A Systematic Survey of Claim Verification: Corpora, Systems, and Case Studies

Anonymous ACL submission

Abstract

Automated Claim Verification (CV), where claim's veracity is assessed against explicitly provided reference materials, is crucial in combating escalating online misinformation. This survey carefully analyzed 198 studies published between January 2022 and March 2025 to summarize recent work on corpus creation, system architectures, and the integration of large language models. We also conducted two case studies: the first looks at the relationship between claims and references. The second examines issues in claim decomposition. Our findings illuminate common corpus construction strategies and emerging trends in system architectures while highlighting remaining challenges in CV research.

1 Introduction

The growing scale of online misinformation has led to a surge of research in automated fact-checking and claim verification, which assess whether a given claim is supported by accompanying references. A key milestone in this field was the release of the FEVER dataset (Thorne et al., 2018), which formalized claim verification as a benchmark task and sparked the development of new datasets such as Xfever (Chang et al., 2023), FEVEROUS (Aly et al., 2021) and many more. Since then, shared tasks like AVeriTeC (Schlichtkrull et al., 2024) have further advanced research by providing standardized datasets and evaluation frameworks for verifying claims against textual evidence.

Many recent surveys have reviewed system designs of claim verification from different angles, including system overviews (Bhuiyan et al., 2025; Guo et al., 2022; Yang et al., 2024), justification generation (Eldifrawi et al., 2024), LLM integration (Dmonte et al., 2024), and multimodal approaches (Akhtar et al., 2023b). Several surveys touch upon some elements in datasets such as size, input, and output format (Yang et al., 2024; Panchendrarajan and Zubiaga, 2024; Gusdevi et al., 2024), but few have examined the corpora creation process and its impact on system design. We fill this gap by providing a review of recent corpus-creation practices, together with system design across key components.

In this study, we conduct a systematic survey of claim verification (CV) research in order to answer the following research questions: (1) What corpora are available for CV research and how are they created? (2) What are common approaches in building CV systems? (3) What are the main issues and challenges in corpus construction and system development and what are some future directions to address the issues? We will answer the first two questions in Section 4-5 and the last question in Section 6-8 with two case studies.

2 Task Setting

The input to a CV system consists of a **claim** and optionally some reference documents. The documents are sometimes called *evidence* or *context*. In this study we will call them **reference documents** or **reference** in short, and use the term *evidence-bearing sentence* to refer to evidence in the reference. The output of a CV system includes a **veracity label** and optionally a **justification** to explain the veracity label.

The task has two settings. In the first (also called *open-domain fact-checking*), only a claim is provided as input and the CV system needs to retrieve relevant documents from external sources such as the Internet. In the second setting, the reference documents are provided as input. In this survey, we focus on the latter as we will study the relationship between claims and references and its effect on corpus creation and system development.

3 Paper Selection

To ground our analysis, we first collected a set of research papers on claim verification.

3.1 The initial set of papers

We collected papers from three main sources: ACL Anthology[1], Semantic Scholar[2], and Google Scholar[3]. We used query terms *(fact OR claim) AND (checking OR verification)* to retrieve papers published between Jan 2022 and March 2025.[4] After removing duplicates, there were 315 papers left, forming our initial set of papers.

3.2 Manual Screening and Categorization

We read all the 316 papers and divide them into three groups: (a) 62 papers that are not on CV; (b) 56 papers are on the first CV setting (i.e., references are not provided); (c) 198 papers on the second setting of CV, forming the **main collection** of studies covered in this survey.

For the rest of the paper, we will report findings from our main collection, but we will mention important work published before 2022 and papers from (b) when appropriate.

4 Corpus Creation

Out of 198 papers in the main paper collection for this survey, 65 created new CV corpora. Among them, 47 focus on corpus construction while the remaining 18 are on system development but have built new corpora for evaluation. In this section, we report findings from these 65 papers.

4.1 Main components of a CV corpus

An instance in a CV corpora consists of a claim, a reference, a veracity label, and very often a justification. In addition, it may include some metadata such as author name, publication date, and publication platform of the claim or the reference.

Claim: A claim is a statement being verified. In almost all corpora in our collection, a claim is text, but there exist several corpora with multi-modal claims such as FACTIFY (Mishra et al., 2022), FACTIFY 2 (Suryavardan et al., 2023), and ClaimReview2024+ (Braun et al., 2024). For instance, a claim can be a (text, image) pair, extracted from public websites such as Twitter.

[1]https://aclanthology.org/

[2]https://www.semanticscholar.org/

[3]https://scholar.google.com/, using SerpAPI

[4]Appendix A provides details of our scraping setup.

Reference: A claim is verified against some reference documents. While references in most corpora in our collection are text (e.g., paragraphs or documents), 12 corpora go beyond text and use images (e.g., (Yao et al., 2022; Mishra et al., 2022; Rangapur et al., 2023; Braun et al., 2024; Chakraborty et al., 2023; Chen et al., 2024b)), charts (Akhtar et al., 2023a, 2024), tables (Akhtar et al., 2022; Yilun Zhao et al., 2024), or videos (Liu et al., 2023).

Veracity Label: Most CV corpora use three labels for veracity: *supported*, *refuted*, and *NEI (not enough information)*. Seventeen corpora use binary labels: *true* or *false*. The rest extend these label sets by adding labels such as *partially supported* (Li et al., 2024), *Conflicting evidence/cherry-picking* (Schlichtkrull et al., 2023), and *Misleading* (Braun et al., 2024). used by the FCTR dataset

Justification: Although justification is not a required field in a CV corpus, it provides explanation to the veracity label and majority of the corpora in our collection include justification. Common types of justification are *evidence-bearing sentences (EBS)* in the original reference (e.g., (Evans et al., 2023; Vladika et al., 2024)), summaries of the EBSs (Chakraborty et al., 2023), or other types such as free-form, deductive and argumentative explanation (e.g., (Cekinel et al., 2024; Chen et al., 2024b; Kotonya and Toni, 2024)).

4.2 Corpus properties

At the corpus level, 12 corpora have 1,000 or fewer instances, 20 have 1,000 to 10,000 instances, and the remaining 35 each have over 10,000 instances.

Modality Fifty-two corpora are text only and 13 corpora are multi-modal where their references include images, charts, tables, or videos. In FACTIFY (Mishra et al., 2022), FACTIFY 2 (Suryavardan et al., 2023), FACIFY3m (Chakraborty et al., 2023), and ClaimReview2024+ (Braun et al., 2024), both claims and references are (text, image) pairs. While the justification in all these corpora are text only, we believe there will be many use cases where multi-modal justification is beneficial (e.g., an image that marks errors in the claim or the reference).

Languages: The majority (50) of the corpora are English only, five are Chinese only (Hu et al., 2022; Lin et al., 2024; Zhang et al., 2024a,b; Wu et al., 2023), two are Vietnamese only (Hoa et al., 2024; Le et al., 2024), and one each in German (Deck et al., 2025), Italian (Scaiella et al., 2024), Indonesian (Muharram and Purwarianti, 2024), Czech (Ullrich et al., 2023) Arabic(Haouari et al., 2024) Bangla(Rahman et al., 2025) and Turkish (Cekinel et al., 2024). In addition, several corpora are multilingual (e.g., (Chang et al., 2023; Zeng et al., 2024; Chung et al., 2025; Pikuliak et al., 2023)).

Domain: Data in the CV corpora come from various domains, such as politics (e.g., (Zeng et al., 2024; Nanekhan et al., 2025; Suryavardan et al., 2023), health (e.g., (Vladika et al., 2024; Akhtar et al., 2022; Gupta et al., 2023; Liu et al., 2023)), science and technology (e.g., (Wadden et al., 2022; Lu et al., 2023; Fu et al., 2024)), and finance (e.g., (Yilun Zhao et al., 2024; Rangapur et al., 2023)). Majority of corpora collect data from multiple domains as Wikipedia is a major source (e.g, (Lin et al., 2024; Ma et al., 2024; Kamoi et al., 2023)).

4.3 Corpus Construction Approaches

CV corpora are rarely built from scratch; they are often built on existing datasets. Each of the four main components (namely, claim, reference, veracity label, and justification) is (1) inherited from existing datasets, (2) created or modified manually by annotators, or (3) generated by NLP systems. Often multiple methods are applied; for instance, claims in FEVERFact (Ullrich et al., 2025) originated from a Wikipedia page, then were modified by systems, and finally checked by annotators.

