# The Alignment Bottleneck in Decomposition-Based Claim Verification

**Mahmud Elahi Akhter**[1] , **Federico Ruggeri**[2] , **Iman Munire Bilal**[3] , **Robert Procter**[3,4] and **Maria Liakata**[1,4]
[1]Queen Mary University of London, UK
[2]University of Bologna, Italy
[3]University of Warwick, UK
[4]The Alan Turing Institute, UK
{m.akhter, m.liakata}@qmul.ac.uk, federico.ruggeri6@unibo.it, {iman.bilal, rob.procter}@warwick.ac.uk

## Abstract

Structured claim decomposition is often proposed as a solution for verifying complex, multi-faceted claims, yet empirical results have been inconsistent. We argue that these inconsistencies stem from two overlooked bottlenecks: evidence alignment and sub-claim error profiles. To better understand these factors, we introduce a new dataset of real-world complex claims, featuring temporally bounded evidence and human-annotated sub-claim evidence spans. We evaluate decomposition under two evidence alignment setups: Sub-claim Aligned Evidence (SAE) and Repeated Claim-level Evidence (SRE). Our results reveal that decomposition brings significant performance improvement only when evidence is granular and strictly aligned. By contrast, standard setups that rely on repeated claim-level evidence (SRE) fail to improve and often degrade performance as shown across different datasets and domains (PHEMEPlus, MMM-Fact, COVID-Fact). Furthermore, we demonstrate that in the presence of noisy sub-claim labels, the nature of the error ends up determining downstream robustness. We find that conservative "abstention" significantly reduces error propagation compared to aggressive but incorrect predictions. These findings suggest that future claim decomposition frameworks must prioritize precise evidence synthesis and calibrate the label bias of sub-claim verification models.

## 1 Introduction

The task of *claim verification* involves determining the *veracity* for a given *claim* based on associated *evidence*. It is crucial for a number of different tasks, such as rumour or misinformation verification and fact-checking [Zubiaga *et al.*, 2018; Glockner *et al.*, 2022; Guo *et al.*, 2022; Pan *et al.*, 2023]. Claim verification is particularly challenging when dealing with complex claims which conflate multiple facts, requiring granular reasoning over evidence for accurate verification. To verify complex claims, humans identify several sub-parts (sub-claims), gather multiple pieces of evidence and perform multi-step reasoning [Nakov *et al.*, 2021].

In such cases mapping the claim to a complete set of evidence is non-trivial; A common approach is to disentangle the sub-claims from the initial claim, (claim decomposition), which is then followed by sub-claim verification and rationale generation for each resulting sub-claim [Chen *et al.*, 2022; Pan *et al.*, 2023]. Although laborious, this process offers more transparency in the decision-making process, especially in cases of insufficient or contradictory evidence relevant to different parts of a claim. However, this introduces an additional error propagation factor since errors in sub-claim decomposition or sub-claim labels can mislead claim-level inference.

Performing claim verification using Large Language Models (LLMs) [Wang and Shu, 2023; Quelle and Bovet, 2023] has recently gained popularity. Apart from single LLMs, Agentic systems have also been applied to perform complex claim verification [Ma *et al.*, 2025; Liu *et al.*, 2025]. However, these still inherit the error propagation caveats due to decomposition and misalignment of granular evidence.

Real world claim verification requires complex multi-hop reasoning steps. Reasoning with LLMs suffers from important shortcomings. For instance, [Dougrez-Lewis *et al.*, 2025] show that LLMs struggle with abductive reasoning while LLM performance on claim verification is very much domain-dependent. Specifically, LLMs are better at handling simple claims from Wikipedia compared to real-world data which manifest complex claims and hierarchical structure. This leads to a major bottleneck in claim verification and fact-checking since most datasets, where the state-of-the-art (SOTA) is trained and evaluated, are sourced from Wikipedia [Guo *et al.*, 2022]. Studies that have evaluated the benefits of claim decomposition for fact-checking report conflicting results [Tang *et al.*, 2024; Wanner *et al.*, 2024; Chen *et al.*, 2024]. According to [Tang *et al.*, 2024] claim decomposition is not necessary to improve LLM performance on fact checking. By contrast, [Wanner *et al.*, 2024] show that the impact of decomposition is method-specific, especially when considering LLM-based decomposition. [Hu *et al.*, 2024] carried out an in-depth analysis of claim decomposition, revealing several deciding factors, including how the decomposition process is defined and how evidence information is associated with each sub-claim. Thus, the benefits of claim decomposition for complex claims are currently poorly understood. We argue these conflicting findings largely stem from two factors: (i) how evidence is aligned and aggregated at the sub-claim

