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

Zhaxi Zerong*, **Chenxi Li***, **Xinyi Liu**, **Ju-hui Chen**, **Fei Xia**
University of Washington
Seattle, WA, USA
{tashi0, cl91, xliu118, juhuic, fxia}@uw.edu

Abstract

Automated Claim Verification (CV)—the task of assessing a claim's veracity against explicitly provided evidence—is a critical tool in the fight against growing misinformation. This survey offers a comprehensive analysis of 198 studies published between January 2022 and March 2025, synthesizing recent advances in CV corpus creation and system design. Through two in-depth case studies, we illuminate persistent challenges in veracity annotation, limitations of conventional CV pipelines, and pitfalls in recent claim decomposition approaches. We conclude by identifying key unresolved challenges and proposing productive directions for future research.[1]

1 Introduction

The growing scale of misinformation has led to a surge of research in automated fact-checking and claim verification (CV), which assess whether a given claim is supported by accompanying references. A key milestone in this field was the release of FEVER (Thorne et al., 2018), a synthetic dataset for CV which sparked the development of other synthetic 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 designs of CV systems 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 CV 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 corpus 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 recent studies on CV 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 in-depth case studies.

2 Task Setting

The input to a CV system consists of a **claim** and one or more **reference documents** (**reference** in short). The latter is called **evidence** or **context** in some previous studies. To avoid confusion, in this study, we use the term **evidence-bearing sentences** to refer to sentences in the reference that support or refute a claim. The output of a CV system includes a **veracity label** and optionally a **justification** to provide support or explanation to the veracity label.

A related task is called **fact checking** (aka **open-domain fact-checking**), where only a claim is provided as input and the system needs to retrieve relevant documents (i.e. *references*) from external sources such as the Internet. In this survey, we will focus on CV, not fact checking, because one can easily build a fact checking system on top of a CV system by adding a document-retrieval module.

* Equal contribution.
[1]The list of papers included in this survey and the annotations for the two case studies are available at https://github.com/CLINEEK/EMNLP2025-Claim-Verification-Survey

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21452–21474
November 4-9, 2025 ©2025 Association for Computational Linguistics

Also, we want to 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[2], Semantic Scholar[3], and Google Scholar[4]. We used query terms *(fact OR claim) AND (checking OR verification)* to retrieve papers published between January 2022 and March 2025.[5] After removing duplicates, there were 316 papers left, forming our initial set of papers.

3.2 Manual screening and categorization

We read all the 316 papers and divided them into three groups: (a) 62 papers that are not on fact checking or CV; (b) 56 papers on fact checking; (c) 198 papers on CV, which form the **main collection** of studies covered in this survey.

We categorize the 198 papers in our main collection into four groups based on their focus: (G1) 47 papers on corpus construction, (G2) 141 on system development, (G3) eight survey papers, and (G4) two miscellaneous papers. Notably, 15 papers in G1 also developed CV systems, while 18 in G2 created CV corpora.

We discuss all 47 papers from G1 and the 18 corpus-building papers from G2 in the corpus construction section (Section 4). Similarly, all 141 papers from G2 and the 15 system-building papers from G1 are covered in the system development section (Section 5). The eight survey papers (G3) are reviewed in the related work section (Section 9). In addition to these 198 papers, we also reference—where relevant—fact checking and influential pre-2022 CV studies such as FEVER (Thorne et al., 2018), FEVEROUS (Aly et al., 2021), and HoVer (Jiang et al., 2020).

[2]https://aclanthology.org/

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

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

[5]Because the terms 'fact-checking' and 'claim verification' are sometimes used interchangeably in the literature, we included both terms in our search query to ensure comprehensive paper retrieval and then filter out fact checking papers through manual screening. Appendix A provides details of our scraping setup.

4 Corpus Creation

In this section, we report findings from 65 papers in our collection that create new CV corpora.

4.1 Main components of a CV corpus

An instance in a CV corpus 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 and publication date.

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 Claim-Review2024+ (Braun et al., 2024). For instance, a claim can be a (text, image) pair, extracted from public websites such as Twitter.

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).

Justification: Although justification is not a required field in a CV corpus, it provides explanation to the veracity label and the 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 (e.g., (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

Below are some basic properties of the 65 newly created corpora.

Size: Twelve corpora have 1,000 or fewer instances, 20 have 1,000 to 10,000 instances, and the remaining 33 each have over 10,000 instances.

Modality: Fifty-two corpora are text only and 13 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 justificatios in all these corpora are text only, we believe there are use cases where multi-modal justification would be beneficial (e.g., an image that highlights 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) Bengali (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; 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)). The majority of corpora collect data from multiple domains with Wikipedia being 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 entirely from scratch; rather, their core components—claims, references, veracity labels, and justification—are (1) created, collected, and/or refined by annotators, (2) generated by NLP systems through paraphrasing, translation, or prompting, (3) directly inherited from existing datasets, or through a combination of those strategies. Based on the sources of the claims and references, there are three common approaches.

Corpora with real-world claims: In this approach, claims occur naturally and are collected from sources such as social media platforms, news, podcasts, political speeches, or fact-checking archives. References are retrieved with claim-based queries and filtered for relevance by humans or NLP systems. Corpora such as Check-COVID (Wang et al., 2023), MSVEC (Evans et al., 2023), and HealthFC (Vladika et al., 2024) exemplify this method.

Corpora with artificial claims: Here, claims are generated from existing references, such as Wikipedia articles. FEVER (Thorne et al., 2018) pioneered this method by asking annotators to create factual, refuted, and unverifiable claims from Wikipedia sentences, and many CV corpora (e.g., (Jiang et al., 2020; Aly et al., 2021)) follow this paradigm. An example of the generation process is in Appendix B. More recently, corpora such as EX-FEVER (Ma et al., 2024), DIALFACT (Gupta et al., 2022), and FeverFact (Ullrich et al., 2025) used automated transformations or LLM prompting to expand and diversify claim sets. This strategy enables control over label balance, claim complexity, and reasoning types—supporting tasks like multi-hop verification or subclaim decomposition.

Corpus inheritance: In this approach, both claims and references are drawn from existing CV corpora and then cleaned, transformed, or extended. 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).

5 System Development

Of the 198 papers in our survey, 156 build or evaluate CV systems. In this section, we report on the traditional pipeline adopted by many systems and other strategies that go beyond the pipeline.

5.1 The traditional pipeline

The traditional CV systems has four steps.

