrieval beyond only semantic relevance. To overcome the limitations of existing IR systems in scientific fact-checking scenarios, it is imperative to develop specialised IR systems capable of handling the specific challenges of verification tasks. However, training an IR system on a single fact-checking dataset risks poor generalizability and potential overfitting, particularly due to data imbalance, a common issue in the relatively small datasets characteristic of scientific fact-checking [97, 116]. Furthermore, poor verification performance deteriorates retrieval accuracy, creating a vicious feedback loop that further degrades overall system effectiveness. A multi-pronged strategy could mitigate these challenges by *pooling verification signals* from various high-performing verifier models, *leveraging large-scale datasets* such as FEVER [90] to improve training robustness, and *providing a shared retrieval checkpoint* enable subsequent studies to fine-tune the model for specific scenarios or datasets while reducing training cost. Recent work [51, 74] has explored unified retrieval models for knowledge-intensive NLP tasks, focusing on retrieval quality and downstream task utility [73, 115], including question answering (QA) and fact-checking. Similarly, we propose a verification-driven IR system for evidence retrieval, which explicitly incorporates evidential informatics. This approach follows a two-step training paradigm: general pre-training on large, diverse datasets followed by domain-specific fine-tuning. This approach balances scalability and domain specificity, ensuring IR models are both robust across different contexts and highly effective in targeted fact-checking applications. Additionally, a corresponding benchmark should employ a diverse set of evaluation metrics beyond for the fact-checking task to ensure comprehensive assessment of performance within fact-checking [6]. These metrics could include *verification accuracy*, reflecting the downstream utility of retrieved evidence; *decision latency*, measuring the computational efficiency of retrieval models; and *robustness to real-world conditions* such as noisy data and incomplete evidence, to improve system resilience.

Integrating verification feedback into evidence retrieval improves relevance assessment beyond binary labels, enhancing retrieval performance. Future research should focus on developing a benchmark IR system tailored for fact-checking, incorporating fine-grained relevance labels and verification-driven retrieval models. A scalable pre-training and fine-tuning approach has the potential to improve retrieval robustness and generalizability thereby producing more accurate and efficient fact-checking systems.

3 Imbalanced resources of scientific topics

In existing general fact-checking datasets, such as FEVER which is based on Wikipedia, the distribution of gold evidence per claim is relatively even and sufficient. However, a significant imbalance of evidence is observed in scientific fact-checking. While the SciFact corpus (~5K documents) maintains a relatively balanced number of evidence documents per claim, this balance was disrupted when the dataset was expanded to create SciFact-Open (~500K documents). In this larger corpus, the majority of claims have none or only one piece of supporting evidence while others have over. Claims related to less-researched topics are generally associated with fewer scientific publications, as illustrated by the examples in Table 3.

Table 3: Sufficient-evidence claim and none-evidence claim examples in SciFact-Open. 'Evidence' is the number of evidence in SciFact-Open corpus and 'Entities' is the number of entities by searching bold-keyword in PubMed.

Claim	Evidence	Entities
Obesity is determined in part by genetic factors.	24	499k
LRBA controls CTLA - 4 expression.	0	0.27k

However, most fact-checking systems do not explicitly account for evidence imbalance. A common approach is to use a fixed retrieval cut-off (i.e., selecting a predefined number of top-ranked documents for verification). One of the most inefficient approaches is setting the cut-off equal to the maximum number of evidence per claim in the dataset, ensuring that all possible evidence is retrieved. This heuristic has not previously caused major issues since general fact-checking datasets contain a relatively balanced number of supporting documents per claim. However, the imbalance in SciFact-Open suggests that the simple approach may not be suitable for open-domain scientific fact-checking with two major drawbacks:

(1) **Inefficiency.** Although the maximum number of gold evidence documents in the SciFact-Open dataset is 24, less than one third of claims have more than two. Using the maximum number as a cut-off would be inefficient due to the large number of documents that would have to be processed by the verifier.

(2) **Inaccuracy.** Introducing irrelevant evidence into downstream verification degrades fact-checking performance, whether the noise is semantically related or completely random [76] (as discussed in Section 2).

To address these challenges in the current and future studies of scientific fact-checking, we proposed a research direction based on flexible cut-off strategy for retrieving evidence based on claim characteristics.

RD.2. Flexible cut-off for Retrieved Evidence

Ranked List Truncation (RLT) refers to the task of selecting an optimal prefix of a ranked list of retrieved documents, with the goal of balancing retrieval effectiveness and efficiency. Prior work explores both heuristic and learned approaches, using either relevance labels or features derived from score distributions to determine the cut-off point [10, 52, 59, 61, 103, 109]. A related line of work is stopping methods in technology-assisted review (TAR) [14, 15, 85], which aim to retrieve as much relevant information as possible while minimising the effort spent on examining irrelevant documents. Both approaches aim to optimise an expected metric over candidate cut positions, typically using metrics such as $F1@k$ or $recall@k$. The datasets used in these prior studies on RLT and stopping methods, such as CLEF and TREC [25–27, 33, 43–45], exhibit imbalance but

typically contain enough relevant items per query to support recall-based supervision and evaluation. This dependence on sufficient relevance labels becomes problematic in fact-checking scenarios, where gold evidence documents are typically rare, making both recall-based stopping and supervised RLT approaches unsuited to this problem.

To explore whether relevance score distributions indicate evidence sufficiency, we compare well-studied and less-studied claims from SciFact-Open. Following Wadden et. al. [101], claims with four or more gold evidence documents are considered to be well-studied and those with none to be less-studied. Using monoT5-3B, we compute several statistics over the ranked document scores, including first-document relevance, average and total scores, score decay, and the initial-to-final score ratio. Table 4 shows that well-studied claims tend to have higher top-ranked scores and sharper decay patterns, suggesting that relevance distributions may serve as indicators of evidence sufficiency.

Table 4: Statistical analysis of average relevance scores for well-studied and less-studied claims. 'I/F' ratio is Initial-to-Final ratio. 'Exp k' denotes the exponential decay factor k.