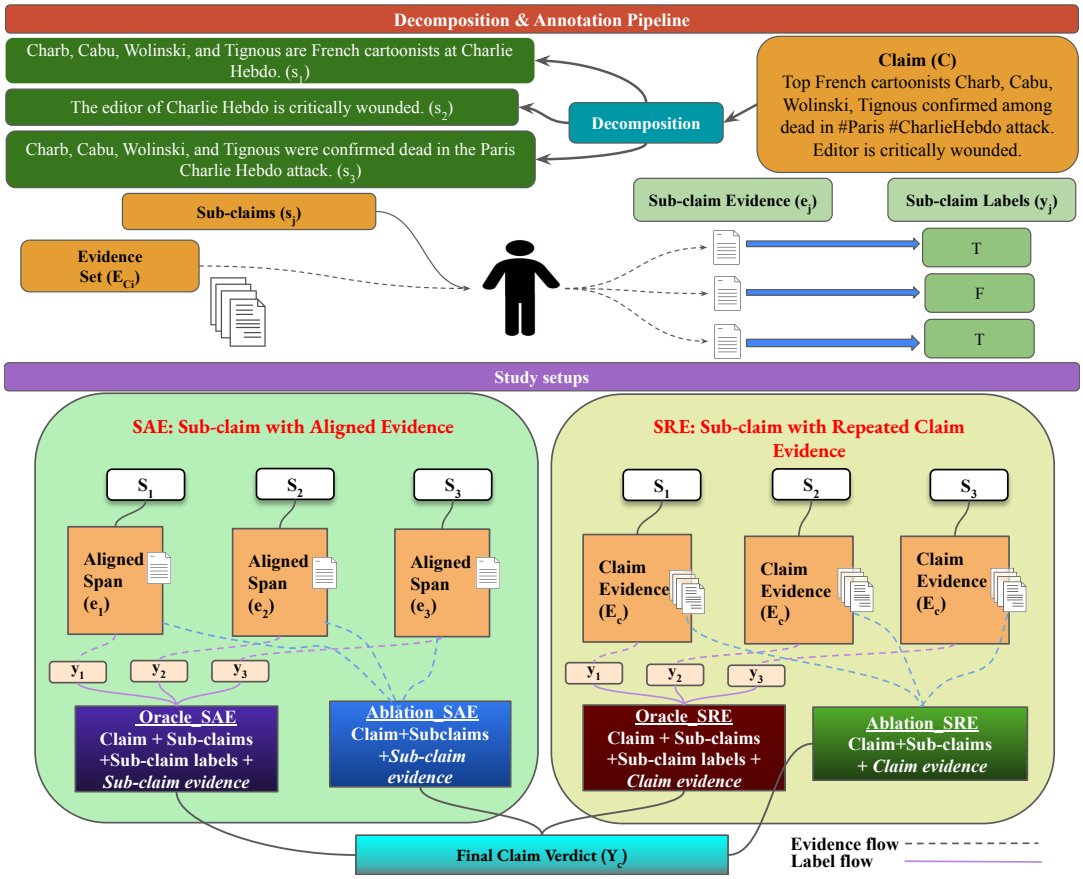


Figure 1: Our annotation and claim verification pipeline and different setups for the study. Oracle_(SAE/SRE) setups use gold sub-claim labels, ablation models do not use any sub-claim labels and noisy setup (not shown in figure) uses predicted sub-claim labels. We use predicted labels instead of oracle labels to evaluate the influence sub-claims have on claim veracity prediction. For that we swap the oracle labels to predicted labels in Oralce_SAE & Oracle_SRE.

level, and (ii) whether sub-claim veracity signals are reliable or noisy. Both of these can directly affect error propagation to claim-level inference. Based on this, we hypothesize that decomposition helps only when evidence is aligned at the sub-claim level, and sub-claim veracity labels are reliable; otherwise, decomposition can degrade performance in claim verification.