Based on whether claims and references existed before corpus construction, there are three common scenarios. First, both claims and references (and even veracity labels) came from datasets. They are cleaned, transformed and extended to form a new CV corpus. For instance, Xfever (Chang et al., 2023) translated the claims and the references in the FEVER dataset (Thorne et al., 2018) from English into five languages to form a multi-lingual corpus. LIAR++ (Russo et al., 2023) started from the LIAR-PLUS dataset (Alhindi et al., 2018).

In the second scenario, claims were pre-existing (e.g., ones made by podcasters). To acquire references, one can retrieve documents with claim-based queries and then filter out irrelevant ones (e.g., (Schlichtkrull et al., 2023; Wadden et al., 2022; Vladika et al., 2024)).

In the third scenario, references are from existing sources such as Wikipedia; claims are generated from the references by humans or systems. In FEVER (Thorne et al., 2018), claims are human-generated by paraphrasing or distorting sentences from Wikipedia to create factual, refuted, or unverifiable statements. Many corpora (e.g., (Diggelmann et al., 2020; Wadden et al., 2022; Jiang et al., 2020)) follow this paradigm. An example is in Appendix B.

For quality control, human inspection and automatic evaluation are conducted at the instance level and the component level with measures such as inter-annotator agreement on veracity labels and ROUGE scores for summaries as justification.

5 System Development

Of the 198 papers in our survey, 156 build or evaluate CV systems.

5.1 The traditional pipeline

The traditional CV systems has four steps.

Document Selection/Evidence Retrieval: This initial step (done by 76 papers) focuses on identifying the most relevant documents or passages for the claim. Recent work emphasizes robust retrieval through methods like multi-stage reranking (Malviya and Katsigiannis, 2024), specialized extraction pipelines (Wuehrl et al., 2023), and sophisticated question enrichment strategies (Churina et al., 2024).

Sentence Selection/Ranking: From the retrieved documents, sentences or snippets pertinent to the claim are selected (68/156 papers). Hu et al. (2023) proposed a latent variable model for better sentence retrieval. (Zheng et al., 2024) demonstrated the importance of accurate evidence retrieval.

Veracity Label Prediction: Considered the core of claim verification (144 papers), this step involves predicting a veracity label based on selected sentences. Recently there is a shift from traditional supervised classifiers to LLMs (Guan et al., 2024; Li et al., 2024; Zeng and Gao, 2023; Zhang and Gao, 2023), which often combine retrieved evidence with instruction-tuned prompting (Alvarez et al., 2024).

Justification Generation: Many systems (56 papers) now generate justification. Extractive approaches use retrieved evidence snippets (Wadden et al., 2022; Vladika et al., 2024), while abstractive methods generate new textual explanations, often using LLMs (Zarharan et al., 2024).

5.2 Other Strategies

In addition to the traditional pipeline, other strategies have been proposed for building CV systems. Below we summarize several common strategies.

Decomposition. As an alternative, recent systems decompose complex claims into subquestions or subclaims (Chen et al., 2024a; Sahu et al., 2024; Schlichtkrull et al., 2023; Kamoi et al., 2023). Liu et al. (2024a) employ "Claim Split" modules for this, guiding targeted verification questions (Xu et al., 2024). However, such atomic units risk losing essential context and they may become ambiguous or unverifiable (Hu et al., 2024). (Gunjal and Durrett, 2024) directly tackles this, defining criteria like decontextuality (ensuring unique specification for stand-alone status) and minimality (adding only essential context). We will examine decomposition more in Section 7.

Temporal Reasoning. Claims that mention dates or event order require temporal consistency checks (Mori et al., 2022). Barik et al. (2024a) extracts event–time pairs from both claim and evidence and aligns them on a shared timeline. Barik et al. (2024b) adds a rule-based filter that discards evidence outside the relevant time window.

Knowledge Graph-Based Reasoning. Graph structures are used to model relationships between evidence and claims (Kim et al., 2023; Lin and Fu, 2022; Lan et al., 2025), enabling reasoning over interconnected facts. In this approach, claims and evidence are represented as nodes (e.g., entities, facts), and verification is framed as graph traversal or subgraph matching (Lin and Fu, 2022).

Iterative self-revision and flaw identification. A newer trend equips verifiers with a "quality-control" loop, where systems self-revise an initial veracity and explanation before user presentation. These extra verification loops improve factual alignment and explanation quality compared to single-shot pipelines. For instance, Zhang et al. (2024b) let GPT-4 provide initial explanations, which a second LLM then scans and revises until fully citation-backed. Kao and Yen (2024a) train a module to detect rhetorical fallacies (e.g., cherry-picking) and apply fallacy-specific corrections.

5.3 Evaluation practices

Claim verification systems are typically evaluated using standard metrics such as accuracy and F1 scores (Nguyen et al., 2025; Bazaga et al., 2023; Zeng and Zubiaga, 2022). For datasets like FEVER (Thorne et al., 2018), FEVEROUS (Aly et al., 2021), and AVeriTeC (Schlichtkrull et al., 2024), a stricter FEVER-style score is used, which requires both the correct label and at least one complete evidence set (Gong et al., 2024; DeHaven and Scott, 2023; Zheng et al., 2024; Liu et al., 2024b).

Extractive justifications are evaluated by measuring precision, recall and F1(Krishna et al., 2022). Abstractive justifications rely on n-gram overlap metrics such as BLEU and ROUGE alongside semantic similarity scores like BERTScore (Zhang et al., 2024b,c; Yao et al., 2022).

6 Case Study #1: Claim and Reference

Claims in early CV corpora were typically based on single documents. For example, 87% of the claims in the FEVER dataset (Thorne et al., 2018) are supported by evidence from one single article, and in many cases, verification relies on a single sentence within that article. This contrasts with real-world scenarios where verifying a claim often requires synthesizing information from multiple sources and multiple pieces of evidence (Ma et al., 2024). In this case study, we aim to investigate the number of evidence-bearing sentences (EBSs) needed to verify a claim.

6.1 Case Study Design

To that end, we randomly sampled 3 corpora - MSVEC (Evans et al., 2023), HealthFC (Vladika et al., 2024), WiCE (Kamoi et al., 2023) - from 12 corpora in which the justifications include multiple EBSs. Figure 1 shows the distribution of the number of EBSs per claim. Notably, in MSVEC, 19.6% of claims have only one EBS in justification. Among the instances in which the number of EBSs is greater than one, we want determine how many of the gold-standard EBSs are truly needed to verify the claim. To answer this question, we randomly sampled from HealthFC (Vladika et al., 2024) 50 instances that have more than one EBSs, for manual analysis.

We examined every (claim, veracity, EBS) triple in our samples and found that EBSs sometimes fail to support the veracity label. We identify six types of scenarios for the triples based on whether an EBS justifies the veracity label given to a claim. We provide full examples of those scenarios in Appendix C.

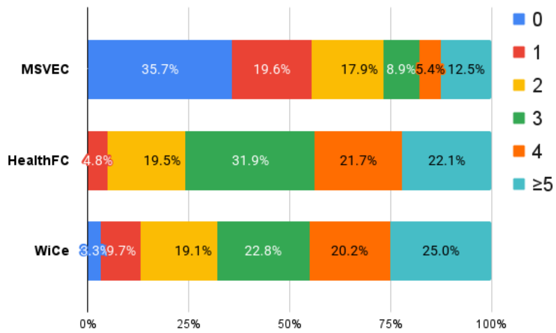


Figure 1: The distribution of the number of EBSs per claim in three corpora

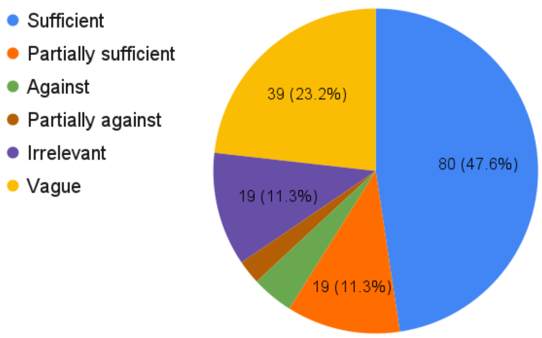


Figure 2: The distribution of six types of relations for claim-veracity-EBS triples. The raw count for Against and Partially against are 7 and 4.

Sufficient: The EBS alone is sufficient to justify the veracity label.

Partially sufficient: The EBS contributes to the veracity label but is not sufficient by itself.

Against: The EBS is against the veracity label and is sufficient for a different label.

Partially against: The EBS is partially sufficient for a different label.

Irrelevant: The EBS is unrelated to the claim.

Vague: It is not clear whether the EBS is related to the claim due to some ambiguity (e.g., due to unsolved coreference).