Metric	1st Doc	Mean	Sum	I/F ratio	Exp k
Less-	0.941	0.502	25.097	3.514	1.544
Well-	0.995	0.741	37.052	1.963	0.651

Based on these findings, one possible direction is to leverage existing techniques to estimate whether a claim is less-studied or well-studied. A prediction module could utilise statistical features of relevance distribution such as those presented in Table 4. In addition, metadata such as retrieved entity counts in PubMed (Table 3) can serve as auxiliary signals to refine the prediction. Claims predicted as less-studied – e.g., with low total relevance or steep score decay – may be assigned smaller cut-offs to reduce verification cost, while RLT and stopping techniques could be applied to well-studied claims where concentrated high scores suggest richer evidence.

While this naive strategy relies on heuristic features, it does not explicitly optimise verification performance. To address this, future approaches could explore learning a cut-off policy using feedback from the verification stage. Specifically, truncation points may be selected based on reward signals, such as whether the claim is correctly verified or the confidence of the verifier. This would bypass the need for relevance-labelled supervision, which is often infeasible in scientific fact-checking due to sparse annotations. Inspired by prior RLT and stopping method work, such a learned policy could optimise both efficiency and factual accuracy by aligning truncation decisions with downstream verification performance.

4 Time and Citation

General fact-checking evidence corpus such as Wikipedia and fact-checking websites, often lack sentence-level evidence timestamps, making it difficult to determine the original publish time of sentence evidence in verification and hindering the development of time-aware retrieval methods. Timeliness is important in fact-checking, but in science, evolving evidence makes outdated studies particularly problematic. For instance, early COVID-19 treatment studies were later refuted, and outdated evidence may lead to harmful decisions. This section discusses whether scientific publications are more suitable for time-aware fact-checking and explores possible ways to leverage their inherent temporal characteristics.

Table 5: Results of health QA task considering the different thresholds of the published time of literature [99]

Year	Precision	Recall	F1 score
≥2020	59.7	**60.3**	**58.7**
≥2018	59.6	58.0	57.9
≥2015	61.1	56.0	53.9
≥2010	63.4	55.6	52.8
≥2005	**68.1**	56.5	52.0
≥2000	66.1	56.8	51.8
≥1990	65.6	55.4	51.3
≥1980	64.2	54.7	50.0

4.1 Timeliness of evidence

Unlike general fact-checking, where historical and static facts remain unchanged, scientific knowledge continuously evolves. This fundamental difference necessitates time-aware retrieval in scientific fact-checking to ensure that retrieved evidence remain valid and reflective of the latest scientific consensus. A time-sensitive retrieval mechanism should prioritise recent publications to ensure that fact-checking systems incorporate the most up-to-date methodologies and factual updates. This is especially crucial in fields like healthcare, where relying on outdated information could lead to misleading conclusions or incorrect decisions. For example, during a rapidly evolving pandemic, a medical treatment initially considered effective might later be deemed unreliable. This section discusses the challenges and opportunities of integrating the evidence timestamp into scientific fact-checking.

Scientific fact-checking aims to find evidential information in the literature to verify a claim. Intuitively, considering outdated literature negatively affects verification. Research in healthcare question answering (QA) has demonstrated that time-aware retrieval improves system performance [99], as shown in Table 5. The F1 score of the healthcare QA system improves as the publication year of evidence documents becomes more recent. By extension, scientific fact-checking on a large-scale corpus may also suffer from incorporating outdated evidence. These findings highlight the need for time-aware filtering in scientific fact-checking systems to enhance reliability. Fact-checking datasets could explicitly incorporate timestamps as metadata to facilitate research into temporal relevance in retrieval and verification. A general fact-checking study [9] collected the 'timestamp of last update' of claims and evidential documents, allowing later research [8] to explore the impact of temporal data on verification. This study found that a time-aware system achieved a 15% improvement in macro F1 score, underscoring the importance of temporal information. However, this work was limited by: (1) *focusing on verification only*, where evidence documents had already been retrieved in a separate initial step, without ensuring that retrieval prioritised high-quality and temporally relevant evidence, and (2) *the use of fragmented evidence* –

the general fact-checking datasets usually only contain small snippets of web documents, omitting many important time expressions present in full texts.

Given these findings, it is essential to develop a fact-checking system that explicitly incorporates temporal awareness across both retrieval and verification stages to meet real-world applications.

RD.3. Time-Aware Retrieval and Verification

Traditional fact-checking approaches often handle conflicting evidence for a single claim by assigning neutral veracity labels, such as "mixture," "unproven," or "not enough information" [37, 97]. However, conflicting evidence often arises due to outdated studies included in the retrieval process, which introduces noise and adversely affects prediction accuracy [99]. This issue has been observed in healthcare QA systems, where outdated evidence degrades performance [99]. To address this issue, we propose a time-aware approach that incorporates temporal information into both retrieval and verification stages:

1.Retrieval Stage: Outdated evidence should be filtered or deprioritised during retrieval to ensure that the retrieved evidence set is temporally aligned with the latest scientific findings.

2.Verification Stage: After filtering by time-aware retrieval, all retrieved evidence should be considered but with differentiated weighting based on temporal relevance. Recent evidence should be prioritised through higher weights, while older evidence should serve as supplementary or contextual information rather than primary evidence.

By processing outdated evidence differently across retrieval and verification components, this direction explores how temporal information can reduce noise and improve the reliability of scientific fact-checking systems.

4.2 Indirect evidence through citation

General fact-checking faces a number of challenges when attempting to determine the timeliness of evidence: (1) *Lack of publication timestamp* [8]. Many fact-checking sources, such as Wikipedia and fact-checking websites, do not provide precise publication dates for individual sentences or paragraphs. Instead, they record only the last edited timestamp, which does not accurately reflect when a fact was first published. (2) *Tracking the origin of evidence.* Evidence is often copied or paraphrased across multiple sources, making it difficult to determine the original publication date of a statement. (3) *Search engine bias in retrieval.* Pre-established fact-checking datasets retrieve evidence using top-k search engine results, where ranking mechanisms may prioritise recent documents due to time-aware ranking biases. This can misrepresent the actual chronology of claims and lead to fragmented evidence, making time extraction unreliable.