Here we investigate how sub-claim decomposition impacts complex claim verification in real-world data. Fig. 1 illustrates our approach. We assume that a complex claim $C$ corresponding to an evidence set $E_C$ is decomposed into sub-claims $\{s_j\}_{j=1}^{m}$, each of which are paired with corresponding evidence sets $\{E_j\}_{j=1}^{m}$ and annotated with sub-claim veracity labels $\{y_j\}_{j=1}^{m}$. These structured triples $(s_j, E_j, y_j)$ serve as inputs for sub-claim-level prediction and, when aggregated, form the basis for claim-level veracity inference under both oracle and noisy labels. Our approach connects human annotation, claim decomposition, and model evaluation into a unified framework for complex claim verification. We show that decomposition can improve performance under ideal labels (oracle), but can be neutral or harmful when labeling is noisy. Furthermore, to test the generalisability of the decomposition effect, we evaluate the same pipelines not only on temporally bounded rumour verification ( [Dougrez-Lewis *et al.*, 2022]), but also on two additional fact-checking datasets (MMM-Fact [Xu *et al.*, 2025] and COVID-Fact[Saakyan *et al.*, 2021]), which differ with respect to domain and evidence characteristics. We make the following contributions,

- We introduce a dataset of real-world claims decomposed into sub-claims accompanied by manually annotated sub-claim-level evidence and fine-grained veracity labels (§3).
- We formalize two complementary experimental settings (i) *claim veracity* under oracle sub-claim labeling (upper bound) (§5.1) and under predicted sub-claim labeling (real-world scenario) (§5.4),and (ii) *sub-claim veracity* (§5.3) as the upstream task; together these quantify how sub-claim labeling errors propagate to claim-level verification.
- We show decomposition improves performance under sub-claim aligned evidence and reliable sub-claim signals, but can also be neutral or harmful under noisy labeling and coarse evidence. Thus, we highlight that the benefits of decomposition are conditional, dependent on

evidence quality and sub-claim error signals.

## 2 Related Work

**Automatic Claim Decomposition.** Several works have proposed decomposing fact-checking claims into atomic text segments. Such segments (sub-claims/atomic facts/atomic propositions), are typically extracted by prompting LLMs [Zhang and Gao, 2023], via hybrid approaches relying on keywords [Gong *et al.*, 2024] or predefined templates [Pan *et al.*, 2023; Liu *et al.*, 2024]. Prompts are either textual propositions [Kamoi *et al.*, 2023; Tang *et al.*, 2024] or queries for information retrieval [Chen *et al.*, 2022; Chen *et al.*, 2024]. Apart from prompting methods, orchestrated multi-agent systems have also been used for decomposition-based claim verification [Zhao *et al.*, 2024; Ma *et al.*, 2025]. [Hu *et al.*, 2024] provide valuable insights for evaluating the effects of claim decomposition. However, their work pertains to datasets containing automatically extracted sub-claim evidence and does not include real-world complex claims, unlike the task we address here.

**Claim Decomposition Datasets.** Researchers have applied claim decomposition for several purposes, including fine-grained evidence retrieval [Kamoi *et al.*, 2023; Zhang and Gao, 2023], and factual verification of generated content to mitigate hallucinations [Min *et al.*, 2023; Wang, 2024; Zhang *et al.*, 2024].

To this effect, several fact-checking datasets containing sub-claims have been proposed. Notable examples are BingCheck, for evaluating the actuality of LLM-generated texts in response to fact-checking [Li *et al.*, 2024], LLM-AggreFact [Tang *et al.*, 2024] containing a suite of fact-checking corpora, ClaimDecomp centered on extracting explicit and implicit sub-claims [Chen *et al.*, 2022], WiCE for fine-grained evidence retrieval on Wikipedia [Kamoi *et al.*, 2023], and AVERITEC for real-world claim verification [Schlichtkrull *et al.*, 2023a].