6.2 Results and Issues

Among the 168 EBSs in the 50 instances, 99 EBSs are sufficient for or contribute to justifying the veracity label; 39 are vague, most of which are due to coreference issues; surprisingly, we have found 7 EBSs directly against the assigned label, supporting a different label. Figure 2 shows the full distribution.

Overall, we agree with the veracity labels in 38 instances. Among them, 35 claims need only one EBS to fully justify the assigned label; the other 3 need a combination of two EBSs. We disagree with the label assigned for the remaining 12 instances: either the EBSs are supporting a different label (2) or the EBSs are not useful for assigning any labels due to contradictory information (4) or irrelevant and vague EBS (6). For each of these cases, we provide detailed examples in Appendix C.

6.3 Discussion

As discussed earlier, not only do some EBSs fail to support the assigned label, but they can also actively suggest a different one. In our sample of 50 instances, we disagreed with nearly a third of the assigned labels. To address this, we suggest improving both annotation guidance and label design.

More study is needed to categorize claim types and understand their annotation needs. For instance, when claims contain qualitative judgments, but the supporting evidence is quantitative, disagreements can easily occur. Many claims in our sample involve subjective interpretations. For example, one claim asks, "Do health benefits increase with the duration and intensity of exercise?" One EBS states, "Compared to inactive people, slight activity prolongs life by 0.7 year." However, is a 0.7-year increase considered significant or minimal? This ambiguity can lead to inconsistent annotations. In cases like this, domain-specific guidance and clearly defined criteria for interpreting evidence would help align annotators' decisions. Moreover, the design of veracity labels should reflect the complexity of real-world data. In this case study, 10 instances could not be mapped to any of the predefined labels. Adding categories like "contradictory" or "irrelevant" could better capture these edge cases.

7 Case Study #2: Claims and Subclaims

As discussed in section 5.2, a common pattern in LLM-driven fact verification is the *Decompose-Then-Verify* paradigm, where complex claims are split into simpler subclaims before verification. While this modular approach improves scalability and interpretability, the quality of decomposition remains a key bottleneck (Hu et al., 2024). Ideally, subclaims should be semantically equivalent to the original claim. In this case study, we examine common decomposition strategies and associated

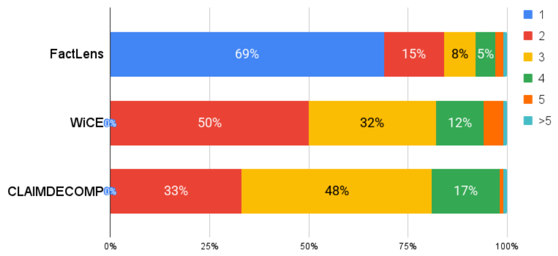


Figure 3: The distribution of number of subclaims in each dataset. For CLAIMDECOMP and WiCE, training dataset is used; for FactLens, whole dataset is used.

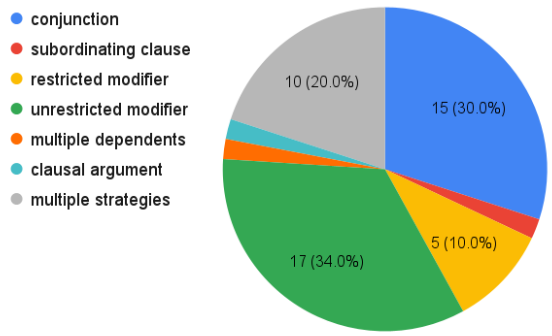


Figure 4: The distribution of strategies we observed in the sample. More detailed analysis of the strategies is shown in the Appendix D.

issues.

7.1 Case study design

To investigate these questions, we reviewed existing corpora that provide aligned claim–subclaim structures and identified three publicly available datasets: CLAIMDECOMP (Chen et al., 2024a), WICE (Kamoi et al., 2023) and FACTLENS (Mitra et al., 2024). In all three datasets, LLMs are used to generate subclaims from complex claims via prompting, followed by human evaluation to ensure the quality of decomposition.

Figure 3 shows the distribution of subclaims per claim. We excluded instances with only one subclaim in FACTLENS, as they do not reflect true decomposition. Across datasets, most claims are decomposed into two or three subclaims, reflecting a tendency toward minimal yet tractable breakdowns.

We then randomly sampled 50 decomposable claims from FACTLENS. For each, we examined the generated subclaims, annotated the decomposition strategy (Section 7.2), and assessed whether the subclaims (1) entailed the original claim and (2) introduced any decomposition errors.

7.2 Common patterns and issues

Based on our analysis of 50 decomposed claims, we identified several recurring decomposition strategies, the distribution of which is shown in Figure 4. Here, we illustrate some of the common patterns and the corresponding issue using a representative example.

Consider the original claim:

> *"Mickey Mansell played in his second World Cup of Darts with Brendan Dolan, he reached the quarter-finals of a PDC event but lost in the UK Open which was held at the Reebok Stadium in Bolton."*

This was decomposed into the following subclaims:

SC1: *Mickey Mansell played in his second World Cup of Darts with Brendan Dolan.* SC2: *Mickey Mansell reached the quarter-finals of a PDC event.* SC3:*Mickey Mansell lost in the UK Open.* SC4: *The UK Open was held at the Reebok Stadium in Bolton.*

One issue with this decomposition is that the connector between SC1 and SC2 is lost. As a result, the temporal or causal relationship between events becomes ambiguous—it is unclear whether these events occurred in sequence, simultaneously, or are otherwise related. Furthermore, by isolating events into standalone subclaims, important contextual information such as temporal scope is stripped away. SC2, SC3, and SC4 all become difficult to verify in isolation, as they lack sufficient temporal anchoring to be accurately matched against the reference material. This observation highlights a broader implication: **context must be preserved when generating and verifying claims and subclaims**. In particular, contextual information should be part of the input to claim verification models, as it is often essential for determining whether a subclaim is truly supported by the evidence.

7.3 Results

Results are shown in Figure 5. Among the 50 sampled claim-subclaim pairs, five sets of subclaims did not entail the original claim, while nine entailed it, but were not semantically equivalent.

Similarly to what we have discussed in Section 7.2, 7 claims were ungrammatical, often formed by join-

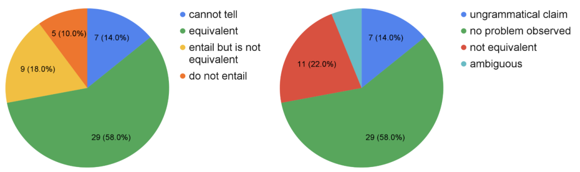


Figure 5: Summary of decomposition analysis. The left pie chart shows the distribution of entailment types in our annotated sample; the right pie chart summarizes the overall presence of problems.

ing two independent sentences without appropriate conjunctions or punctuation. Such cases introduce structural ambiguities that hinder both manual and automatic decomposition.

Accounting for all cases where (1) subclaims were not semantically equivalent, (2) decomposition introduced errors, or (3) the original claim was malformed, we estimate that 42% of the samples exhibit some form of decomposition failure. Given the increasing reliance on subclaim decomposition in fact verification pipelines, these quality issues raise concerns about the validity of this approach and its potential to negatively impact downstream verification performance.

8 Challenges and Future Directions

This study has revealed several issues with corpus creation and system development.

8.1 Issues with corpus creation

Context Dependency of Claim: Very few CV corpora in our survey provide context information to help resolve ambiguities in the claims. For instance, in order to verify the Mickey Mansell claim in our case study 2, we need to know which year the claim refers to, what *PDC* stands for, what was considered a *PDC event* in that year, and so on. As a result, some claims cannot be verified without additional information (Ousidhoum et al., 2022) Therefore, corpus designers should try to eliminate such ambiguities by changing their ways of generating claims or references or by adding context as a new component of the corpora.

Claim type and veracity label: Setty and Becker (Setty and Becker, 2025) created a dataset for fact-checking podcasts and categorized claims into four types of *Checkable* claims (i.e., factual descriptions, cause and effect, numerical claims, and quotations) and five types of *Not Checkable* claims.

Our survey shows that the large majority of CV corpora use binary or ternary veracity labels. For some claim types (e.g., numerical claims), more fine-grained label sets are needed, as discussed in Section 6.3. Thus, our field will benefit from more studies on claim types and veracity label sets and more detailed guidelines for veracity annotation.

Modality and language: As our survey shows, English is unsurprisingly the dominant language in CV corpora and text remains the most common modality. However, this dominance does not reflect the complexity of the real-world information ecosystem, where claims are made in many languages and supported by evidence drawn from what people read, hear, and watch. Expanding beyond English and text should be a collective priority in the field, encouraging the inclusion of multilingual and multimodal data to better align with real-world contexts.