While these challenges also affect scientific fact-checking, they can be naturally mitigated by leveraging full-text papers rather than abstract-only sources. The main advantages of using full-text papers include: (1) *Explicit publication metadata*: Each piece of literature has a clearly defined publication date, ensuring accurate temporal tracking. (2) *Citation tracing for indirect evidence*: Mandatory citation rules in academic publications facilitate source tracing, even when statements are referenced indirectly. (3) *Structured nature of academic papers*: provides an indication of the origin of statements. For example, evidence in the background section typically references prior studies, whereas those in the abstract or results sections represent findings from the current publication. Given these inherent advantages, incorporating timestamps as metadata offers a promising research direction for full-paper-based scientific fact-checking. Exploring the temporal dynamics of claims and evidence should further enhance retrieval accuracy. Additionally, citation tracking to trace the original source of paraphrased evidence, ensuring the first-published timestamp is accurately recorded.

Citation-based tracking provides a promising approach to estimating evidence timestamps. However, the widespread presence of multiple citations in scientific literature makes it difficult to identify which references should be tracked, increasing computational costs and reducing efficiency. Moreover, indirect citations and paraphrased references, particularly in introductory sections, further obscure the retrieval of the first-published source. To address these issues, we propose the following research direction to develop a more effective approach for citation tracking and timestamp attribution.

RD.4. Citation-Based Evidence Tracking

Intuitively, self-contained evidence refers to information directly presented in the body of the current paper, while cited evidence is derived from external sources referenced by the paper. To explore this distinction, we analysed 22 accessible full papers out of 24 gold evidence for the claim in Table 2. We prompted GPT-4o to search for supporting/refuting evidence and determine whether each piece of evidence was paraphrased/summarised from a citation or is self-contained.

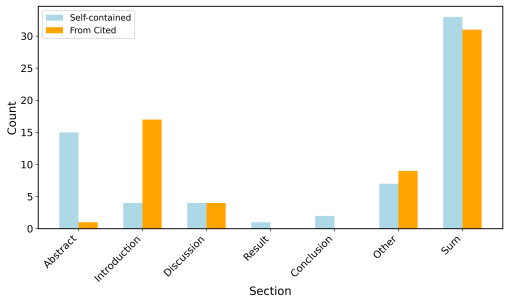


Figure 2: Evidence sources in scientific literature.

The result shown in Figure 2 reveals, as expected, that cited evidence is predominantly located in the introduction, while self-contained evidence is more common in the results and conclusion sections. While we observed a few inaccurate outcomes (e.g., one 'From Cited' evidence appearing in 'Abstract' is misjudged), the overall distribution remains discernible and interpretable. In addition, we also observed that sentences in scientific literature frequently contain multiple citations, making it costly to manually extract the original timestamp of cited evidence. To address this challenge, we propose a two-step approach:

1. Identify track-worthy citations. Not all citations are equally important for fact verification. Track-worthy citations should include: *conclusive evidence* directly influences claim verification and

plausible evidence that may impact claim assessment [108]. Since checking every cited document is expensive, an initial filtering step is required. A potential solution is to rank citations based on their relevance and function. Citation recommendation [30, 34] is a similar task to identify relevant publications for a given statement using retrieval models. Beyond relevance, the function of citation can be referred to as a signal to adjust priority. Citations are classified into eight categories: background, motivation, uses, extends, similarities, differences, compare/contrast, and future work [42]. While single evidence has multiple citations, the function of citation can help identify the citations that align with the evidence. For example, for predicted evidence 'Experiments show model A outperforms previous SOTA model B [citation 1,2,3]', cited papers for 'model A/B' in the 'introduction' function are less check-worthy than the citation in the 'compare/contrast' function. By prioritising high-impact citations, retrieval costs can be significantly reduced.

2. Track the original timestamp of evidence. Once trackworthy citations are identified, the next step is to trace the original timestamp of cited evidence via direct and indirect citation tracking. *Direct citation tracking* can be applied if the publication date of the cited paper is straightforward to retrieve. However, some evidence is paraphrased or indirectly cited, requiring a deep tracking mechanism to trace the citation path. Citation graph analysis [17, 96] can help map citation paths using directed graphs and applying search algorithms to identify the earliest relevant source. A recent study [118] found that reference errors – references do not include information to support statement – frequently appear ranging from 11% to 41% across domains. Addressing these errors introduces a sub-task for scientific fact-checking: verifying whether a cited reference truly supports the claim. To ensure feasibility, an early explorative study can assume that scientific literature generally adheres to citation conventions, preventing infinite citation loops in verification.

In summary, to explore time-aware fact-checking for scientific literature, we propose two potential research directions: (1) Integrating temporal information into retrieval and verification to handle outdated evidence and (2) Developing citation-based tracking methods to identify the original source and timestamp of evidence. These directions provide a basis for studying the impact of time-aware mechanisms in scientific fact-checking systems.

5 Structured Long-Context Evidence Retrieval

Existing scientific fact-checking datasets construct evidence corpora using fragmented sentences, paragraphs, or abstracts [16, 49, 62, 72, 75, 100, 104]. However, scientific literature is typically presented in structured, visually rich formats, often as PDF documents, where different sections serve distinct functions: abstracts, results, and conclusions summarise key findings, while background and introduction sections provide prior research context. With the diverse and unique functionalities of scientific literature components, this section explores challenges and potential research directions for advancing scientific fact-checking at the full-paper level. Existing fact-checking pipelines typically follow a document retrieval and then sentence selection paradigm [37, 97, 116]. For general fact-checking, evidence retrieval often uses top-ranked sentences from top-ranked documents, treating them as self-contained evidence

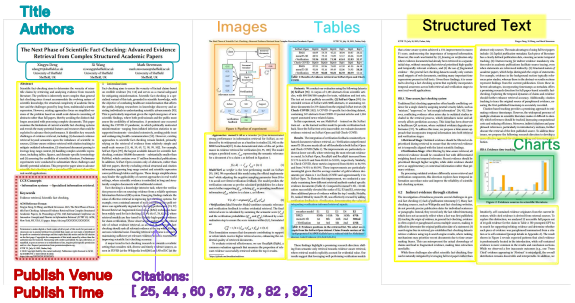


Figure 3: An example for parsing the scientific literature

units. However, this approach neglects long-range context, as using the extracted sentences ignores surrounding information to support verification. In contrast, scientific documents exhibit higher document-level consistency of verdict, commonly one paper having a sole standpoint to a given question, making document-level processing necessary for scientific fact-checking [50, 102, 119].