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

Sentence Selection/Ranking: From the retrieved documents, sentences or snippets pertinent to the claim are selected (used by 68 papers). For instance, 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 (used by 144 papers), this step involves predicting a veracity label based on selected sentences. Recently, there has been a shift from using traditional supervised classifiers to LLM prompting (Guan et al., 2024; Li et al., 2024; Zeng and Gao, 2023; Zhang and Gao, 2023), which often combine evidence retrieved with instruction-tuned prompting (Alvarez et al., 2024).

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

5.2 Other strategies

In addition to the traditional pipeline, other strategies have been proposed for building CV systems. Several common strategies are described below.

Decomposition: A common strategy to handle complex claims is to decompose them into subquestions or subclaims (e.g., (Chen et al., 2024a; Sahu et al., 2024; Schlichtkrull et al., 2023)). Liu et al. (2024a) employed ”Claim Split” modules for this, guiding targeted verification questions (Xu et al., 2024a). However, such atomic units risk losing essential context and they may become ambiguous or unverifiable (Hu et al., 2024). (Gunjal and Durrett, 2024) directly tackled this by defining criteria such as decontextuality (ensuring unique specification for stand-alone status) and minimality (adding only essential context). We will examine claim decomposition more closely in Section 7.

Temporal Reasoning: Claims that mention dates or event order require temporal consistency checks (Mori et al., 2022). Barik et al. (2024a) extracted event–time pairs from both claim and evidence and aligns them on a shared timeline. Barik et al. (2024b) added 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) asked GPT-4 to provide initial explanations, which were then scanned and revised by a second LLM. Kao and Yen (2024a) trained a module to detect rhetorical fallacies (e.g., cherry-picking) and applied fallacy-specific corrections.

5.3 Evaluation practices

Veracity labels produced by CV systems are evaluated with 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 veracity 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 (e.g., evidence-bearing sentences) are evaluated by measuring precision, recall and F1 (Krishna et al., 2022). Abstractive justifications (e.g., explanation) are often evaluated with n-gram overlap-based metrics such as BLEU and ROUGE, alongside semantic similarity scores like BERTScore (Zhang et al., 2024b; Yao et al., 2022).

6 Case Study #1: Claim, Selected Sentences, and Veracity

As discussed in Section 5.1, 68 out of 156 system development papers in our survey included a sentence selection/ranking module, which identifies *evidence-bearing sentences (EBSs)* in the references. Once EBSs are identified, the veracity label module is a classifier that predicts the label at either the **sentence level** or the **instance level**. That is, the input to the classifier is either a single EBS or all the EBSs together, plus the claim. The majority of the studies (e.g., (Zhang et al., 2023; Momii et al., 2024; Mohammadkhani et al., 2024)) built instance-level classifiers directly, while others (e.g., (Fajcik et al., 2023; Olivares et al., 2023; Özge Sevgili et al., 2024)) created sentence-level classifiers and then obtained instance-level veracity labels by combining sentence-level labels (e.g., through weighted voting).

To better understand the need of sentence selection in the CV pipeline and the difficulty of accurate veracity prediction with EBSs only, in our first case study (CS1), we look into the following questions: **(CS1-Q1)** What is the average number of EBSs per claim in existing CV corpora? If that number is small for a corpus, that implies the CV task on that corpus is relatively easy as only a small number of sentences are relevant to the claim. **(CS1-Q2)** How hard is it for human annotators to determine the veracity label at the sentence level and the instance level? What are the main sources of annotation difficulty? Answering those questions can shed light on the difficulty of EBS-based veracity prediction by CV systems.

6.1 Average number of EBSs per claim

Among the 65 corpora discussed in Section 4, twelve include justification in the form of EBSs, from which we randomly sampled three corpora. They are HealthFC (Vladika et al., 2024), MSVEC (Evans et al., 2023), WiCE (Kamoi et al., 2023).

Table 1 shows the distribution of the number of EBSs per claim. For instance, in MSVEC, no EBS is marked for 35.7% of claims and 19.6% of claims have only one EBS. This table shows that the numbers of EBSs for most claims are indeed very low, which may contrast with real-world scenarios where verifying a claim often requires synthesizing information from multiple sources and multiple pieces of evidence (Ma et al., 2024).

# of EBSs	0	1	2	3	4	≥5
HealthFC	0.0	4.8	19.5	31.9	21.7	22.1
MSVEC	35.7	19.6	17.9	8.9	5.4	12.5
WiCE	3.3	9.7	19.1	22.8	20.2	25.0

Table 1: Case Study #1: The distribution of the number of EBSs per claim in three corpora; the corresponding raw count for each cell is in Table 5, Appendix C.

6.2 Veracity annotation design

To answer **CS1-Q2**, we randomly sampled 50 claims from HealthFC (Vladika et al., 2024) that each have more than one EBS and used them for manual annotation.

The original HealthFC dataset employs a ternary label set {Support, NEI, Refute}. To capture EBSs' different degrees of support or refutation of the claim, we used a more fine-grained label set, as defined in Appendix C. An abridged version of the definitions is as follows:

1 (Support): The EBS(s) strongly confirm or support the claim.

2 (Partially Support): The EBS(s) support some parts or scenarios of the claim, but other parts or scenarios are either unsupported or contradicted.

3 (Undecided): The evidence in EBS(s) is too limited or ambiguous or the evidence contains conflicting information.

4 (Partially Refute): The EBS(s) refute some parts or scenarios of the claim, but not all.

5 (Refute): The EBS(s) strongly refute the claim.

6 (Irrelevant): The EBS(s) are irrelevant to the claim.

Two annotators manually annotated veracity at the sentence level first and then at the instance level, using the same label set as defined above. For instance-level annotation, annotators were asked to ignore sentence-level labels and make the decision based on the claim and all its EBSs as a whole.

6.3 Annotation results

At the sentence level, there are 168 EBSs for the 50 claims combined (i.e., 168 (claim, EBS) pairs). The inter-annotator agreement (IAA) is 98/168 = 58.3% when using the 6 labels; the IAA increases to 125/168 = 74.4% when we use 4 labels (that is, label 1 and 2 are merged, so are label 4 and 5). See Table 6 in Appendix C for details.

At the instance level, the IAA with 4 labels is 38/50 = 76% (see Table 7 in Appendix C). We also compare each annotator's labels with the gold standard labels from HealthFC. Coincidentally, the agreements are also 76% (see Table 2-3).