8.2 Issues with system development

Multi-hop reasoning and decomposition: They are common strategies adopted in CV systems. As shown in Section 7, the decomposition process can be error-prone; e.g., the conjunction of subclaims might not be equivalent to the claim. Even when they are equivalent, some subclaims might be unverifiable based on the available references. Furthermore, some claims can be difficult to decompose. Thus, more studies are needed on when and how decomposition should be performed in the CV task.

Use of LLMs Nowadays many CV systems are built on top of LLMs. One issue is how LLMs' *prior knowledge* would affect their "judgment" of the claims, especially when the prior knowledge is in conflict with the information in the reference. Will LLMs be able to temporarily suspend its own prior knowledge when dealing such conflict? More studies are required to better understand LLMs' behavior.

Shared task, evaluation corpora and deployment The results of our survey, as well as the overall system designs observed in the field, are strongly shaped by the structure and requirements of shared tasks. For instance, the AVeriTeC shared task (Schlichtkrull et al., 2024) focuses not only on veracity accuracy, but also on evaluating the quality of questions and their corresponding answers generated from given claims. Consequently, all participating teams were incentivized to include a question generation component in their systems. Moreover, the task mandated evaluation of an intermediate step—sentence selection—even though our survey indicates that this step is not typically emphasized in standard claim verification pipelines. In other words, the specific design and evaluation criteria imposed by shared tasks like AVeriTeC significantly influence the development of systems in this subfield, often introducing components that would not otherwise be prioritized.

Similarly, the design of CV systems can be greatly affected by the choice of evaluation corpora; for instance, if the corpora were created by aggregating multiple evidence-bearing sentences, CV systems are more likely to "reverse engineer" by decomposing the claims.

As the ultimate goal of building CV systems is to deploy them to check real-world claims, more work is needed on facilitating the deployment efforts and testing system performance in real world.

9 Related Work

The release of the FEVER dataset (Thorne et al., 2018) marked a turning point in automated claim verification. Follow-up datasets like HoVer (Jiang et al., 2020) and EX-FEVER (Ma et al., 2024) introduced multi-hop reasoning and structured evidence. These resources inspired a variety of corpora today and spurred the development of pipeline systems that typically include document retrieval, sentence selection, and veracity prediction.

Our selection includes 8 surveys, which reviewed different aspects of claim verification. Many (Bhuiyan et al., 2025; Guo et al., 2022; Yang et al., 2024) provided an overview of the CV systems. Some adopted a more focus angle: Eldifrawi et al. (2024) specifically explored the methods on justification production generation; Dmonte et al. (2024) focuses exclusively on how LLMs are adopted into the CV system. Two surveys (Panchendrarajan and Zubiaga, 2024; Gusdevi et al., 2024) examined claim verification systems in non-English and region-specific contexts, whereas another (Akhtar et al., 2023b) focused on multimodal approaches. While these surveys touch on certain aspects of corpus creation like size, label, and annotation(Yang et al., 2024; Panchendrarajan and Zubiaga, 2024; Gusdevi et al., 2024), none provides a comprehensive analysis of how CV corpora are constructed.

Our work differs in scope and focus: we survey only tasks where both a claim and reference are present as input. Apart from synthesizing common system approaches, we provide a detailed account of how CV datasets are constructed to address a gap in existing surveys by foregrounding the role of dataset design in CV landscape. Furthermore, we conduct two case studies to explore the number of EBS used in verification and the quality of claim decomposition.

10 Conclusion

Our survey of 198 claim verification (CV) papers (January 2022 - March 2025) offers a novel fine-grained analysis of corpus creation, system design, and pipeline vulnerabilities investigated through two detailed case studies. We described common strategies and challenges in CV corpus construction, with our first case study highlighting the relationship between claims and references. For system development, we detailed the pipeline's evolution and emerging strategies like claim decomposition, where our second case study found various problems with decomposition.

While most studies in the NLP field focus on proposing novel systems, our findings underscore the need to better understand the data, as how the corpora were created can affect whether certain system design strategy would be effective. We hope this survey motivates future research to apply new techniques with critical awareness of these identified issues. Future research directions include developing corpora with richer context, ensuring LLM faithfulness to reference materials, and expanding into multilingual and multimodal claim verification.

Limitations

This survey included only papers in English published from Jan 2022 to March 2025, and thus may have missed studies published in other languages or outside this time period.

Due to the large number of papers in the initial set, most papers were manually checked by one annotator in the the screening and annotation stage; thus, annotation errors or inconsistencies are inevitable. Finally, due to page limits for submission, while XX papers are included in this survey from which we gathered our statistics, only a small subset of them are discussed individually in our paper.

Ethical Consideration

All the papers covered in our survey and the corpora used in our two case studies are publicly available. The screening process in Section 3 and manual checking for the case studies were performed by researchers on our team. We are not aware of any ethical issues that arose while conducting our work.

References

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*.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023a. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023b. Multimodal automated fact-checking: a survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

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*.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*.

Carlos Alvarez, Maxwell Bennett, and Lucy Wang. 2024. Zero-shot scientific claim verification using llms and citation text. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.

A. Barik, W. Hsu, and M. Lee. 2024a. Chronofact: Timeline-based temporal fact verification. DOI:10.48550/arXiv.2410.14964.

A. Barik, W. Hsu, and M. Lee. 2024b. Evidence-based temporal fact verification. DOI:10.48550/arXiv.2407.15291.

A. Bazaga, Pietro Lio, and G. Micklem. 2023. Unsupervised pretraining for fact verification by language model distillation. In *International Conference on Learning Representations*.

Varad Bhatnagar, Diptesh Kanojia, and Kameswari Chebrolu. 2022. Harnessing abstractive summarization for fact-checked claim detection. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Maniruzzaman Bhuiyan, Farzana Sultana, and Aha Mudur Rahman. 2025. Fake news classifier: Advancements in natural language processing for automated fact-checking. *Strategic Data Management and Innovation*, pages 181–201.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2024. Defame: Dynamic evidence-based fact-checking with multimodal experts. DOI:10.48550/arXiv.2412.10510.

Ramón Casillas, Helena Gómez-Adorno, V. Lomas-Barrie, and Orlando Ramos-Flores. 2022. Automatic fact checking using an interpretable bert-based architecture on covid-19 claims. *Applied Sciences*, 12(20).

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2024. Cross-lingual learning vs. low-resource fine-tuning: a case study with fact-checking in turkish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Megha Chakraborty, Khushbu Pahwa, Anku Rani, Shreyas Chatterjee, Dwip Dalal, Harshit Dave, Ritvik G, Preethi Gurumurthy, Adarsh Mahor, Samahriti Mukherjee, Aditya Pakala, Ishan Paul, Janvita Reddy, Arghya Sarkar, Kinjal Sensharma, Aman Chadha, Amit Sheth, and Amitava Das. 2023. Factify3m: a benchmark for multimodal fact verification with explainability through 5w question-answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Yi-Chen Chang, Canasai Kruengkrai, and Junichi Yamagishi. 2023. Xfever: exploring fact verification across languages. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024a. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Ting-Chih Chen, Chia-Wei Tang, and Chris Thomas. 2024b. Metasumperceiver: multimodal multi-document evidence summarization for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Zacharias Chrysidis, Stefanos-Iordanis Papadopoulos, Symeon Papadopoulos, and P. Petrantonakis. 2024. Credible, unreliable or leaked?: Evidence verification for enhanced automated fact-checking. In *Proceedings* of the 3rd ACM International Workshop on Multimedia AI against Disinformation.*

Yi-Ling Chung, Aurora Cobo, and Pablo Serna. 2025. Beyond translation: Llm-based data generation for multilingual fact-checking. DOI:10.48550/arXiv.2502.15419.

Svetlana Churina, Anab Maulana Barik, and Saisamarth Rajesh Phaye. 2024. Improving evidence retrieval on claim verification pipeline through question enrichment. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*.

Oliver Deck, Z. M. Hüsünbeyi, Leonie Uhling, and Tatjana Scheffler. 2025. Annotation and linguistic analysis of claim types for fact-checking. *Linguistics Vanguard*.

Mitchell DeHaven and Stephen Scott. 2023. Bevers: a general, simple, and performant framework for automatic fact verification. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*.

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

A. Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of large language models: A survey. DOI:10.48550/arXiv.2408.14317.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact checking: a survey on architectures and approaches. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Michael Evans, Dominik Soós, Ethan Landers, and Jian Wu. 2023. Msvec: A multidomain testing dataset for scientific claim verification. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*.