LLMs have recently demonstrated growing capability to process long contexts. However, it remains challenging to fact-check an entire full-text document in a single pass [105]. LLMs are prone to hallucinating content that is not grounded in the provided documents [2, 20, 23, 88] and are often susceptible to distraction from irrelevant context [82]. Their reasoning capabilities also degrade as text length increases [46]. A potential solution is context distillation, as used in retrieval-augmented generation (RAG), to filter high-quality context and mitigate hallucination, which improves the performance of downstream QA tasks [106]. However, unlike the QA task, fact-checking requires explicit retrieval of explicit. Filtering context may remove crucial supporting evidence, leading to incomplete verification. Moreover, much of the historical scientific literature exists in PDF format. Although recent multimodal LLMs are capable of consuming PDFs and conducting reasoning tasks, their capabilities are still limited to surface-level understanding. For example, they frequently fail to capture cross-page content and complex layout structures [60, 87, 95]. By prompting GPT-4o to locate evidence in PDF literature with the results presented in Section 4.2, we observed that it mislocated evidence in incorrect sections or non-existent sections. These limitations indicate that existing LLMs are not capable of supporting end-to-end fact-checking for long-context scientific literature – not only for raw PDFs but also for plain-text documents.

In addition to their length, scientific papers follow a structured format that introduces additional challenges for verification. Background and introduction sections often cite prior work, which may conflict with conclusions drawn later, while discussion sections highlight limitations that can cast doubt on earlier findings. These internally inconsistent signals may introduce misleading information, demanding reasoning that accounts for both the factual assertions and the functional roles of different sections. Parsing scientific documents into structured units enables a modular pipeline where retrieval and verification can be independently optimised. This modular structure facilitates interpretability, robustness, and denoising of conflicting or irrelevant content. Layout-aware tools offer a practical foundation for such structure-aware processing [19, 32, 41, 56, 81], as illustrated in Figure 3.

To address these challenges, we advocate for a retrieval framework that explicitly considers the document structure commonly found within scientific literature. Such a system should identify targeted evidence, capture long-range context across sections, and suppress irrelevant or conflicting content. We next outline a direction toward adaptive, section-aware retrieval strategies designed to meet these requirements.

RD.5. Adaptive Section-Aware Evidence Retrieval

Scientific literature follows a structured format where different sections serve distinct functions, presenting challenges for traditional evidence retrieval and verification in fact-checking systems. Existing retrieval methods often operate at the sentence or paragraph level, neglecting the long-range context and the structured nature of scientific documents. Additionally, large language models (LLMs) struggle to accurately associate claims with the appropriate sections, leading to potential misinterpretations and inconsistencies.

A promising research direction is **adaptive section-aware evidence retrieval**, which dynamically adjusts retrieval strategies based on document structure and claim types. This approach consists of two key components:

Claim-evidence matching. Claims should first be matched to the most relevant sections. For instance, experiment-driven claims, such as "X method improves accuracy compared to Y," should primarily retrieve evidence from the 'Results' and 'Conclusion' sections, as these contain empirical findings. In contrast, background or theoretical claims, such as "X method is widely used in Y applications," should focus on the Introduction and Background sections, which provide foundational knowledge. Prioritising section-aware evidence retrieval helps filter irrelevant context and reduces retrieval noise.

Contextual expansion. After retrieving primary evidence, the system should augment it with relevant contextual information from other sections to improve interpretability. For example, methodological details from the 'Analysis' section can provide additional support for experimental claims, while historical context from the Background section can clarify theoretical claims. Access to the full document allows for retrieving finer-grained evidence or complementary details that fragmented approaches may overlook.

In summary, an effective adaptive section-aware retrieval system must overcome challenges in accurately parsing document structures, prioritising relevant sections based on claim types, and efficiently handling conflicting evidence. By integrating structured document parsing, hierarchical retrieval strategies, and context-aware reasoning, future systems can leverage richer evidence from full-paper scientific literature while reducing LLM hallucinations. Advancing these techniques will enhance the reliability and interpretability of scientific fact-checking systems.

6 Multimodal content in Science

Scientific literature often conveys key evidence using non-textual elements such as tables, charts, and figures, which are commonly used to present experimental results, statistical analyses, and theoretical models, as shown in Figure 3. For full-text scientific fact-checking, especially across various fields of science and technology, it is crucial to move beyond text and accurately interpret these structured elements to ensure comprehensive verification. Fact-checking and misinformation detection on individual modalities – such as figures [1, 65, 66, 69, 93, 112], charts [4, 7], and tables [3, 13, 22, 29, 36, 57, 79, 83, 113] – has been studied independently [5]. To improve tabular reasoning, transformer-based approaches such as TAPAS [38] and Table-BERT [114] have been developed. However, these techniques perform poorly in SCITAB [57], a dataset for scientific fact-checking on tables, with results barely above random. One possible cause is the lack of contextual grounding for tables, which are rarely self-contained. Similarly, figures and charts in scientific papers often require surrounding textual explanations for correct interpretation. Recent datasets like AVerImaTeC [18] address image-text verification using web-sourced data, but scientific domains present greater challenges: figures are densely structured, often span multiple sections, and require domain-specific understanding capability.

We argue that scientific fact-checking should shift toward full-paper analysis, where structured elements are interpreted alongside their textual context. Unlike standalone multimodal models, document-level processing enables cross-referencing between figures/tables/charts and their descriptions, facilitating more faithful and complete verification.

RD.6. Multi-modal Evidence Alignment

Scientific fact-checking requires integrating evidence across multiple modalities, including text, tables, and figures, to ensure consistency and completeness. Recent multimodal information retrieval datasets in the scientific domain [71, 110] provide aligned pairs of textual and structured content, offering a foundation for cross-modal reasoning. However, in real scientific documents, structured elements, such as figures and tables, are not always located close to their descriptive text, making alignment a non-trivial challenge.