6.4 Sources of annotation difficulty

As discussed in Section 6.3, both IAAs and the agreement between each annotator and gold standard are 76% or lower. Even after lengthy discussion, the two annotators could not resolve some

	1+2	3	4+5	6	Total
Support	18	1	0	0	19
NEI	2	15	7	1	25
Refute	0	1	5	0	6
Total	20	17	12	1	50

Table 2: Case Study #1: Confusion matrix on instance-level veracity label between the gold labels from HealthFC and labels provided by **Annotator 1**. Row labels are from HealthFC, column labels are from Annotator 1, and each cell shows the number of instances with the row and column labels. Mapping of two label sets: 1+2 = *Support*, 3 = *NEI*, 4+5 = *Refute*, 6 = *Irrelevant*. The agreement is 38/50 = 76%.

	1+2	3	4+5	6	Total
Support	16	3	0	0	19
NEI	0	17	6	2	25
Refute	0	1	5	0	6
Total	16	21	11	2	50

Table 3: Case Study #1: Confusion matrix on instance-level veracity label between the gold labels from HealthFC and labels provided by **Annotator 2**. The agreement is 76% too, purely by coincidence.

of the disagreed cases, indicating that the veracity annotation is quite challenging for humans. There are several reasons for annotation difficulties.

First, veracity annotation often requires domain knowledge. For example, a claim talks about *colorectal cancer*, while its EBS discusses *colon cancer*. Annotators without medical knowledge will not know the relationship between those two cancer terms and have to google the terms first, which results in slower annotation speed and potential lower IAA due to different interpretation of search results.

Second, annotators may differ in their interpretation of expressions such as numerical values (e.g., *"5 out of 100"*), modals (e.g., *"could"*), hedging (e.g., *"give indications"*), and degree adverbs (e.g., *"slightly"*). For instance, a claim states that *"Taking antibiotics speeds up the healing of the infection"*. One of its EBSs says *"Sickness duration: only 5 out of 100 benefit"*. One annotator feels that the EBS *partially supports* the claim because it acknowledges the benefit of taking antibiotics on 5% of the patients, while the other annotator chooses the label *undecided* as she believes the adverb *"only"* in the EBS emphasizes the benefit is very small and might be negligible.

Third, EBSs and sometimes even the claims can be hard to interpret due to missing context. For instance, an EBS may contain a pronoun such as *them* but not its antecedent, making it hard to know what the pronoun refers to. Similarly, without the context, we will not know whether a common noun such as *cancer* in an EBS refers to cancer in general or the same type of cancer mentioned in the claim.

Fourth, instance-level veracity labels cannot always be correctly inferred from the sentence-level labels. For example, a claim states *"health benefits increase with duration of exercise"*. Its two EBSs are *"150 minutes of physical activity per week reduced mortality by 9%"* and *"less than 150 minutes per week can reduce risk of death by 34% compared to inactive people."* One problem with this instance is that it is not clear what is the comparison group in the first EBS due to missing context. Assuming that the comparison group is *inactive people*, we label each EBS as *partially supporting* the claim as exercise reduces mortality in both EBSs. However, at the instance level, two EBSs combined *refute* the claim because more exercise (*150 minutes* vs. *less than 150 minutes*) results in less reduction of mortality (*9%* vs. *34%*).

6.5 Summary

To summarize, this case study demonstrates two points. First, the average numbers of EBSs per claim in the three corpora we examined are very low, which may contrast with real-world scenarios.

Second, veracity annotation at both sentence and instance levels can be quite challenging. To address the first two reasons for annotation difficulties, it is important for corpus designers to provide detailed annotation guidelines that clearly define criteria for interpreting claims and EBSs and the guidelines may need to be tailored to the specific domain of the corpus (e.g., how should annotators handle degree adverbs and numerical expressions in claims and EBSs in a medical CV corpus). The third and fourth reasons for annotation difficulties indicate that the traditional CV pipeline (which selects relevant sentences and then aggregates sentence-level results to obtain the instance-level labels) needs to address the issues of missing context and the complex relationship between sentence-level and instance-level labels. While this case study demonstrates the difficulty of veracity annotation, most of the same challenges also hinder automatic veracity prediction by CV systems.

7 Case Study #2: Claim Decomposition

As discussed in Section 5.2, a common strategy to handle complex claims is to decompose the original claims into subclaims; subclaims are then verified in order to obtain a veracity label for the original claim. While this approach can potentially improve system performance and interpretability of system output, the quality of decomposition remains a key bottleneck (Hu et al., 2024).

In our second case study (CS2), we examine the following questions: **(CS2-Q1)** What is the average number of subclaims per claim in existing CV corpora? **(CS2-Q2)** How are subclaims generated and used in current CV systems? **(CS2-Q3)** What is the quality of decomposition? The first two questions are easy to answer, and the last one requires a close examination.

7.1 Average number of subclaims per claim

Out of the 65 corpora in our survey, twelve provide subclaims for each claim. To answer **CS2-Q1**, we randomly picked three from these twelve corpora; they are CLAIMDECOMP (Chen et al., 2022), WICE (Kamoi et al., 2023), and FACTLENS (Mitra et al., 2024). Table 4 shows the distribution of the number of subclaims per claim; the average number of subclaims per claim in each corpus is relatively small, ranging from 2.7 to 3.9.

# of subclaims	1	2	3	4	≥5	Avg
ClaimDecomp	0	33.6	47.6	16.9	1.9	2.8
FactLens	68.5	14.9	8.3	4.6	3.7	3.9
WiCE	0	50.0	31.9	12.1	6.0	2.7

Table 4: Case Study #2: Subclaim distribution across three datasets, with percentages by subclaim count and the average shown in the last column.

7.2 Generation and usage of subclaims

To answer **CS2-Q2**, we examine how subclaims are generated in these three corpora and how they are later used in the process of predicting the veracity label of the original claims.

FACTLENS, derived from COVERBENCH (Jacovi et al., 2024), sampled complex claims from diverse domains and then generated subclaims by few-shot prompting. The subclaims were then evaluated with metrics such as atomicity, sufficiency, and coverage. WICE, based on Wikipedia claims, also used few-shot prompting to generate subclaims; the quality of subclaims were evaluated manually with measures of *completeness* and *correctness*. CLAIMDECOMP relies on human annotators to create yes/no subquestions from PolitiFact claims and justifications, with quality evaluated on *comprehensiveness* and *conciseness*.

In all three studies, subclaims are verified first and claim-level labels are derived from subclaim-level labels with different aggregation rules: FACTLENS applies a strict veto rule, WICE allows partially-supported, and CLAIMDECOMP uses proportions of "yes" answers.