Yu Fu, Shunan Guo, J. Hoffswell, V. S. Bursztyn, R. Rossi, and J. Stasko. 2024. ""the data says otherwise""-towards automated fact-checking and communication of data claims. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. Ambifc: fact-checking ambiguous claims with evidence. In *Transactions of the Association for Computational Linguistics, Volume 12*.

Haisong Gong, Weizhi Xu, Shu Wu, Q. Liu, and Liang Wang. 2024. Heterogeneous graph reasoning for fact checking over texts and tables. In *AAAI Conference on Artificial Intelligence*.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: desiderata for decontextualization in llm fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

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

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Dialfact: a benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Vishwani Gupta, Astrid Viciano, Holger Wormer, and Najmehsadat Mousavinezhad. 2023. Exploring unsupervised semantic similarity methods for claim verification in health care news articles. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*.

Harya Gusdevi, A. Setyanto, Kusrini, and Ema Utami. 2024. Systematic literature review on technology-based fact verification. In *2024 Ninth International Conference on Informatics and Computing (ICIC)*.

Fatima Haouari, Tamer Elsayed, and Reem Suwaileh. 2024. Aured: Enabling arabic rumor verification using evidence from authorities over twitter. In *ARABICNLP*.

Tran Thai Hoa, Tran Quang Duy, Khanh Quoc Tran, and Kiet Van Nguyen. 2024. Vifactcheck: A new benchmark dataset and methods for multi-domain news fact-checking in vietnamese. DOI:10.48550/arXiv.2412.15308.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2024. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? DOI:10.48550/arXiv.2411.02400.

Xuming Hu, Zhijiang Guo, Guan-Huei Wu, Lijie Wen, and Philip S. Yu. 2023. Give me more details: Improving fact-checking with latent retrieval. DOI:10.48550/arXiv.2305.16128.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. Chef: a pilot chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

3460, Online. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. In *Conference on Empirical Methods in Natural Language Processing*.

Wei-Yu Kao and An-Zi Yen. 2024a. How we refute claims: Automatic fact-checking through flaw identification and explanation. In *Companion Proceedings of the ACM on Web Conference 2024*.

Wei-Yu Kao and An-Zi Yen. 2024b. Magic: multi-argument generation with self-refinement for domain generalization in automatic fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

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)*.

Neema Kotonya and Francesca Toni. 2024. Towards a framework for evaluating explanations in automated fact verification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. Proofver: natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Yuqing Lan, Zhenghao Liu, Yu Gu, Xiaoyuan Yi, Xiaohua Li, Liner Yang, and Ge Yu. 2025. Multi-evidence based fact verification via a confidential graph neural network. *IEEE Transactions on Big Data*, 11:426–437.

Hung Tuan Le, Long Truong To, Manh Trong Nguyen, and Kiet Van Nguyen. 2024. Viwikifc: Fact-checking for vietnamese wikipedia-based textual knowledge source. DOI:10.48550/arXiv.2405.07615.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-checker: plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Hongbin Lin and Xianghua Fu. 2022. Heterogeneous-graph reasoning and fine-grained aggregation for fact checking. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*.

Ying-Jia Lin, Chun Lin, Chia-Jen Yeh, Yi-Ting Li, Yun-Yu Hu, Chih-Hao Hsu, Mei-Feng Lee, and Hung-Yu Kao. 2024. Cfever: A chinese fact extraction and verification dataset. In *AAAI Conference on Artificial Intelligence*.

Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. 2023. Covid-vts: fact extraction and verification on short video platforms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Jiayu Liu, Junhao Tang, Hanwen Wang, Baixuan Xu, Haochen Shi, Weiqi Wang, and Yangqiu Song. 2024a. Gprooft: a multi-dimension multi-round fact checking framework based on claim fact extraction. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*.

Jin Liu, Steffen Thoma, and Achim Rettinger. 2024b. Fzi-wim at averitec shared task: real-world fact-checking with question answering. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*.

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*.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024. Exfever: a dataset for multi-hop explainable fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Shrikant Malviya and Stamos Katsigiannis. 2024. Evidence retrieval for fact verification using multi-stage reranking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Shreyash Mishra, S. Suryavardan, Amrit Bhaskar, P. Chopra, Aishwarya N. Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, A. Sheth, and Asif Ekbal. 2022. Factify: A multi-modal fact verification dataset. In *DE-FACTIFY@AAAI*.

Kushan Mitra, Dan Zhang, Sajjadur Rahman, and Estevam R. Hruschka. 2024. Factlens: Benchmarking fine-grained fact verification. DOI:10.48550/arXiv.2411.05980.

Marco Mori, Paolo Papotti, Luigi Bellomarini, and Oliver Giudice. 2022. Neural machine translation for fact-checking temporal claims. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*.

Arief Purnama Muharram and Ayu Purwarianti. 2024. Enhancing natural language inference performance with knowledge graph for covid-19 automated fact-checking in indonesian language. DOI:10.48550/arXiv.2409.00061.

Kevin Nanekhan, V. Venktesh, Erik Martin, Henrik Vatndal, Vinay Setty, and Avishek Anand. 2025. Flashcheck: Exploration of efficient evidence retrieval for fast fact-checking. In *European Conference on Information Retrieval*.

Nam V. Nguyen, Dien X. Tran, Thanh T. Tran, Anh T. Hoang, Tai V. Duong, Di T. Le, and Phuc-Lu Le. 2025. Semviqa: A semantic question answering system for vietnamese information fact-checking. DOI:10.48550/arXiv.2503.00955.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Rrubaa Panchendrarajan and A. Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Faviq: Fact verification from information-seeking questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Md. Rashadur Rahman, Rezaul Karim, M. Arefin, P. K. Dhar, Gahangir Hossain, and Tetsuya Shimamura. 2025. Facilitating automated fact-checking: a machine learning based weighted ensemble technique for claim detection. *Discover Applied Sciences*, 7:73.

Aman Rangapur, Haoran Wang, and Kai Shu. 2023. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation. DOI:10.48550/arXiv.2309.08793.

Anku Rani, S.M Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. Factify-5wqa: 5w aspect-based fact verification through question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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

Pritish Sahu, Karan Sikka, and Ajay Divakaran. 2024. Pelican: correcting hallucination in vision-llms via claim decomposition and program of thought verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Antonio Scaiella, Stefano Costanzo, Elisa Passone, Danilo Croce, and Giorgio Gambosi. 2024. Leveraging large language models for fact verification in italian. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*.

M. Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Neural Information Processing Systems*.

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)*, pages 1–26, Miami, Florida, USA. Association for Computational Linguistics.

Vinay Setty and Adam James Becker. 2025. Annotation tool and dataset for fact-checking podcasts. DOI:10.48550/arXiv.2502.01402.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022a. Empowering the fact-checkers! automatic identification of claim spans on twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Megha Sundriyal, Ganeshan Malhotra, Md Shad Akhtar, Shubhashis Sengupta, Andrew Fano, and Tanmoy Chakraborty. 2022b. Document retrieval and claim verification to mitigate covid-19 misinformation. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*.

Suryavardan Suresh, Anku Rani, Parth Patwa, Aishwarya N. Reganti, Vinija Jain, Aman Chadha, Amitava Das, Amit P. Sheth, and Asif Ekbal. 2024. Overview of factify5wqa: Fact verification through 5w question-answering. DOI:10.48550/arXiv.2410.04236.

S. Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya N. Reganti, Aman Chadha, Amitava Das, Amit P. Sheth, Manoj Kumar Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Factify 2: A multimodal fake news and satire news dataset. In *DE-FACTIFY@AAAI*.

Fiona Anting Tan, Jay Desai, and Srinivasan H. Sengamedu. 2024. Enhancing fact verification with causal knowledge graphs and transformer-based retrieval for deductive reasoning. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*.

Neset Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan, and Michael Witbrock. 2023. Multi2claim: generating scientific claims from multi-choice questions for scientific fact-checking. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Xin Tan, Bowei Zou, and Ai Ti Aw. 2025. Improving explainable fact-checking with claim-evidence correlations. In *Proceedings of the 31st International Conference on Computational Linguistics*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

(Long Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Herbert Ullrich, Jan Drchal, Martin R'ypar, Hana Vincourov'a, and Václav Moravec. 2023. Csfever and ctkfacts: acquiring czech data for fact verification. *Language Resources and Evaluation*, pages 1571–1605.