ClaimDecomp [Chen *et al.*, 2022] and WiCE [Kamoi *et al.*, 2023] are the closest to our setting. In ClaimDecomp, sub-claims are presented as questions to help evaluate the veracity of complex claims extracted from Politifact. Both the claim and the evidence-based justification, written by journalists, are used to reverse-engineer the verification process, thus introducing both literal (produced from the claim) and implied sub-claims (produced from evidence). Here we adhere to a more realistic setting of decomposing claims with no prior knowledge of existing evidence. In WiCE, claim decomposition is carried out via LLMs, and each sub-claim is associated with evidence manually annotated from Wikipedia articles. Similarly, our dataset provides manually annotated evidence texts for sub-claims but unlike WiCE, we focus on real-world claim verification, consisting of social media content, where rumours often emerge and spread [Zubiaga *et al.*, 2018].

## 3 Dataset

### 3.1 Task Overview

Our task involves assessing the veracity of complex claims through structured decomposition into sub-claims, each supported by relevant evidence. We first decompose claims into distinct sub-claims, then independently verify each sub-claim using associated evidence. We then analyze how this pipeline behaves under different evidence association strategies (claim-level vs sub-claim-aligned evidence) and different supervision regimes (oracle vs predicted sub-claim labels), to quantify error propagation to claim-level inference.

### 3.2 Dataset Creation

We filter complex claims within the PHEME dataset [Zubiaga *et al.*, 2016] by removing claims containing fewer than three verbs and two sentences. This heuristic is a proxy method for multi-fact composition introduced by [Chen *et al.*, 2022]. PHEME contains a collection of claims posted on Twitter as time-sensitive breaking news events were unfolding within a specific time-frame. It contains 1972 rumourous claims related to five events and each of the rumours are annotated with a veracity label (True, False, Unverified). To decompose complex claims into sub-claims, we adopted a fixed decomposition procedure to isolate the effects of evidence association and label noise on verification performance, rather than optimizing the decomposition model itself. We used an LLM-based schema introduced by FactScore [Min *et al.*, 2023] with gpt-3.5-turbo-instruct to break down text into short statements covering only one piece of information. This is followed by a manual evaluation process using the claim detection guidelines in [Sundriyal *et al.*, 2022] to ensure that the resulting sub-claims contain check-worthy information and are comprehensive when considered without the original claim. In total, we extract 423 complex claims from PHEME, covering five distinct events. We then leveraged PHEMEplus [Dougrez-Lewis *et al.*, 2022] annotations to map these claims to relevant evidence from news articles. PHEMEplus contains evidence-providing articles for the claims within PHEME in a time constrained manner (within 2 days of each event). We obtained 399 claims after accounting for evidence quality and dataset artifacts in PHEMEplus. We filtered out image-only evidence, hyperlinks and embedded ads. We do not consider claims with associated images in the tweet since decomposing such claims result in non-sense sub-claims (URL links or pictures).

**Domain generalisation.** We additionally use MMM-Fact and COVID-Fact to test generalisability of our findings beyond event-centric rumour verification. MMM-Fact is a large-scale multi-domain fact-checking benchmark with multi-source evidence. It has a full size of 125,449 samples with evidence grouped into retrieval-difficulty levels (Basic/Intermediate/Advanced). We use intermediate level which had 6-10 evidence sources per claim (21,873). COVID-Fact is a domain-specific (health/COVID-19) real-world fact-checking dataset with 4,086 claims paired with single source evidence. Their goal was to test multi-document synthesis (MMM-Fact), to better assess evidence alignment and to see the decomposition and alignment effect in a narrower-domain with less structural complexity (COVID-Fact).