7.3 Methods for evaluating decomposition quality

Our last question, **CS2-Q3**, concerns the quality of decomposition. Beyond traditional criteria like *correctness* and *completeness*, other criteria such as simplicity mattertoo: subclaims should be easier to verify than the original claim. Due to space limit, here we only focus on correctness of decomposition; that is, whether the conjunction of subclaims is *semantically equivalent* to the original claim. Instead of asking annotators to judge equivalence directly, we identify common decomposition strategies employed by LLMs or humans and note where they may introduce errors. We randomly sampled 50 instances from FACTLENS (Mitra et al., 2024), and identified six common strategies along with the conditions under which each strategy fails (see Appendix D). Next, we have two annotators independently labeled each case for (i) strategies used, (ii) errors introduced, and (iii) semantic equivalence.

7.4 An example of decomposition strategy

Consider the original claim:

> *"Chest wall irradiation is informative after mastectomy and negative node breast cancer."*

which was decomposed into two subclaims:

> SC1: *"Chest wall irradiation is informative after mastectomy"*
> SC2: *"Chest wall irradiation is informative after negative node breast cancer"*

We refer to this as the *coordinating conjunction (CC) strategy*, where the original claim contains a CC phrase *X1 and X2*, and each subclaim is identical to the claim except that the CC phrase is replaced by one of its components (*X1* or *X2*).

This strategy is simple but fails to produce semanti-

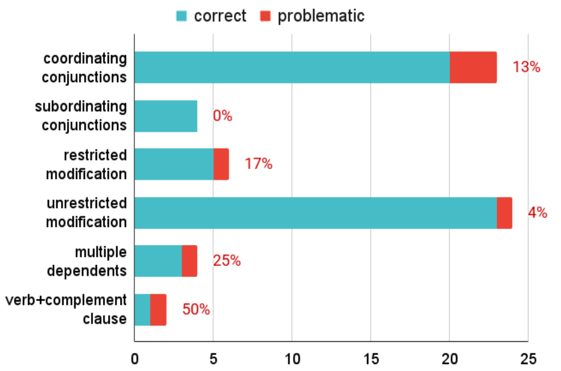


Figure 1: Case study #2: Frequency of decomposition strategy application across 50 FactLens instances (Based on Annotator A). For each strategy, the blue bar represents the number of decompositions that maintained semantic equivalence, while the red bar represents those that violated it. The percentage above each bar indicates the error rate (i.e., red bar / (red bar + blue bar).

cally equivalent subclaims under some conditions: (a) When the CC phrase denotes a single entity (e.g., Barnes & Noble is one bookstore chain, not two). Dropping one component produces an incorrect subclaim. (b) When the CC phrase is ambiguous and decomposition forces one reading. For example, "Smart boys and girls are present" could mean [smart boys + girls] or [smart boys + smart girls]. Splitting into two subclaims assumes one interpretation and discards the other. (c) When the intended meaning is collective rather than distributive. In the irradiation example, if the treatment is informative only after both mastectomy and negative node breast cancer, the two subclaims are not equivalent to the original claim.

7.5 Annotation results

Two annotators independently labeled the 50 instances, reaching 72% agreement on semantic equivalence judgments (36/50; see Table 8 in Appendix D).

Figure 1 summarizes the application frequency and error rate of each decomposition strategy across the 50 annotated instances. Among the six strategies, coordinating conjunctions and unrestricted modification were the most frequently applied. The error rates for the strategies varied widely, ranging from 0% to 50%.

Figure 2 shows the distribution of semantic equivalence judgments across the 50 instances. Decompositions maintained semantic equivalence in 76% of cases. Among the non-equivalent instances, the conjunction of the subclaims entailed the original claim in half of the cases (6 instances) but did not in the other half. Note that Figure 1-2 are based on the annotation from Annotator A. The error rates based on Annotator B's annotation are higher.

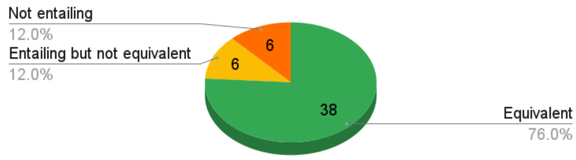


Figure 2: Case study #2: The semantic equivalence judgment on the 50 FactLens Instance (Based on Annotator A).

7.6 Summary

Through a close examination of 50 FactLens instances, we identify six common decomposition strategies and delineate the conditions that lead to a loss of semantic equivalence. A key implication is that the efficacy of decomposition-based claim CV systems is contingent on decomposition quality; erroneous decompositions inevitably propagate to produce erroneous verification results.

8 Challenges and Future Directions

This survey has revealed a number of challenges in corpus creation and system development. In this section, we focus on the most pressing ones.

8.1 Issues with corpus creation

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.

Annotation difficulty: As discussed in Section 6.5, the annotation process is challenging due to various reasons. To mitigate this issue, we recommend the development of detailed, domain-specific annotation guidelines. Furthermore, claims and EBSs are often difficult to interpret in isolation. Corpus designers should therefore consider adding relevant context as a new component of a CV corpus.

Artificial claims: Due to the high cost of manual annotation, a common approach to corpus construction involves generating artificial claims from existing references (see Section 4.3). However, a critical gap in the literature exists, concerning the systematic analysis of the divergence between real-world and artificial claims and the consequential effects of this divergence on the generalization and performance of CV systems.

8.2 Issues with system development

The traditional CV pipeline: Our case study #1 shows that EBSs can be hard to interpret without context, and aggregating sentence-level labels to determine instance-level veracity is error-prone. The traditional CV pipeline needs to evolve to overcome these limitations.

Claim decomposition: Although claim decomposition is a common technique in CV systems, our case study #2 reveals significant limitations. The decomposition process often fails to maintain semantic equivalence between the original claim and the conjunction of its subclaims. Even when equivalence is preserved, some subclaims may be unverifiable given the provided references. Furthermore, many claims resist decomposition using standard strategies. Consequently, further research is crucial to determine when and how decomposition should be applied effectively in CV tasks.

Use of LLMs in CV systems: A growing number of contemporary CV systems are built upon LLMs. A crucial issue is the influence of an LLM's prior knowledge on its judgment, particularly when that knowledge conflicts with the provided reference. Can LLMs temporarily suppress their prior beliefs to objectively verify claims in such conflicting scenarios? More studies like (Xu et al., 2024b) are needed to better understand LLMs' behaviors and adjust the CV systems accordingly.