Herbert Ullrich, Tomás Mlynár, and Jan Drchal. 2025. Claim extraction for fact-checking: Data, models, and automated metrics. DOI:10.48550/arXiv.2502.04955.

V. Venktesh, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. Healthfc: verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

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*.

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*.

Lianwei Wu, Dengxiu Yu, Pusheng Liu, Chao Gao, and Zhen Wang. 2023. Heuristic heterogeneous graph reasoning networks for fact verification. *IEEE Transactions on Neural Networks and Learning Systems*, 35:14959–14973.

Amelie Wuehrl, Lara Grimminger, and Roman Klinger. 2023. An entity-based claim extraction pipeline for real-world biomedical fact-checking. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*.

Bangrui Xu, Fuhui Sun, Xiaoliang Liu, Peng Wu, Xiaoyan Wang, and Li-Li Pan. 2024. Complex claim verification via human fact-checking imitation with large language models. In *2024 19th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*.

Song Yang, Xue Yuan, Tong Gan, and Yue Wu. 2024. A survey of automatic fact verification research. In *2024 7th World Conference on Computing and Communication Technologies (WCCCT)*.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2022. 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*.

Yitao Long Yilun Zhao, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. 2024. Findver: explainable claim verification over long and hybrid-content financial documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Majid Zarharan, Pascal Wullschleger, Babak Behkam Kia, Mohammad Taher Pilehvar, and Jennifer Foster. 2024. Tell me why: explainable public health fact-checking with large language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*.

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

Fengzhu Zeng and Wei Gao. 2024. Justilm: few-shot justification generation for explainable fact-checking of real-world claims. In *Transactions of the Association for Computational Linguistics, Volume 12*.

Xia Zeng and A. Zubiaga. 2022. Aggregating pairwise semantic differences for few-shot claim verification. *PeerJ Computer Science*, 8:e1137.

Yirong Zeng, Xiao Ding, Yi Zhao, Xiangyu Li, Jie Zhang, Chao Yao, Ting Liu, and Bing Qin. 2024. Ru22fact: optimizing evidence for multilingual explainable fact-checking on russia-ukraine conflict. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. 2024a. Do we need language-specific fact-checking models? the case of chinese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Xiaocheng Zhang, Xi Wang, Yifei Lu, Zhuangzhuang Ye, Jianing Wang, Mengjiao Bao, Peng Yan, and Xiaohong Su. 2024b. Augmenting the veracity and explanations of complex fact checking via iterative self-revision with llms. DOI:10.48550/arXiv.2410.15135.

Xiaocheng Zhang, Xi Wang, Yifei Lu, Zhuangzhuang Ye, Jianing Wang, Mengjiao Bao, Peng Yan, and Xiaohong Su. 2024c. Verification with transparency: The trendfact benchmark for auditable fact-checking via natural language explanation. DOI:10.48550/arXiv.2410.15135.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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*.

A Scraping and Filtering Details

We collected papers from three sources:

- **Semantic Scholar**: Queried via their public API with keyword queries like "fact checking" and "claim verification". We retrieved up to 400 papers and filtered the first 200 titles that matched either an exact keyword phrase or at least two unigrams after stopword removal.
- **Google Scholar**: Accessed via SerpAPI. Titles were filtered using the same logic as above. Due to SerpAPI limits and noisier metadata, fewer papers passed the filter.
- **ACL Anthology**: Parsed locally from metadata in the official ACL Anthology GitHub repository. XML files were searched for titles with exact keyword phrases or ($\geq$2) keyword unigrams.

Across all sources, abstract matching was enabled (via the '–check-abstracts' flag) to increase relevance. Deduplication was performed using normalized titles, with preference given to papers from ACL Anthology, followed by Semantic Scholar, then Google Scholar.

B An Example of Claim Generation

Figure 6 shows an example from Feverous dataset (Aly et al., 2021), which is used as original claims in FactLens (Mitra et al., 2024) dataset. The claim is generated by using information from three sentences on the first Wikipedia article[5] and a table on the second article[6]. The colors show the connection between the claim and the sources. The purple highlights are about context information relevant to the claim. Specifically, together with these cues, temporal information "2013" can be also inferred from the fact that the paragraph shown in (a) is between two paragraphs that talked about Mansell's career in 2012 and 2014.

C Details of Case Study #1

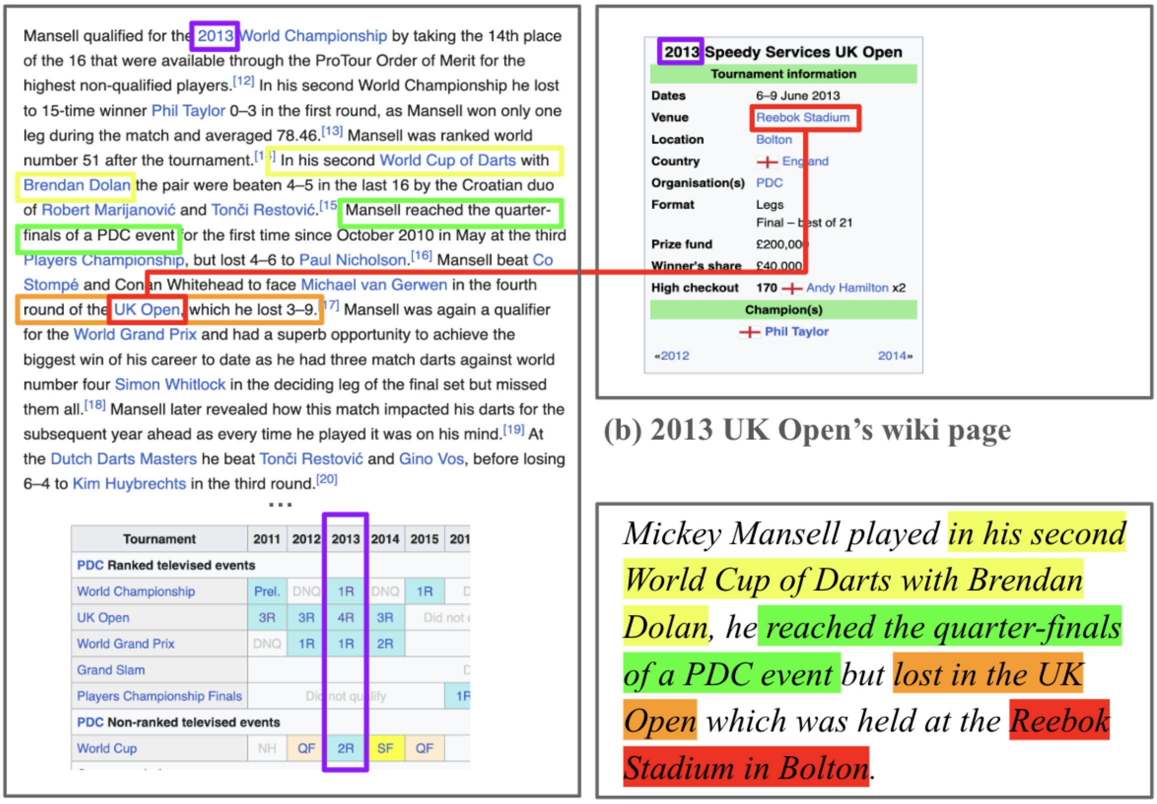
In this appendix, we provide full examples of six types of relations regarding claim-veracity-EBS triples. They are presented in table 1. As for whether we agree with the labels given by the author, we also provide examples for the following 5 scenarios: we agree with the label and believe only one EBS is needed for justifying the label; we agree with the label and believe a combination of two EBSs are needed for justification; we disagree with the label and believe the EBSs are supporting a different label; we disagree with the label and are unsure what label to put due to contradictory information; disagree with the label and are unsure what label to put due to irrelevant and vague EBSs. The examples are given in table 2.

D Details of Case Study #2

Conjunction: One of the most common decomposition strategies is to split coordinated structures, a pattern observed in approximately half of our sample. This strategy is generally safe when the conjunction connects two independent clauses. However, it becomes problematic when the coordination occurs at the noun or modifier level. In two cases, we observed that decomposing noun-level conjunctions resulted in a loss of essential combined meaning. For example, the claim *"Analysis of A and B shows C"* was split into *"Analysis of A shows C"* and *"Analysis of B shows C"*, leading to subclaims that no longer entail the original claim. Another type of issue arises when prepositional phrases (PPs) or adjectives are involved in the conjunction. Splitting such constructions can force a disambiguation not present in the original claim. For instance, in the phrase *"A and B of C"*, the decomposition can yield either *"A of C; B of C"* or *"A; B of C"*, each carrying a distinct semantic interpretation. In such cases, the decomposition introduces ambiguity or alters the intended meaning.