### 3.3 Annotation Details

Given a claim, associated sub-claims and evidence providing articles for the original claim, the task was to label the veracity of the sub-claims and highlight relevant spans of evidence supporting the assigned veracity label. The highlighted evidence

| Category | Split | T% | U% | F% |
|---|---|---|---|---|
| **Claim Veracity** | **Total** | 48.37 | 31.33 | 20.3 |
| | **Train** | 47.98 | 31.78 | 20.25 |
| | **Test** | 50.00 | 29.49 | 20.51 |
| **Sub-claim Veracity** | **Total** | 57.66 | 34.05 | 8.3 |
| | **Train** | 57.11 | 34.70 | 8.19 |
| | **Test** | 59.75 | 31.54 | 8.71 |

Table 1: Label distributions (%) for claims and sub-claims per splits.

was later extracted as sub-claim level evidence. In particular, each sub-claim was labeled as true (T), false (F), and Unverified (U). Starting with 1169 sub-claims, these were evenly distributed to three annotators, Phd students in Computer Science, fluent in English, so that 300 samples were annotated by all. The latter were used to calculate the Inter Annotator Agreement (IAA) score. In short, we double-annotated 300 sub-claims for IAA; the remaining sub-claims were single-annotated and distributed evenly, resulting in ≈290 unique items per annotator. Annotators labeled a sub-claim as (T) if directly supported by the evidence, (F) if contradicted, and (U) if the evidence was insufficient or ambiguous. Evidence spans were selected as the minimal set of sentences justifying the label. We computed IAA as Bennett's S score to account for label imbalance, resulting in a score of 0.81. We also analyzed sentence-level overlap among annotators using BLEU and BERTScore. The Average BLEU[Papineni *et al.*, 2002] ranged from ≈ 0.40–0.48 across pairs, and average symmetric BERTScore [Zhang *et al.*, 2019] F1 was around 0.49–0.56, indicating moderate to strong overlap in selected evidence spans.

Unlike Averitec, [Schlichtkrull *et al.*, 2023b], our verification process was not constructed as a QA process but rather the annotators had to verify sub-claims based on the provided evidence. Contrary to Factcheck-Bench's [Wang, 2024] automatic evidence retrieval, we manually selected evidence spans to isolate the verification problem from retrieval errors[Ge *et al.*, 2025].

Our dataset differs from previous work in that we provide: (i) human-annotated evidence texts per extracted sub-claim rather than evidence automatically extracted by LLMs and (ii) evidence documents per claim with a closed retrieval setting. In particular, we consider only evidence documents that were restricted to articles published within a fixed temporal window around the event [Dougrez-Lewis *et al.*, 2022], preventing leakage from post-hoc reporting and ensuring a realistic real-time verification setting. [Zubiaga *et al.*, 2018]. Furthermore, this controlled evidence setting allows us to isolate how sub-claim evidence alignment and sub-claim label noise affect decomposition-based verification, without conflating effects from open-domain retrieval.

### 3.4 Dataset Details

Dataset statistics are in Table 1, including distribution of veracity labels for both complex claims and sub-claims, with 399 complex claims corresponding to 1169 sub-claims. We used the full dataset for oracle and ablation setups in Tables 2 and 3 with 929 samples for training the baseline and GNN models in §4.3 and 240 as testing. The results in Tables 3, 4, and 5, were calculated using this test set. On MMM-Fact we extracted 1181 (5.40% of intermediate subset) complex claims and 163 complex claims in COVID-Fact (3.99% full dataset). These resulted in 2326 and 447 sub-claims respectively for each dataset. Note that these datasets do not contain manually annotated sub-claim labels or corresponding sub-claim aligned evidence.

## 4 Experiments

While claim verification has increasingly been addressed alongside explanation generation [Atanasova *et al.*, 2020; Jolly *et al.*, 2022; Bilal *et al.*, 2024], here we focus on assessing the merits of claim decomposition given (i) claim-level verification and (ii) sub-claim verification labels, while considering two evidence alignment settings **(SRE Eq.** (5) **vs SAE Eq.** (6)**)** (*§5.3*) and two levels of labeling quality for sub-claims (oracle vs predicted/noisy). Evidence is temporally bounded and pre-linked (no open-domain retrieval) to isolate decomposition effects from open retrieval errors.