A promising direction is to explicitly align structured elements with their corresponding textual explanations within the same document. In scientific articles, tables and figures are typically explained through captions or surrounding sentences. Layout-aware parsing techniques can help identify these elements and link them to relevant text spans. Once aligned, their contents can be jointly encoded, enabling claim verification that draws on both structured data and contextual text. This alignment facilitates more coherent retrieval and reasoning across modalities, improving the reliability of multimodal fact-checking.

This unified framework contrasts with traditional multimodal systems that process each modality in isolation. Effective alignment demands progress in scientific document understanding [24, 60], visual structure parsing [32], and domain-specific retrieval [71]. Integrating these efforts will support full-document, multimodal verification, where structured evidence is faithfully grounded in its textual context.

7 Credibility of scientific literature

Existing evidence retrieval models often favour high-ranking documents based on semantic relevance, often overlooking scientific rigour [98]. As a result, low-quality documents may be retrieved as evidence, undermining the credibility of scientific fact-checking. The reliability of a claim verification process is inherently tied to

the quality of the supporting literature, making evidence credibility a crucial factor in scientific fact-checking.

In the domain of scientific literature, credibility assessment is influenced by multiple factors, including *peer-review status*, which ensures methodological scrutiny, *citation impact* which indicates how influential a study is within its field, and *experimental rigour*, reflecting the robustness of a study's methodology. However, the proliferation of non-peer-reviewed manuscripts and publications from venues with varying editorial standards, particularly in open-access repositories, poses a growing challenge. Such sources may lack the rigorous methodological scrutiny necessary to ensure reliable scientific conclusions. Therefore, incorporating additional quality indicators is essential for enhancing the robustness of scientific fact-checking.

RD.7. Extending Indicators of Evidence Quality

Evaluating scientific literature quality extends beyond content reliability, and should consider factors including venue reputation, methodological rigour, and experimental transparency. High-impact journals and prestigious conferences generally enforce stringent peer-review standards, contributing to the credibility of published research. Similarly, the expertise and prior contributions of an author, particularly in reputable venues, can provide further insight into the credibility of a study. In addition to traditional metadata, emerging indicators such as replication status, data availability, and adherence to reporting guidelines can further reflect methodological soundness. Although metadata-based credibility assessment is a useful heuristic, it is not foolproof. For example, selective reporting and statistical manipulation still exist, as some widely cited studies have later been retracted due to methodological flaws [31]. While the integration of such metadata remains a reasonable approach, as these indicators generally correlate with literature quality, their limitations must be acknowledged, given that even widely cited studies can occasionally be subject to retraction due to undetected methodological flaws.

8 Scientific terminology complexity

Scientific fact-checking systems often encounter challenges when aligning claims with supporting evidence due to mismatches in terminology granularity [101, 107]. The prevalence of hierarchical and synonymous scientific terms introduces significant challenges in fact verification. Many concepts exist at multiple levels of specificity, where broader categories encompass more specific subtypes, leading to ambiguity in claim-evidence alignment. This issue is further exacerbated by high token-level similarity among related terms, making it difficult for models to differentiate between general and specific concepts. As a result, models often misinterpret evidence relevance, increasing the likelihood of incorrect verification outcomes.

This issue arises when a claim uses a broad term, while the supporting evidence provides a more specific instance, or vice versa. As in the following example, such mismatches can lead to incorrect veracity assignments, as existing models struggle to recognise hierarchical relationships between concepts.

Claim: **Cancer** risk is lower in individuals with a history of alcohol consumption.

Supports: Alcohol consumption was associated with a decreased risk of **thyroid cancer**.

This issue is common within scientific fact checking and has been reported to occur within 44% of annotated examples in SciFact-Open [101]. Hence, we argue that capturing the hierarchical relationship could be a research direction to improve verification performance in the scientific domain, by solving the mismatch problem.

RD.8. Hierarchical Concept Modelling

To address this issue, ontology-based reasoning can be integrated into fact-checking pipelines. Structured ontologies such as MeSH [54] and UMLS [53] define hierarchical relationships that help systems infer term specificity. Recognising that 'lung cancer' is a subtype of 'cancer' enables better claim-evidence alignment, mitigating errors caused by lexical similarity. Beyond that, the knowledge graph can enrich ontological reasoning by encoding both hierarchical and associative relationships among scientific concepts [39, 48]. However, its potential for resolving terminology granularity mismatches in scientific fact-checking remains unexplored. Additionally, representation learning techniques such as contrastive learning can embed these hierarchical relationships into vector space representations, reducing reliance on token-level similarity. Domain-specific models like SciBERT [11] and PubMedBERT [35] can further enhance contextual understanding by incorporating structured knowledge into retrieval and verification processes. By leveraging ontological reasoning, knowledge graphs, and structured embeddings, scientific fact-checking systems can better align claims with relevant evidence, reducing verification errors caused by terminology granularity mismatches. In addition to enhancing precision, this also has potential to improve interpretability by making model decisions more transparent.

9 Conclusion

This paper explores the evolution of scientific fact-checking methodologies from abstract-level approaches to full-paper frameworks on large-scale corpora. As the volume and diversity of scientific knowledge continue to grow, the challenges of verifying claims across heterogeneous sources become increasingly complex. By addressing the complexities inherent in scientific literature, including its evolving nature, structured format, and the necessity for precise evidence retrieval, we underscore the importance of developing specialised retrieval systems capable of managing large, multimodal, and time-sensitive evidence retrieval. Furthermore, this work proposes several research directions aimed at improving scientific fact-checking efficiency and reliability. They include the integration of time-aware evidence retrieval to ensure the use of the most relevant and up-to-date findings, adaptive document processing to enable context-sensitive retrieval strategies, and multi-modal evidence alignment to integrate text, tables and figures to enhance verification accuracy. Overcoming these challenges has potential to improve the accuracy of fact-checking processes and also facilitate the scalability and applicability of these systems in diverse scientific fact-checking scenarios. By bridging the gap between scientific fact-checking and effective evidence retrieval, these advancements will contribute to more robust, interpretable, and trustworthy scientific fact-checking methodologies for real-world applications.