Head + Restricted Modifier: In six examples, decomposition involved noun phrases with restricted modifiers, such as relative clauses, tightly scoped adjectives or restricted phrases. In three of these cases, we observed redundancy issues. Specifically, the system added the original claim as a subclaim alongside a version that included only the head noun without its modifier. Alongside this problem, the head noun was also included as a standalone subclaim, resulting in misleading entailments. For example, the subclaim *T-cell deficiency can affect spatial learning ability"* may be true, while the full original claim *"T-cell deficiency can affect spatial learning ability following toluene exposure"* may not. In such cases, the subclaim set entails but is

[5]https://en.wikipedia.org/wiki/Mickey_Mansell
[6]https://en.wikipedia.org/wiki/2013_UK_Open



(a) Mickey Mansell's wiki page (b) 2013 UK Open's wiki page (c) A claim generated from (a) and (b)

Figure 6: A claim from the Feverous corpus, which was generated from two Wikipedia articles

Claim	Label	EBS	Relation
Do heat patches help with lower back pain?	Support	Carrying self-warming patches for three days has on average improved the pain in the lower back by 18 points on the 100s scale [1].	Sufficient
Do heat patches help with lower back pain?	Support	In addition to movement exercises or painkillers, the heat patches are probably pain-relieving	Partially Sufficient
Does light freezing help with weight loss?	NEI	Nevertheless, weight loss was not significantly higher than after the same exercise program at more pleasant temperatures.	Against
Does taking magnesium salts reduce the frequency and intensity of exercise-induced muscle cramps during sports?	NEI	In muscle spasms without obvious cause, the symptoms were not easier and did not occur less frequently compared to placebo when participants had taken magnesium supplements.	Partially Against
Do milk or dairy products promote colon cancer and rectal cancer?	Refute	There is also the possibility that dairy products will reduce the likelihood of bladder cancer.	Irrelevant
Do milk or dairy products promote colon cancer and rectal cancer?	Refute	However, the study situation is still too unclear to draw definitive conclusions, which requires more and more meaningful studies.	Vague

Table 1: Full example of six types relations for claim-veracity-EBS-triples

Claim	Label	EBSs	Agreement	Rational
Can arthroscopy reduce pain or improve mobility?	Refute	1. Studies clearly speak against a benefit A research team summarized the most meaningful of all previously published studies on arthroscopy in knee arthritis. 2. In these studies, patients treated after arthroscopy had no noticeably less pain or movement restrictions than those treated only for appearance or not at all. 3. Arthroscopy against osteoarthritis: not effective, but also not very risky After all: Undesirable events were not conspicuously common in the arthroscopy groups either.	Agree	Each EBS is sufficient
Does taking antibiotics for acute sinusitis speed up the healing of the infection?	Support	1. They say that antibiotics can shorten acute sinus inflammation a little – but only in a few people. 2. Sickness duration: only 5 out of 100 benefit What is the benefit of taking an antibiotic on the cure, i.e. 3. This means that only 5 out of 100 people with acute rhinosinusitis benefit from taking an antibiotic instead of a dummy medication.	Agree	EBS 1 and 3 combined are sufficient to justify the label
Does taking magnesium salts reduce the frequency and intensity of exercise-induced muscle cramps during sports?	NEI	1. Anyone suffering from nocturnal calf cramps without known cause will probably not feel relief from magnesium preparations [1] [2]. 2. In muscle spasms without obvious cause, the symptoms were not easier and did not occur less frequently compared to placebo when participants had taken magnesium supplements. 3. Accordingly, the authors also came to similar conclusions: no effect of magnesium salts was detectable in the general population compared to placebo.	Disagree	EBS 1 or 3 suggests the label "Refute"
Can antibiotic-resistant germs from animal husbandry be transferred to humans?	Support	1. However, studies indicate that transmission to humans is possible. 2. For example, persons such as farmers, veterinarians or slaughterhouse workers who have frequent contact with farm animals for professional reasons are likely to be more likely to be populated with resistant bacteria than persons from the general population [1] [8] [10–12]. 3. Their summarized results show that people with close contact with animals such as farmers, veterinarians or slaughterhouse workers are actually more frequently populated than the average population with the so-called "livestock-associated MRSA". 4. From this, the study authors conclude that a transfer of resistant germs from animals to humans is in principle possible. 5. Although this type of study may give indications that antibiotic use in animal husbandry will transfer resistant pathogens to humans, it is not possible to provide clear evidence.	Disagree	EBS 5 suggests the label "NEI" rather than "Refute". It contradicts with EBS 2, 3, or 4.
Do green smoothies promote health?	NEI	1. However, studies on green smoothies are not yet available. 2.In other words, the claim that they promote health is not substantiated. 3.They cannot easily be transferred to humans. 4. From the point of view of evidence-based reporting, the topic would be already eaten.	Disagree	All EBSs are vague and thus are not contributory to any label.

Table 2: Full example of five scenarios in which we agree or disagree with the label provided by the authors, with rationals for our opinions.

not semantically equivalent to the original claim.

Head + Unrestricted Modifier: In approximately half of our samples, the decomposition involved head noun phrases with *unrestricted modifiers*, such as unrestricted relative clauses, appositive clauses, and prepositional phrases that are not semantically essential to the head. This strategy is generally safe, as the unrestricted modifier contributes supplementary information without altering the scope or truth conditions of the main proposition. However, care must be taken when decomposing appositive constructions, particularly when a *be*-verb is inserted to form a standalone subclaim. These cases are often **tense-sensitive**. For example, the claim: *"Cuba, a member of the Commonwealth Realms under the monarchy of Queen Elizabeth II, ..."*may be incorrectly decomposed into:*"Cuba is a member of the Commonwealth Realms"; "Cuba is under the monarchy of Queen Elizabeth II."* Using the present tense here may introduce factual inaccuracies, particularly if the context implies a historical or past-tense reading.

Head with Multiple Dependents: A critical issue we observed involves cases where a single head element (such as a predicate or noun phrase) has multiple dependent phrases, and the decomposition splits these dependents into separate subclaims. This results in a loss of meaning that arises from their joint contribution. For example, consider the original claim: *"HIV-infected patients should be screened for silent myocardial ischaemia using gated myocardial perfusion SPECT."* which was decomposed into: *"HIV-infected patients should be screened for silent myocardial ischaemia"; "HIV-infected patients should be screened using gated myocardial perfusion SPECT."* In this decomposition, the link between the method (SPECT) and the target condition (ischaemia) is severed. Each subclaim is independently verifiable, but the original intent—screening for a specific condition using a specific method—is not preserved. In such cases, the subclaim set does not entail the original claim.

Clause-taking Verbs: Another issue arises when decomposing constructions in which a verb takes a clause as its complement. This occurred in two of our annotated samples. Consider the claim: *'X, as determined by histological evaluation'* which was decomposed into: *"X"; "Histologic evaluation determined X."* This decomposition is problematic because the subclaim 'X' is no longer supported by any evidential attribution. It presents the proposition as a standalone fact, rather than one dependent on an evaluative process. In contexts where the original claim relies on such attribution (e.g., evaluation, belief, reporting), the stripped-down subclaim can overstate the certainty or factual status of the information.

E Claim Verification Corpora in Our Collection

In this section, we curated an extensive collection of corpora used in the papers in our survey. These datasets span diverse modalities (text, image, video, and audio), languages, and application domains, offering a broad foundation for both benchmarking and qualitative assessment. The full list is detailed in Table 3 to 6.

Corpus Name	Corpus Size	Modality	Language	Seed dataset	Veracity	Justification	Link
Bangla Claim Detection Dataset(Rahman et al., 2025)	4	1	ben	fact-checking websites, interviews, speeches	1	0	Avialable upon request
FEVERFact(Ullrich et al., 2025)	5	1	eng	podcast episodes	1	0	link
GCC(Deck et al., 2025)	3	1	ger	WhatsApp	3	0	Available upon request
2024 Presidential Debate Claims(Nanekhan et al., 2025)	1	1	eng	presidential debates	1	1	link
Fact-Checking Podcasts Dataset(Setty and Becker, 2025)	1	1,4	eng, ger, nor	podcast episodes	N/A	0	link
MultiSynFact(Chung et al., 2025)	5	1	eng, spa, ger, low	LLMs	2	1	link
CorFEVER(Tan et al., 2025)	2	1	eng	online sources	2	3	link
CHEF-EG, TrendFact(Zhang et al., 2024b)	4	1	chi	CHEF, Weibo	2	3	N/A
T-FEVER, T-FEVEROUS(Barik et al., 2024b)	5	1	eng	FEVER, FEVEROUS	2	1	N/A
ChronoClaims(Barik et al., 2024a)	5	1	eng	Wikipedia	2	1	N/A
FactLens(Mitra et al., 2024)	2	1	eng	CoverBench	1	1,3	N/A
Factify5WQA(Suresh et al., 2024)	5	1	eng	fact-checking datasets	2	1	link
ViFactCheck(Hoa et al., 2024)	4	1	vie	newspwpers	2	1	link
ViWikiFC(Le et al., 2024)	5	1	vie	Wikipedia	2	0	link
TrendFact (Zhang et al., 2024c)	5	1	chi	social media, fact-checking websites	2	2, 3	link

Table 3: Claim Verification Corpora in Our Collection (1 of 4).