We denote as $C$ a complex claim and as $E_C = \{E_{Ci}\}_{i=1}^{n}$ the complete evidence set corresponding to the original claim $C$. $C$ is decomposed into a set of $m$ sub-claims $\{s_j\}_{j=1}^{m}$, each associated with an evidence set $\{E_j\}_{j=1}^{m}$ and a veracity label $y_j \in \{T, F, U\}$. Here, $y_j$ denotes either oracle sub-claim labels (upper bound), predicted labels (realistic), or is omitted in label-free ablation variants (Eq. (7) & Eq. (8)). The overall claim-level veracity $Y_C$ will depend on the veracity and evidence of its sub-claims as follows:

$$Y_C = f(C, \{(s_j, E_j, y_j)\}_{j=1}^{m}) \tag{1}$$

where each sub-claim evidence $E_j$ derives from some claim evidence $E_{Ci}$. To ensure temporal and contextual fidelity, evidence sets are restricted to information available before the claim timestamp:

$$E_{Ci} \subseteq \{e \in \mathcal{D} : t(e) \leq t(C)\} \tag{2}$$

where $\mathcal{D}$ is the document collection, $t(e)$ is the publication time of evidence $e$, and $t(C)$ is the time when the claim $C$ emerged. This constraint prevents open-domain retrieval to avoid compromising the time-sensitive nature of the dataset as it would otherwise retrieve evidence not available at the time of unfolding events; events happening in real time usually generate unverified claims that are later resolved within a few hours to a few days. As our goal was not to propose a new retrieval method, we fix evidence to measure the causal effect of decomposition choices instead of compounding it with retrieval errors.

### 4.1 Claim Veracity

**Experiment details.** Here the task is to verify a claim (label as T/F) given a set of evidence texts, known to be relevant to the claim where evidence documents are pre-linked to each claim via PHEMEPLUS within a temporally bounded window. We do not consider unverified (U) labels and drop them as these were not typically supported by corresponding evidence within the PHEMEPlus subset. All experiments at claim-level are conducted using the Qwen3-14B [Yang *et al.*, 2025] reasoning model.

We first perform claim-level veracity prediction using **oracle sub-claim labels** [Huang *et al.*, 2024; Yang *et al.*, 2024] to establish an upper bound on performance. Here, $y_j$ the gold sub-claim is provided as part of the structured input to test the best-case scenario for decomposition; these are later replaced with predicted labels Eq. (10) or omitted in ablations (Eq. (7) & Eq. (8)) to reflect realistic usage. The model predicts claim veracity as:

$$\hat{Y}_C = \mathcal{M}(C, \mathcal{E}) \tag{3}$$

where $\mathcal{E}$ represents the evidence configuration used. We consider the following evidence alignment configurations per setting: Claim-level evidence ($E_{Ci}$), for the Vanilla and **Sub-claims with Repeated Claim-level Evidence (SRE)** experimental settings, and Sub-claim level evidence ($E_j$) for the **Sub-claims with Aligned Evidence (SAE)** setting:

$$\textbf{(i) Vanilla:} \quad \mathcal{E} = E_C = \{E_{Ci}\}_{i=1}^{n} \tag{4}$$

$$\textbf{(ii) SRE:} \quad \mathcal{E} = (s_j, \{E_{Ci}\}_{i=1}^{n}, y_j)_{j=1}^{m} \tag{5}$$

$$\textbf{(iii) SAE:} \quad \mathcal{E} = \{(s_j, E_j, y_j)\}_{j=1}^{m} \tag{6}$$

In practice, SRE repeats the same claim-level evidence block after each sub-claim in the prompt, whereas SAE inserts the sub-claim-specific evidence block $E_j$ next to its sub-claim $s_j$.

**Ablation.** We conduct a set of targeted ablation studies, with the aim of isolating the impact of the oracle sub-claim veracity labels on downstream model performance and quantifying the influence of evidence alignment. Here, we use the same model setting as Equation 3, but in the ablations the evidence for SRE and SAE isn't paired with any sub-claim veracity labels.

$$\textbf{(i) Abl. SRE:} \quad \mathcal{E} = (s_j, \{E_{Ci}\}_{i=1}^{n})_{j=1}^{m} \tag{7}$$

$$\textbf{(ii) Abl. SAE:} \quad \mathcal{E} = \{(s_j, E_j)\}_{j=1}^{m} \tag{8}$$

Ablation SRE also mirrors real-world use cases where we may have sub-claims available, with the parent claim and corresponding evidence but no sub-claim labels or sub-claim level evidence. Indeed, Ablation SRE is the setup used in the experiments for the MMM-Fact and COVID-Fact datasets. The SAE setting was not possible for these as contrary to the dataset introduced in this paper, they lack human-annotated sub-claim evidence spans, a resource-intensive annotation process.