References

[1] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14940–14949.

[2] Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering. *Transactions of the Association for Computational Linguistics* 12 (2024), 681–699. doi:10.1162/tacl_a_00667

[3] Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2022. PubHealthTab: A Public Health Table-based Dataset for Evidence-based Fact Checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 1–16. doi:10.18653/v1/2022.findings-naacl.1

[4] Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and Reasoning over Chart Images for Evidence-based Automated Fact-Checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 399–414. doi:10.18653/v1/2023.findings-eacl.30

[5] Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal Automated Fact-Checking: A Survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5430–5448. doi:10.18653/v1/2023.findings-emnlp.361

[6] Mubashara Akhtar, Michael Schlichtkrull, and Andreas Vlachos. 2024. Ev2R: Evaluating Evidence Retrieval in Automated Fact-Checking. *arXiv preprint arXiv:2411.05375* (2024).

[7] Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. 2024. ChartCheck: Explainable Fact-Checking over Real-World Chart Images. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13921–13937. doi:10.18653/v1/2024.findings-acl.828

[8] Liesbeth Allein, Marlon Saelens, Ruben Cartuyvels, and Marie-Francine Moens. 2023. Implicit Temporal Reasoning for Evidence-Based Fact-Checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 176–189. doi:10.18653/v1/2023.findings-eacl.13

[9] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4685–4697. doi:10.18653/v1/D19-1475

[10] Dara Bahri, Yi Tay, Che Zheng, Donald Metzler, and Andrew Tomkins. 2020. Choppy: Cut transformer for ranked list truncation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1513–1516.

[11] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3615–3620. doi:10.18653/v1/D19-1371

[12] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150* (2020).

[13] Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. 2013. Methods for exploring and mining tables on wikipedia. In *Proceedings of the ACM SIGKDD workshop on interactive data exploration and analytics*. 18–26.

[14] Reem Bin-Hezam and Mark Stevenson. 2023. Combining Counting Processes and Classification Improves a Stopping Rule for Technology Assisted Review. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2603–2609. doi:10.18653/v1/2023.findings-emnlp.171

[15] Reem Bin-Hezam and Mark Stevenson. 2024. RLStop: A Reinforcement Learning Stopping Method for TAR. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2604–2608.

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

[17] Peter Buneman, Dennis Dosso, Matteo Lissandrini, and Gianmaria Silvello. 2021. Data citation and the citation graph. *Quantitative Science Studies* 2, 4 (2021), 1399–1422.

[18] Rui Cao, Zifeng Ding, Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2025. AVerImaTeC: A Dataset for Automatic Verification of Image-Text Claims with Evidence from the Web. arXiv:2505.17978 [cs.CL] https://arxiv.org/abs/2505.17978

[19] Catherine Chen, Zejiang Shen, Dan Klein, Gabriel Stanovsky, Doug Downey, and Kyle Lo. 2023. Are Layout-Infused Language Models Robust to Layout Distribution Shifts? A Case Study with Scientific Documents. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13345–13360. doi:10.18653/v1/2023.findings-acl.844

[20] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17754–17762.

[21] Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–5.

[22] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. TabFact : A Large-scale Dataset for Table-based Fact Verification. In *International Conference on Learning Representations (ICLR)*. Addis Ababa, Ethiopia.

[23] Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. 2023. The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 947–959. doi:10.18653/v1/2023.findings-acl.60

[24] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952* (2024).

[25] Gordon V Cormack and Maura R Grossman. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 153–162.

[26] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. arXiv:2102.07662 [cs.IR] https://arxiv.org/abs/2102.07662

[27] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820* (2020).

[28] Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim Verification in the Age of Large Language Models: A Survey. *arXiv preprint arXiv:2408.14317* (2024).

[29] Haoyu Dong and Zhiruo Wang. 2024. Large language models for tabular data: Progresses and future directions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2997–3000.

[30] Michael Färber and Adam Jatowt. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries* 21, 4 (2020), 375–405.

[31] Aaron HA Fletcher and Mark Stevenson. 2025. Predicting retracted research: a dataset and machine learning approaches. *Research Integrity and Peer Review* 10, 1 (2025), 1–10.

[32] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent skimming support for scientific papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 476–490.

[33] Maura R Grossman, Gordon V Cormack, and Adam Roegiest. 2016. TREC 2016 Total Recall Track Overview.. In *TREC*.

[34] Nianlong Gu, Yingqiang Gao, and Richard HR Hahnloser. 2022. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. In *European conference on information retrieval*. Springer, 274–288.

[35] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv:arXiv:2007.15779

[36] Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. PASTA: Table-Operations Aware Fact Verification via Sentence-Table Cloze Pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4971–4983. doi:10.18653/v1/2022.emnlp-main.331

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

[38] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4320–4333. doi:10.18653/v1/2020.acl-main.398

[39] Jason Hoelscher-Obermaier, Edward Stevinson, Valentin Stauber, Ivaylo Zhelev, Viktor Botev, Ronin Wu, and Jeremy Minton. 2022. Leveraging knowledge graphs to update scientific word embeddings using latent semantic imputation. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, Tirthankar Ghosal, Sergi Blanco-Cuaresma, Alberto Accomazzi, Robert M. Patton, Felix Grezes, and Thomas Allen (Eds.). Association for Computational Linguistics, Online, 43–53. doi:10.18653/v1/2022.wiesp-1.6

[40] Xuming Hu, Zhaochen Hong, Zhijiang Guo, Lijie Wen, and Philip Yu. 2023. Read it twice: Towards faithfully interpretable fact verification by revisiting evidence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2319–2323.