Legend for column codes:

- **Corpus Name:** This is the name of the CV corpus the paper created.
- **Corpus size:** 1: no more than 500 instances, 2: no more than 1,000 instances, 3: no more than 5,000 instances, 4: no more than 10,000 instances, 5: greater than 10,000 instances
- **Modality:** 1 = text, 2 = image, 3 = video, 4 = audio, 5 = chart, 6 = table, 7 = others
- **Language:** eng = English, ben = Bengali, chi = Chinese, jpn = Japanese, spa = Spanish, ger = German, ita = Italian, ind = Indonesian, fre = French, tib = Tibetan, rus = Russian, ukr = Ukrainian, vie = Vietnamese, tur = Turkish, nor = Norwegian, cze = Czech, low = low-resource languages mult = multilingual
- **Seed dataset:** It is the seed dataset used by the CV corpus.
- **Veracity:** 1 = binary (true/false), 2 = ternary (supported/refuted/NEI), 3 = more than 3 labels, 4 = numerical scale, 5 = others
- **Justification:** 0 = N/A, 1 = evidence-bearing sentences, 2 = summary, 3 = explanation, 4 = others
- **Link:** the link to access the dataset

Corpus Name	Corpus Size	Modality	Language	Seed dataset	Veracity	Justification	Link
CREDULE(Chrysidis et al., 2024)	5	1	eng	MultiFC, Politifact, PUBHEALTH, NELA-GT, Fake News Corpus	3	3	link
CFEVER(Lin et al., 2024)	5	1	chi	Wikipedia	2	0	link
CLAIMREVIEW2024+(Braun et al., 2024)	1	1, 2	eng	ClaimReview Project	3	0	link
QuanTemp(Venktesh et al., 2024)	5	1	eng	Google Fact Check Tools API	2	0	link
FlawCheck(Kao and Yen, 2024a)	5	1	eng	WatClaimCheck	3	0	link
Adversarial CHEF(Zhang et al., 2024a)	2	1	chi	CHEF	N/A	3	link
LLMforFV(Guan et al., 2024)	2	1	eng	LLMs	1	0	link
RU22Fact(Zeng et al., 2024)	5	1	eng, chi, rus, ukr	fact-checking websites, news outlets	2	3	link
XClaimCheck(Kao and Yen, 2024b)	5	1	eng	WatClaimCheck, PolitiFact	3	0	link
HealthFC(Vladika et al., 2024)	2	1	eng, ger	Medizin Transparent web portal	2	1, 2	link
FCTR(Cekinel et al., 2024)	3	1	tur	fact-checking organization, Snopes	3	2	link
ChartCheck(Akhtar et al., 2024)	5	1, 5	eng	Wikimedia Commons	2	3	link
EX-Fever(Ma et al., 2024)	5	1	eng	Wikipedia	2	3	link
BINGCHECK(Li et al., 2024)	3	1	eng	ChatGPT prompted user queries	3	0	N/A
EX-Claim(Zeng and Gao, 2024)	4	1	eng	WatClaim Check	1	3	link
UNK(Tan et al., 2024)	5	1	eng	reports from National Transportation Safety Board	1	0	N/A
AMBIFC(Glockner et al., 2024)	5	1	eng	BoolQ dataset	2	0	link

Table 4: Claim Verification Corpora in Our Collection (2 of 4).

Corpus Name	Corpus Size	Modality	Language	Seed dataset	Veracity	Justification	Link
Multi-News-Fact-Checking(Chen et al., 2024b)	5	1, 2	eng	Multi-News summarization dataset	3	2, 3	link
FINDVER(Yilun Zhao et al., 2024)	3	1, 6	eng	company reports through U.S. Securities and Exchange Commission	1	3	link
FEVER-it(Scaiella et al., 2024)	5	1	ita	FEVER	2	0	link
AuRED(Haouari et al., 2024)	1	1	ara	Twitter	2	0	link
Facity 2(Suryavardan et al., 2023)	5	1, 2	eng	Twitter	3	0	link
WICE(Kamoi et al., 2023)	3	1	eng	Wikipdeia	2	1	link
Fin-Fact(Rangapur et al., 2023)	3	1, 2	eng	PolitiFact, Snopes, FactCheck	2	3	link
EFact(Hu et al., 2023)	4	1	eng	fact-checking organization	3	0	N/A
X-Fact(Hu et al., 2023)	5	1	mult	fact-checking organization	3	0	N/A
MSVEC(Evans et al., 2023)	1	1	eng	news outlets, fact-checking websites	1	1	link
AVeriTeC(Schlichtkrull et al., 2023)	3	1	eng	fact-checking organizations	3	3	link
Multi2Claim(Tan et al., 2023)	5	1	eng	scientific multiple-choice QA datasets	N/A	3	link
COVID-VTS(Liu et al., 2023)	4	1, 3	eng	Twitter	1	1, 3	link
FACTKG(Kim et al., 2023)	5	1	eng	WebNLG datase	1	0	link
FACTIFY-5WQA(Rani et al., 2023)	5	1	eng	fact verification datasets	2	1, 3	link
LIAR++; FullFact(Russo et al., 2023)	4	1	eng	LIAR-PLUS, FULL-FACT website	2	3	link
XFEVER(Chang et al., 2023)	5	1	eng, chi, jpn, spa, ind, fre	FEVER	2	0	link
Check-COVID(Wang et al., 2023)	3	1	eng	scientific journal articles	2	0	link

Table 5: Claim Verification Corpora in Our Collection (3 of 4).

Corpus Name	Corpus Size	Modality	Language	Seed dataset	Veracity	Justification	Link
ChartFC(Akhtar et al., 2023a)	5	1, 5	eng	TabFact	1	0	link
MultiClaim(Pikuliak et al., 2023)	5	1	mult	Google Fact Check Explorer, Snopes	1	0	Available upon request
FACTIFY 3M(Chakraborty et al., 2023)	5	1, 2	eng	ChatGPT, visual paraphrases	3	2, 3	N/A
SCITAB(Lu et al., 2023)	3	1, 6	eng	Sci-Gen dataset	2	0	link
German healthcare news articles(Gupta et al., 2023)	1	1	eng, ger	German news sources	N/A	1	N/A
CsFEVER, CTKFacts(Ullrich et al., 2023)	5	1	cze	Czech adaptation of the English FEVER	3	1	link
FACTIFY(Mishra et al., 2022)	5	1, 2	eng	Twitter	3	0	link
Custom COVID-19 Claims Dataset(Casillas et al., 2022)	3	1	eng	WHO Mythbusters, John Hopkins FAQs, CNN QA pages	1	0	link
Mocheg(Yao et al., 2022)	5	1, 2	eng	PolitiFact, Snopes	2	1	link
SCIFACT-OPEN(?)	5	1	eng	SCIFACT-ORIG test set	2	1	link
PubHealthTab(Akhtar et al., 2022)	3	1, 6	eng	fact-checking, news review websites	1	0	link
SufficientFacts(Atanasova et al., 2022)	2	1	eng	FEVER, Vitamin C, HoVer	2	0	link
CHEF(Hu et al., 2022)	5	1	chi	news review sites	2	0	link
FC-Claim-Det(Bhatnagar et al., 2022)	1	1	eng	Fact-checked articles	2	2, 3	link
FAVIQ(Park et al., 2022)	5	1	eng	Natural Questions dataset, AmbigQA	1	0	link
ClaVer(Sundriyal et al., 2022b)	3	1	eng	CORD-19, LESA	2	0	link
DIALFACT(Gupta et al., 2022)	5	1	eng	Wikipedia	2	1	link
CURT(Sundriyal et al., 2022a)	4	1	eng	Twitter	N/A	3	link

Table 6: Claim Verification Corpora in Our Collection (4 of 4).