Shared task, evaluation corpora and deployment: Shared tasks and evaluation corpora heavily shape CV system design. For instance, the AVeriTeC shared task (Schlichtkrull et al., 2024) required systems to incorporate and evaluate question generation and sentence selection modules—components that are not essential to all CV architectures. Similarly, corpora composed of artificial claims, constructed by aggregating information from multiple reference sentences, inherently incentivize the use of claim decomposition strategies. Since the ultimate objective of CV research is to verify real-world claims, future work should prioritize evaluating system performance in realistic deployment scenarios and streamlining implementation for practical use.

9 Related Work

Our main collection of studies includes eight survey papers. Three of them (Bhuiyan et al., 2025; Guo et al., 2022; Yang et al., 2024) provided overviews of the CV systems. Two surveys adopted a more focused perspective: Eldifrawi et al. (2024) specifically examined methods on justification generation; Dmonte et al. (2024) concentrated on the integration of LLMs into CV systems. Another two surveys (Panchendrarajan and Zubiaga, 2024; Gusdevi et al., 2024) examined CV systems in non-English and region-specific contexts and one additional survey (Akhtar et al., 2023b) focused on multimodal verification approaches.

The scope and focus of our survey differ from previous work; it systematically reviews literature pertaining to both corpora construction and system development. To ground this review, we also conducted two case studies that elucidate outstanding research challenges.

10 Conclusion

Our survey of 198 papers (January 2022 - March 2025) provides a detailed analysis of recent advancements in claim verification (CV), focusing on both corpus creation and system design. Through two case studies, we first highlight the difficulties of veracity annotation and the limitations of traditional CV pipelines, and then identify common decomposition strategies along with their associated pitfalls. Our analysis culminates in a discussion of remaining challenges and proposed future directions for the field.

In contrast to the predominant focus in the NLP field on novel system design, our findings underscore the critical importance of data-centric analysis—meticulously examining, annotating, and understanding the data itself. Our case studies demonstrate how such analysis reveals fundamental limitations in existing methodologies. Addressing these limitations will be a primary focus of our future research.

Limitations

This survey included only papers in English published from January 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 screening and annotation stage; thus, annotation errors or inconsistencies are inevitable.

Next, due to page limits for submission, while 198 papers are included in this survey from which we gathered our statistics, only a small subset of them are discussed individually in our paper.

Finally, due to the high cost of manual annotation, we limited double annotation to 50 instances per case study.

Ethical Consideration

All publications included in this survey and the corpora utilized for the case studies are publicly accessible. The authors carried out the screening procedure detailed in Section 3 and manual annotation in the two case studies. We discern no ethical concerns associated with this research.

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)*.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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)*.

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*.

Martin Fajcik, Petr Motlicek, and Pavel Smrz. 2023. Claim-dissector: an interpretable fact-checking system with joint re-ranking and veracity prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*.

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*.

Alon Jacovi, Moran Ambar, Eyal Ben-David, Uri Shaham, Amir Feder, Mor Geva, Dror Marcus, and Avi Caciularu. 2024. Coverbench: A challenging benchmark for complex claim verification. *arXiv preprint arXiv:2408.03325*.

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. Ex-fever: 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.

Mohammad Ghiasvand Mohammadkhani, Ali Ghiasvand Mohammadkhani, and Hamid Beigy. 2024. Zero-shot learning and key points are all you need for automated fact-checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*.

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)*.

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.

Daniel Guzman Olivares, Lara Quijano, and Federico Liberatore. 2023. Enhancing information retrieval in fact extraction and verification. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*.

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, Ganeshan Malhotra, Md Shad Akhtar, Shubhashis Sengupta, Andrew Fano, and Tanmoy Chakraborty. 2022. 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. 2024a. 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)*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024b. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565.

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*.

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*.

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*.

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)*.

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 3 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[6] and a table on the second article[7]. 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 additional materials for case study #1.

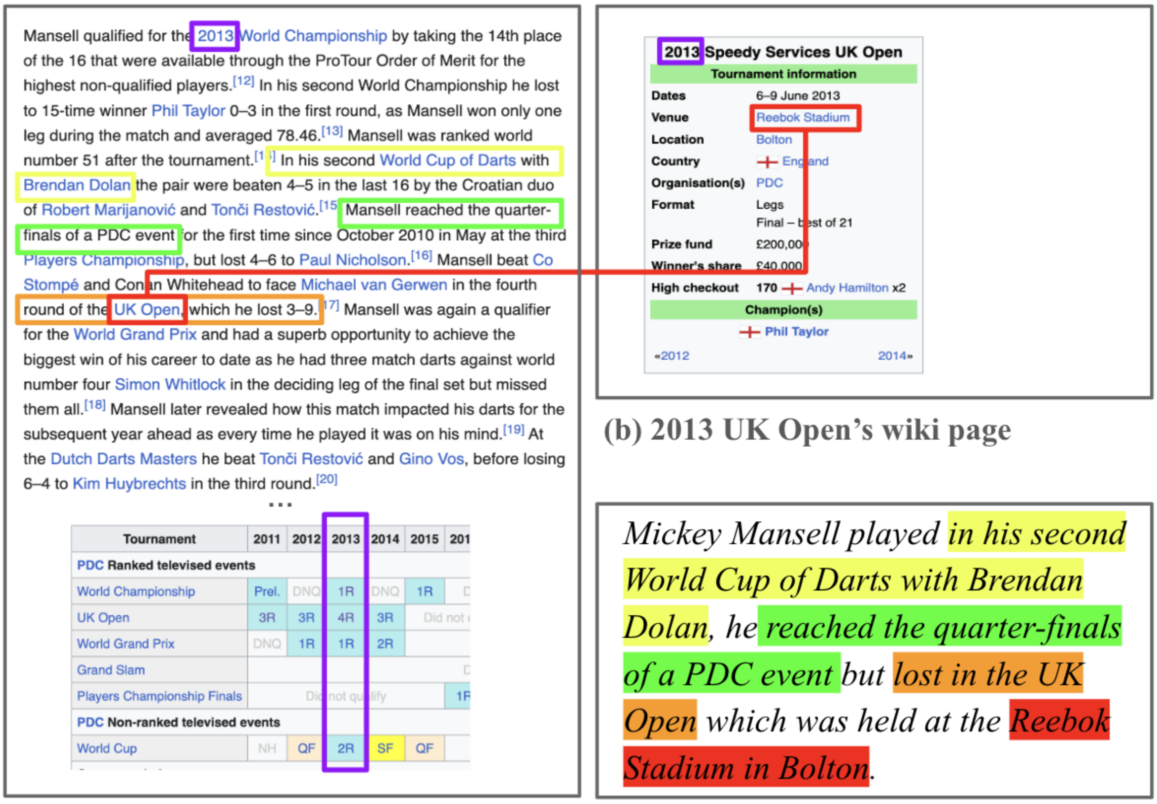
C.1 Number of EBSs per claim in Corpora HealthFC, MSVEC and WiCE

Table 5 shows the distribution of the number of EBSs per claim in Corpora HealthFC, MSVEC, and WiCE.