[41] Po-Wei Huang, Abhinav Ramesh Kashyap, Yanxia Qin, Yajing Yang, and Min-Yen Kan. 2022. Lightweight Contextual Logical Structure Recovery. In *Proceedings of the Third Workshop on Scholarly Document Processing*, Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Drahomira Herrmannova, Petr Knoth, Kyle Lo, Philipp Mayr, Michal Shmueli-Scheuer, Anita de Waard, and Lucy Lu Wang (Eds.). Association for Computational Linguistics, Gyeongju, Republic of Korea, 37–48. https://aclanthology.org/2022.sdp-1.5/

[42] Tomoki Ikoma and Shigeki Matsubara. 2023. On the Use of Language Models for Function Identification of Citations in Scholarly Papers. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, Tirthankar Ghosal, Felix Grezes, Thomas Allen, Kelly Lockhart, Alberto Accomazzi, and Sergi Blanco-Cuaresma (Eds.). Association for Computational Linguistics, Bali, Indonesia, 130–135. doi:10.18653/v1/2023.wiesp-1.15

[43] E. Kanoulas, Dan Li, Leif Azzopardi, and René Spijker. 2018. CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In *Conference and Labs of the Evaluation Forum*. https://api.semanticscholar.org/CorpusID:246440739

[44] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2018. CLEF 2018 technologically assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, Vol. 2125.

[45] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2019. CLEF 2019 technology assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, Vol. 2380. 250.

[46] Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One Thousand and One Pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17048–17085. doi:10.18653/v1/2024.emnlp-main.948

[47] Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. 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)*, Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 280–296. doi:10.18653/v1/2024.fever-1.29

[48] Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact Verification via Reasoning on Knowledge Graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 16190–16206. doi:10.18653/v1/2023.acl-long.895

[49] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7740–7754. doi:10.18653/v1/2020.emnlp-main.623

[50] Xiangci Li, Gully A Burns, and Nanyun Peng. 2021. A Paragraph-level Multi-task Learning Model for Scientific Fact-Verification.. In *SDU@ AAAI*.

[51] Xiaoxi Li, Zhicheng Dou, Yujia Zhou, and Fangchao Liu. 2024. Corpuslm: Towards a unified language model on corpus for knowledge-intensive tasks. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 26–37.

[52] Yen-Chieh Lien, Daniel Cohen, and W Bruce Croft. 2019. An assumption-free approach to the dynamic truncation of ranked lists. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. 79–82.

[53] Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of medical informatics* 2, 01 (1993), 41–51.

[54] Carolyn E Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88, 3 (2000), 265.

[55] Qi Liu, Bo Wang, Nan Wang, and Jiaxin Mao. 2025. Leveraging Passage Embeddings for Efficient Listwise Reranking with Large Language Models. In *Proceedings of the ACM on Web Conference 2025* (Sydney NSW, Australia) *(WWW '25)*. Association for Computing Machinery, New York, NY, USA, 4274–4283. doi:10.1145/3696410.3714554

[56] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu B. Kang, Egor Klevak, Bailey Kuehl, Michael J. Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2024. The Semantic Reader Project. *Commun. ACM* 67, 10 (Sept. 2024), 50–61. doi:10.1145/3659096

[57] Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A Challenging Benchmark for Compositional Reasoning and Claim Verification on Scientific Tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7787–7813. doi:10.18653/v1/2023.emnlp-main.483

[58] María Luengo and David García-Marín. 2020. The performance of truth: politicians, fact-checking journalism, and the struggle to tackle COVID-19 misinformation. *American Journal of Cultural Sociology* 8, 3 (2020), 405.

[59] Yixiao Ma, Qingyao Ai, Yueyue Wu, Yunqiu Shao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2022. Incorporating retrieval information into the truncation of ranking lists for better legal search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 438–448.

[60] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. MMLongBench-Doc: Benchmarking Long-context Document Understanding with Visualizations. arXiv:2407.01523 [cs.CV] https://arxiv.org/abs/2407.01523

[61] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Ranked list truncation for large language model-based re-ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 141–151.

[62] Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A Corpus of Fact-checked Biomedical COVID-19 Tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 244–257. https://aclanthology.org/2022.lrec-1.26

[63] Yuki Momii, Tetsuya Takiguchi, and Yasuo Ariki. 2024. RAG-Fusion Based Information Retrieval for Fact-Checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 47–54. doi:10.18653/v1/2024.fever-1.4

[64] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 708–718. doi:10.18653/v1/2020.findings-emnlp.63

[65] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. 2023. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*. 36–44.

[66] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024. VERITE: a Robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval* 13, 1 (2024), 4.

[67] Andrew Parry, Debasis Ganguly, and Manish Chandra. 2024. In-Context Learning" or: How I learned to stop worrying and love" Applied Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 14–25.

[68] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific Claim Verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, Eben Holderness, Antonio Jimeno Yepes, Alberto Lavelli, Anne-Lyse Minard, James Pustejovsky, and Fabio Rinaldi (Eds.). Association for Computational Linguistics, online, 94–103. https://aclanthology.org/2021.louhi-1.11/

[69] Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13052–13062.

[70] Vatsal Raina and Mark Gales. 2024. Question-Based Retrieval using Atomic Units for Enterprise RAG. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 219–233. doi:10.18653/v1/2024.fever-1.25

[71] Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. SciFIBench: Benchmarking Large Multimodal Models for Scientific Figure Interpretation. *arXiv preprint arXiv:2405.08807* (2024).

[72] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2116–2129. doi:10.18653/v1/2021.acl-long.165

[73] Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2395–2400.

[74] Alireza Salemi and Hamed Zamani. 2024. Towards a search engine for machines: Unified ranking for multiple retrieval-augmented large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 741–751.

[75] Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based Fact-Checking of Health-related Claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 3499–3512. doi:10.18653/v1/2021.findings-emnlp.297

[76] Artsiom Sauchuk, James Thorne, Alon Halevy, Nicola Tonellotto, and Fabrizio Silvestri. 2022. On the role of relevance in natural language processing tasks. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1785–1789.

[77] Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The Automated Verification of Textual Claims (AVeriTeC) Shared Task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1–26. doi:10.18653/v1/2024.fever-1.1

[78] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems* 36 (2024).

[79] Michael Sejr Schlichtkrull, Vladimir Karpukhin, Barlas Oguz, Mike Lewis, Wen-tau Yih, and Sebastian Riedel. 2021. Joint Verification and Reranking for Open Fact Checking Over Tables. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6787–6799. doi:10.18653/v1/2021.acl-long.529

[80] Özge Sevgili, Irina Nikishina, Seid Muhie Yimam, Martin Semmann, and Chris Biemann. 2024. UHH at AVeriTeC: RAG for Fact-Checking with Real-World Claims. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 55–63. doi:10.18653/v1/2024.fever-1.5

[81] Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S. Weld, and Doug Downey. 2022. VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups. *Transactions of the Association for Computational Linguistics* 10 (2022), 376–392. doi:10.1162/tacl_a_00466

[82] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*. PMLR, 31210–31227.

[83] Qi Shi, Yu Zhang, Qingyu Yin, and Ting Liu. 2021. Logic-level Evidence Retrieval and Graph-based Verification Network for Table-based Fact Verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 175–184. doi:10.18653/v1/2021.emnlp-main.16

[84] Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed Fact Checking using RAG and Few-Shot In-Context Learning with LLMs. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 91–98. doi:10.18653/v1/2024.fever-1.10

[85] Mark Stevenson and Reem Bin-Hezam. 2023. Stopping Methods for Technology-assisted Reviews Based on Point Processes. *ACM Transactions on Information Systems* 42, 3 (2023), 1–37.

[86] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14918–14937. doi:10.18653/v1/2023.emnlp-main.923

[87] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 13636–13645.

[88] Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 8818–8847. doi:10.18653/v1/2024.emnlp-main.499

[89] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=wCu6T5xFjeJ

[90] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 809–819. doi:10.18653/v1/N18-1074

[91] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The Fact Extraction and VERification (FEVER) Shared Task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal (Eds.). Association for Computational Linguistics, Brussels, Belgium, 1–9. doi:10.18653/v1/W18-5501

[92] Fangzheng Tian, Debasis Ganguly, and Craig Macdonald. 2025. Is Relevance Propagated from Retriever to Generator in RAG?. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part I* (Lucca, Italy). Springer-Verlag, Berlin, Heidelberg, 32–48. doi:10.1007/978-3-031-88708-6_3

[93] Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. 2024. "Image, Tell me your story!" Predicting the original meta-context of visual misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 7845–7864. doi:10.18653/v1/2024.emnlp-main.448

[94] Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. 2024. AIC CTU system at AVeriTeC: Re-framing automated fact-checking as a simple RAG task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 137–150. doi:10.18653/v1/2024.fever-1.16

[95] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19528–19540.

[96] Vijay Viswanathan, Graham Neubig, and Pengfei Liu. 2021. CitationIE: Leveraging the Citation Graph for Scientific Information Extraction. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 719–731. doi:10.18653/v1/2021.acl-long.59

[97] Juraj Vladika and Florian Matthes. 2023. Scientific Fact-Checking: A Survey of Resources and Approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki

(Eds.). Association for Computational Linguistics, Toronto, Canada, 6215–6230. doi:10.18653/v1/2023.findings-acl.387

[98] Juraj Vladika and Florian Matthes. 2024. Comparing Knowledge Sources for Open-Domain Scientific Claim Verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 2103–2114. https://aclanthology.org/2024.eacl-long.128/

[99] Juraj Vladika and Florian Matthes. 2024. Improving Health Question Answering with Reliable and Time-Aware Evidence Retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 4752–4763. doi:10.18653/v1/2024.findings-naacl.295

[100] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7534–7550. doi:10.18653/v1/2020.emnlp-main.609

[101] David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4719–4734. doi:10.18653/v1/2022.findings-emnlp.347

[102] David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 61–76. doi:10.18653/v1/2022.findings-naacl.6

[103] Dong Wang, Jianxin Li, Tianchen Zhu, Haoyi Zhou, Qishan Zhu, Yuxin Wen, and Hongming Piao. 2022. MtCut: A Multi-Task Framework for Ranked List Truncation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1054–1062.

[104] Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-COVID: Fact-Checking COVID-19 News Claims with Scientific Evidence. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14114–14127. doi:10.18653/v1/2023.findings-acl.888

[105] Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 14199–14230. doi:10.18653/v1/2024.findings-emnlp.830

[106] Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377* (2023).

[107] Jevin D West and Carl T Bergstrom. 2021. Misinformation in and about science. *Proceedings of the National Academy of Sciences* 118, 15 (2021), e1912444117.

[108] Dustin Wright and Isabelle Augenstein. 2021. CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1796–1807. doi:10.18653/v1/2021.findings-acl.157

[109] Chen Wu, Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2021. Learning to truncate ranked lists for information retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4453–4461.

[110] Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. 2024. SciMMIR: Benchmarking Scientific Multi-modal Information Retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12560–12574. doi:10.18653/v1/2024.findings-acl.746

[111] Amelie Wührl and Roman Klinger. 2022. Entity-based Claim Representation Improves Fact-Checking of Medical Content in Tweets. In *Proceedings of the 9th Workshop on Argument Mining*, Gabriella Lapesa, Jodi Schneider, Yohan Jo, and Sougata Saha (Eds.). International Conference on Computational Linguistics, Online and in Gyeongju, Republic of Korea, 187–198. https://aclanthology.org/2022.argmining-1.18/

[112] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. 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*. 2733–2743.

[113] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 174–184.

[114] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8413–8426. doi:10.18653/v1/2020.acl-main.745

[115] Hamed Zamani and Michael Bendersky. 2024. Stochastic rag: End-to-end retrieval-augmented generation through expected utility maximization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2641–2646.

[116] Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass* 15, 10 (2021), e12438.

[117] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. From Relevance to Utility: Evidence Retrieval with Feedback for Fact Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6373–6384. doi:10.18653/v1/2023.findings-emnlp.422

[118] Tianmai M Zhang and Neil F Abernethy. 2024. Detecting Reference Errors in Scientific Literature with Large Language Models. *arXiv preprint arXiv:2411.06101* (2024).

[119] Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3580–3586. doi:10.18653/v1/2021.emnlp-main.290

[120] Liwen Zheng, Chaozhuo Li, Xi Zhang, Yu-Ming Shang, Feiran Huang, and Haoran Jia. 2024. Evidence Retrieval is almost All You Need for Fact Verification. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 9274–9281. https://aclanthology.org/2024.findings-acl.551