C.2 The veracity labels used in our manual annotation

Below are the full definitions of the veracity labels used for our human annotation:

[6]https://en.wikipedia.org/wiki/Mickey_Mansell
[7]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 3: An artificial claim from the Feverous corpus, which was generated by combining information from two Wikipedia articles

# of EBSs	0	1	2	3	4	≥5
HealthFC	0	36	146	239	163	166
MSVEC	20	11	15	3	7	12.5
WiCE	242	699	1379	1648	1460	1807

Table 5: Case Study #1: The distribution of the number of EBSs per claim in three corpora. It is identical to Table 1 except for that the cells here show the raw counts, not the percentages.

1 (Support): the EBS strongly confirms or support the claim. This means the evidence is clear, direct, and comprehensive in validating the claim. No major aspects of the claim are left unaddressed, and the support does not rely on weak inference or speculation.

2 (Partially Support): the EBS backs up some parts or scenarios of the claim, but other parts or scenarios are either unsupported or contradicted. In this case, the evidence may be specific to certain conditions; the evidence is from one single study; or the sentence uses hedging, resulting in a somewhat uncertain tone. The overall stance leans supportive, but gaps or inconsistencies prevent full confirmation. For example, if the claim is asking about the benefit of some treatment, an EBS says states one study shows benefits would be partially support.

3 (Undecided): The evidence in EBS is too limited or ambiguous to judge or the evidence contains conflicting information. Here, the EBS might specifically state that the conclusion cannot be reached. Or the evidence might be vague, incomplete, or equally open to multiple interpretations.

4 (Partially Refute): the EBS refutes some parts or scenarios of the claim, but not all. In this case, the evidence may highlight limitations, negative results, or contradictory findings that apply only under certain conditions; the evidence might come from a single study or source that challenges the claim; or the wording may emphasize exceptions or caveats, giving the evidence a somewhat skeptical tone. The overall stance leans negative, but it does not amount to a full rejection. For example, if the claim is that a treatment is effective, and the EBS states that one study found no benefit in a specific subgroup, this would be partially refuted.

5 (Refute): the EBS strongly refutes the claim. This indicates the evidence clearly and directly contradicts the claim in a broad and decisive way. The refutation is comprehensive and applies to the claim as a whole.

6 (Irrelevant): the EBS is irrelevant to the claim. The evidence neither supports nor refutes the claim, often because it addresses a different topic, is too general, or provides information unrelated to the central issue.

C.3 Sentence-level and instance-level IAA for veracity annotation

Table 6 and Table 7 show the confusion matrix between two annotators for veracity annotation, at the sentence and instance levels, respectively.

At the sentence level, there are 168 EBSs for the 50 claims (i.e., 168 (claim, EBS) pairs). The inter-annotator agreement (IAA) is 98/168 = 58.3% when using the 6 labels; the IAA increases to 125/168 = 74.4% when we use 4 labels (that is, label 1 and 2 are merged, so are label 4 and 5). At the instance level, the IAA is 30/50 = 60% when using the 6 labels; it increases to 76% when using the 4 labels.

	1	2	3	4	5	6	Total
1	3	18	2	1	0	0	24
2	2	27	11	0	0	0	40
3	0	0	52	5	0	0	57
4	0	0	2	9	1	2	14
5	0	1	0	6	5	2	14
6	0	2	15	0	0	2	19
Total	5	48	82	21	6	6	168

Table 6: Case study #1: Confusion matrix between two annotators for veracity annotation at the **sentence** level. Rows correspond to Annotator 2's labels, and columns correspond to Annotator 1's label. Each cell shows the number of **(claim, EBS) pairs** with the corresponding labels. Label 1-6 are defined in Appendix C.2. The IAA is 58.3% with 6 labels and 74.4% with 4 labels (that is, label 1 and 2 are merged, so are label 4 and 5).

D Details of Case Study #2

D.1 Decomposition strategies

Below are six common decomposition strategies that we have identified from the 50 FactLens instances.

Coordinating Conjunction (CC): One of the most common decomposition strategies is to split a coordinating conjunction phrase "*X1 and X2*" in the

	1	2	3	4	5	6	Total
1	3	6	0	0	0	0	9
2	0	6	4	1	0	0	11
3	0	1	14	2	0	0	17
4	0	0	3	3	0	0	6
5	0	0	0	2	3	1	6
6	0	0	0	0	0	1	1
Total	3	13	21	8	3	2	50

Table 7: Case study #1: Confusion matrix between two annotators for veracity annotation at the **instance** level. Rows correspond to Annotator 2's labels, and columns correspond to Annotator 1's labels. Each cell shows the number of **instances** with the corresponding labels. The IAA is 60% with 6 labels and 76% with 4 labels (that is, label 1 and 2 are merged, so are label 4 and 5).

original claim so that each subclaim is identical to the original claim except that the CC phrase is replaced with either *X1* or *X2*.

This strategy is generally safe when the CC phrase is the same as the original claim (e.g., the claim is "Paris is the capital of France and London is the largest city in the UK"). However, this strategy becomes problematic when the CC phrase is a noun phrase due to collective vs. distributive readings of plural expressions (see Section 7.4).

In addition, CC phrases tend to lead to more syntactic ambiguities. For instance, in the expression *"A and B of C"*, the PP phrase *"of C"* may modify either B only or *"A and B"* together. In order to decide whether the subclaim set should be {*"A", "B of C"*} or {*"A of C", "B of C"*}, the decomposition process will be forced into resolving the PP attachment ambiguity first. Failure in PP attachment disambiguation will lead to decomposition errors.

Subordinating conjunctions: In this case, the original claim contains subordinating clauses connected a subordinating conjunction (SC) such as *"S1 SC S2"*. If the subclaim set includes only S1 and S2, the connection between the two clauses expressed by the SC will be lost with this decomposition. Even worse, if the SC is a word such as *when* or *if*, decomposing the original claim into {*S1, S2*} is simply wrong because "*if S1, S2*" being true does not entail that both *S1* and *S2* are true and vice versa.

Restricted Modification: In this case, the original claim includes a head with a restricted modifier, e.g., a noun phrase followed by a restricted relative clause. This decomposition strategy will form two subclaims, one with the modifier removed from the original claim and the other turns the head and the modifier into a sentence. For instance, the claim *"The textbook required by CS101 is very expensive"* is decomposed into {*"The textbook is very expensive", "The textbook is required by CS101 course"*}. This decomposition strategy can be problematic since it is not clear which textbook the subject of each subclaim refers to. Removing restricted modifiers changes the scope of the head that is being modified.

Unrestricted Modification: In this case, the claim includes a head with an unrestricted modifier, such as a noun phrase modified by an unrestricted relative clause. This decomposition strategy will form two subclaims, one with the modifier removed from the original claim and the other turns the head and the modifier into a sentence. For instance, the claim *"John Smith, the CEO of Disney, visited Boston in 2023"* is decomposed into {*"John Smith visited Boston in 2023", "John Smith was/is the CEO of Disney"*}. As removing unrestricted modifier will not change the scope of the head being modified, this decomposition strategy seems to be safe. However, when the modifier is an appositive expression or a reduced relative clause, the *be*-verb has to be inserted into the second subclaim and determining the tense of the *be*-verb can be tricky.

Multiple Dependents: When the claim contains a head with multiple modifiers, this strategy splits these dependents into separate subclaims. For example, the claim *"John lived with his wife in Chicago for two years"* is decomposed into: *"John lived with his wife", "John lived in Chicago for two years"*. While the strategy seems to preserve semantic equivalence, upon close examination, it does not. For instance, imagine that John lived in Chicago by himself for two years, got married and then lived with his wife in Seattle for 10 years. Under this scenario, both subclaims are true but the original claim is not. The reason for the loss of equivalence is that the PP *"for two years"* modifies the phrase *"lived with his wife in Chicago"* in the original claim, but modifies only *"lived in Chicago"* in the second subclaim.

Verb + complement Clause: In this case, the original claim includes a verb with a complement clause as its subject or object. The decomposition strategy treats the complement clause as a subclaim. For instance, suppose the claim is of the form "Subject V S1", whether the claim entails S1 depends on what type of verb V is. If V is a mental state verb such as *believe* or a communication verb such as *argue*, the original claim will not entail S1 (e.g., "John believes/argues that the earth is flat" does not entail "the earth is flat"). On the other hand, if V is a factive verb such as *regret, realize, forget*, the claim is likely to entail S1 (e.g., "John regretted that he missed the meeting" entails that "he missed the meeting").

D.2 Inter-Annotator Agreement for Equivalence labeling

Figure 8 shows the confusion matrix for semantic equivalence judgment on the 50 FactLens instances.

	1	2a	2b	Total
1	28	0	10	38
2a	1	3	2	6
2b	1	0	5	6
Total	30	3	17	50

Table 8: Case study #2: Confusion matrix between two annotators for the semantic equivalence judgment of the 50 FactLens instances. Row labels are from Annotator A, and column labels come from Annotator B. Meanings of the labels: 1 = *Semantic equivalence*, 2a = *Conjunction of subclaims entails the claim, but not vice versa*, 2b = *Conjunction of subclaims does not entail the claim*. The IAA is 36/50 = 72%.

One factor contributing to the low IAA was a discrepancy in applying the subordinating clause strategy. Annotator B adhered to a stricter interpretation of semantic equivalence. For example, when decomposing the claim "*S1, resulting in S2*" into the subclaims $\{S1, S2\}$, Annotator B argued that semantic equivalence was not maintained because the causal relationship between S1 and S2 was lost. In contrast, Annotator A adopted a more lenient view. Under Annotator B's strict definition, preserving equivalence would require adding a new subclaim explicitly stating the causal relation, which would often closely resemble the original claim. This highlights the need for a more precise and operationalized definition of semantic equivalence in the annotation guidelines.

E Details of the New CV Corpora in Our Survey

Table 9-12 provide more information on the 65 corpora discussed in Section 4, all with the following columns:

Corpus Name: This is the name of the CV corpus the paper created.

Corpus size: This is the number of instances in the corpus:

- 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: A corpus may have one or more modalities: 1 = text, 2 = image, 3 = video, 4 = audio, 5 = chart, 6 = table.

Language: We used the 3-letter ISO 639-2 language codes for individual languages: ara = Arabic, ben = Bengali, chi = Chinese, cze = Czech, eng = English, fre = French, ger = German, ind = Indonesian, ita = Italian, jpn = Japanese, nor = Norwegian, rus = Russian, spa = Spanish, tur = Turkish, ukr = Ukrainian, vie = Vietnamese, plus low = low-resource languages and mult = multilingual.

Source: This column shows the source of data used by the CV corpus.

Veracity: This column shows the veracity label set used by the corpus:

- 1 = binary labels (true, false),
- 2 = ternary labels (supported, refuted, NEI),
- 3 = more than 3 labels,

Justification: This column indicates the type of justification:

- 0 = the corpus has no justification field
- 1 = evidence-bearing sentences (EBSs)
- 2 = summary of EBSs
- 3 = free-form explanation

Link: This is the link to access the dataset.

Corpus Name	Corpus Size	Modality	Language	Source	Veracity	Justification	Link
2024 Presidential Debate Claims (Nanekhan et al., 2025)	1	1	eng	presidential debates	1	1	link
Bangla Claim Detection Dataset (Rahman et al., 2025)	4	1	ben	fact-checking websites, interviews, speeches	1	0	Available upon request
CorFEVER (Tan et al., 2025)	2	1	eng	online sources	2	3	link
Fact-Checking Podcasts Dataset (Setty and Becker, 2025)	1	1,4	eng, ger, nor	podcast episodes	1	0	link
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
MultiSynFact (Chung et al., 2025)	5	1	eng, ger, low, spa	Wikipedia	2	1	link
Adversarial CHEF (Zhang et al., 2024a)	2	1	chi	CHEF	2	3	link
AMBIFC (Glockner et al., 2024)	5	1	eng	BoolQ dataset	2	0	link
AuRED (Haouari et al., 2024)	1	1	ara	Twitter	2	0	link
BINGCHECK (Li et al., 2024)	3	1	eng	ChatGPT prompted user queries	3	0	N/A
CFEVER (Lin et al., 2024)	5	1	chi	Wikipedia	2	0	link
ChartCheck (Akhtar et al., 2024)	5	1, 5	eng	Wikipedia Commons	2	3	link
CHEF-EG, TrendFact (Zhang et al., 2024b)	4	1	chi	CHEF, Weibo	2	3	N/A
ChronoClaims (Barik et al., 2024a)	5	1	eng	Wikipedia	2	1	N/A
CLAIMREVIEW2024+ (Braun et al., 2024)	1	1, 2	eng	ClaimReview Project	3	0	link

Table 9: Claim Verification Corpora in Our Collection (1 of 4).

Corpus Name	Corpus Size	Modality	Language	Source	Veracity	Justification	Link
CREDULE (Chrysidis et al., 2024)	5	1	eng	MultiFC, Politifact, PUBHEALTH, NELA-GT, Fake News Corpus	3	3	link
EX-Claim (Zeng and Gao, 2024)	4	1	eng	WatClaim Check	1	3	link
EX-Fever (Ma et al., 2024)	5	1	eng	Wikipedia	2	3	link
Factify5WQA (Suresh et al., 2024)	5	1	eng	fact-checking datasets	2	1	link
FactLens (Mitra et al., 2024)	2	1	eng	CoverBench	1	1,3	N/A
FCTR (Cekinel et al., 2024)	3	1	tur	fact-checking organization, Snopes	3	2	link
FEVER-it (Scaiella et al., 2024)	5	1	ita	FEVER	2	0	link
FINDVER (Yilun Zhao et al., 2024)	3	1, 6	eng	company reports through U.S. Securities and Exchange Commission	1	3	link
FlawCheck (Kao and Yen, 2024a)	5	1	eng	WatClaimCheck	3	0	link
HealthFC (Vladika et al., 2024)	2	1	eng, ger	Medizin Transparent web portal	2	1, 2	link
LLMforFV (Guan et al., 2024)	2	1	eng	LLM-generated text with human annotations	1	0	link
Multi-News-Fact-Checking (Chen et al., 2024b)	5	1, 2	eng	Multi-News summarization dataset	3	2, 3	link
QuanTemp (Venktesh et al., 2024)	5	1	eng	Google Fact Check Tools API	2	0	link
RU22Fact (Zeng et al., 2024)	5	1	chi, eng, rus, ukr	fact-checking websites, news outlets	2	3	link
T-FEVER, T-FEVEROUS (Barik et al., 2024b)	5	1	eng	FEVER, FEVEROUS	2	1	N/A
TrendFact (Zhang et al., 2024c)	5	1	chi	social media, fact-checking websites	2	2, 3	link
ViFactCheck (Hoa et al., 2024)	4	1	vie	newspapers	2	1	link
ViWikiFC (Le et al., 2024)	5	1	vie	Wikipedia	2	0	link

Table 10: Claim Verification Corpora in Our Collection (2 of 4).

Corpus Name	Corpus Size	Modality	Language	Source	Veracity	Justification	Link
XClaimCheck (Kao and Yen, 2024b)	5	1	eng	WatClaimCheck, PolitiFact	3	0	link
UNK (Tan et al., 2024)	5	1	eng	reports from National Transportation Safety Board	1	0	N/A
AVeriTeC (Schlichtkrull et al., 2023)	3	1	eng	fact-checking organizations	3	3	link
ChartFC (Akhtar et al., 2023a)	5	1, 5	eng	TabFact	1	0	link
Check-COVID (Wang et al., 2023)	3	1	eng	scientific journal articles	2	0	link
COVID-VTS (Liu et al., 2023)	4	1, 3	eng	Twitter	1	1, 3	link
CsFEVER, CTKFacts (Ullrich et al., 2023)	5	1	cze	Czech adaptation of the English FEVER	3	1	link
EFact (Hu et al., 2023)	4	1	eng	fact-checking organization	3	0	N/A
Facity 2 (Suryavardan et al., 2023)	5	1, 2	eng	Twitter	3	0	link
FACTIFY 3M (Chakraborty et al., 2023)	5	1, 2	eng	Internet-collected stories paraphrased by ChatGPT	3	2, 3	N/A
FACTIFY-5WQA (Rani et al., 2023)	5	1	eng	fact verification datasets	2	1, 3	link
FACTKG (Kim et al., 2023)	5	1	eng	WebNLG datase	1	0	link
Fin-Fact (Rangapur et al., 2023)	3	1, 2	eng	PolitiFact, Snopes, FactCheck	2	3	link
German healthcare news articles (Gupta et al., 2023)	1	1	eng, ger	German news sources	1	1	N/A
LIAR++; FullFact (Russo et al., 2023)	4	1	eng	LIAR-PLUS, FULL-FACT website	2	3	link
MSVEC (Evans et al., 2023)	1	1	eng	news outlets, fact-checking websites	1	1	link
Multi2Claim (Tan et al., 2023)	5	1	eng	scientific multiple-choice QA datasets	2	3	link
MultiClaim (Pikuliak et al., 2023)	5	1	mult	Google Fact Check Explorer, Snopes	1	0	Available upon request
SCITAB (Lu et al., 2023)	3	1, 6	eng	Sci-Gen dataset	2	0	link

Table 11: Claim Verification Corpora in Our Collection (3 of 4).

Corpus Name	Corpus Size	Modality	Language	Source	Veracity	Justification	Link
WICE (Kamoi et al., 2023)	3	1	eng	Wikipedia	2	1	link
X-Fact (Hu et al., 2023)	5	1	mult	fact-checking organization	3	0	N/A
XFEVER (Chang et al., 2023)	5	1	chi, eng, fre, ind, jpn, spa	FEVER	2	0	link
CHEF (Hu et al., 2022)	5	1	chi	news review sites	2	0	link
ClaVer (Sundriyal et al., 2022)	3	1	eng	CORD-19, LESA	2	0	link
Custom COVID-19 Claims Dataset (Casillas et al., 2022)	3	1	eng	WHO Mythbusters, Johns Hopkins FAQs, CNN QA pages	1	0	link
DIALFACT (Gupta et al., 2022)	5	1	eng	Wikipedia	2	1	link
FACTIFY (Mishra et al., 2022)	5	1, 2	eng	Twitter	3	0	link
FAVIQ (Park et al., 2022)	5	1	eng	Natural Questions dataset, AmbigQA	1	0	link
FC-Claim-Det (Bhatnagar et al., 2022)	1	1	eng	Fact-checked articles	2	2, 3	link
Mocheg (Yao et al., 2022)	5	1, 2	eng	PolitiFact, Snopes	2	1	link
PubHealthTab (Akhtar et al., 2022)	3	1, 6	eng	fact-checking, news review websites	1	0	link
SCIFACT-OPEN (Wadden et al., 2022)	5	1	eng	SCIFACT-ORIG test set	2	1	link
SufficientFacts (Atanasova et al., 2022)	2	1	eng	FEVER, Vitamin C, HoVer	2	0	link

Table 12: Claim Verification Corpora in Our Collection (4 of 4).