**Metrics and statistical significance.** Claim-level verification performance (See Table 2) is gauged using two primary ***metrics:*** macro-averaged F1-score and balanced accuracy [Brodersen *et al.*, 2010]. We use both due to the class imbalance between claim veracity labels. Models were run over three independent random seeds (six for ablations).

**Statistical Significance:** We applied paired statistical testing on claims across model configurations to evaluate the statistical significance of our findings. We performed a paired bootstrap analysis (Primary Performance $\Delta$) to assess the differences in F1 and Accuracy metrics. Claims were selected with replacement ($N = 1000$ resamples) to compute the performance difference ($\Delta = \text{Metric}_{\text{SAE/SRE}} - \text{Metric}_{\text{baseline}}$) for each resample. We report the two-sided bootstrap $p$-value ($p_{\text{boot}}$) (how often $\Delta$s crosses zero). To evaluate the significance of changes in correct/incorrect predictions, we utilized McNemar's test (Categorical Agreement) on the $2 \times 2$ table of paired correctness on the same claims. This focuses specifically on instances where (SAE/SRE)) and baseline disagree. We report both the McNemar $p$-value and the Odds Ratio (OR) of the disagreeing pairs ($b_{01}/b_{10}$) as a measure of effect size, indicating how many times more likely the model is to correct a baseline error than to introduce a new one.

## 4.2 Sub-claim Veracity

**Experiment details:** Sub-claim veracity is formulated as a classification task [Guo *et al.*, 2022] where individual sub-claims are classified with the claim-level evidence (claim level evidence is repeated per sub-claim). Each sub-claim is considered to be a discrete instance. At the sub-claim level, each sub-claim $s_j$ is independently classified as:

$$\hat{y}_j = \mathcal{M}(s_j, \{E_{Ci}\}_{i=1}^{n}) \tag{9}$$

We predict $\hat{y}_j \in \{T, F, U\}$, since decomposed sub-claims can remain unresolved even when the overall claim is verifiable. We evaluate using the macro-averaged F1-score. We do not consider sub-claim veracity prediction given sub-claim evidence as this deviates from real-world scenarios where only claim-level evidence is available e.g. to journalists.

Predicted sub-claim veracity labels are then aggregated to infer claim-level veracity under noisy sub-claim labels (setting SRE_Noisy in Table 3):

$$\hat{Y}_C^{(\text{noisy})} = f(C, \{(s_j, \mathcal{E}, \hat{y}_j)\}_{j=1}^{m}) \tag{10}$$

Here, $f$ is the classifier that aggregates the claim and sub-claim information. This setup allows us to assess how noisy sub-claim predictions $\hat{y}_j$ influence overall claim verification. We use an LLM-based aggregator, and leave a systematic comparison against deterministic or learned aggregation rules (e.g., conjunctive/disjunctive logic over $\hat{y}_j$, or meta-classifiers) as future work.

## 4.3 Models for Sub-claim Veracity Prediction

**Graph Neural Network Baseline.** We chose GNNs as a strong structured baseline. Our Graph Neural Network (GNN) is a variant of [Bilal *et al.*, 2024], able to leverage evidence from both news articles and Twitter threads, tested on PHEME-plus. It is structured as a bipartite graph to perform sub-claim veracity prediction with two distinct node sets: tweet nodes (preserving thread structure) ($\mathbf{x}_t \in \mathbb{R}^{n \times 512}$) and evidence nodes ($\mathbf{x}_e \in \mathbb{R}^{m \times 128}$), interconnected via dedicated edges. Tweet and evidence texts are encoded into initial node embeddings using pretrained encoders. The training was done with leave-one-event-out [Kochkina and Liakata, 2020] (learning rate=$5e^{-5}$, weight decay= $1e^{-4}$, batch size=20).

**Encoder Models.** We use CHEF [Hu *et al.*, 2022] a BERT-based sub-claim verifier that performs evidence selection before veracity classification. We report the two strongest CHEF variants: *CHEF*$_{latent}$ (hard selection via Hard Kumaraswamy [Bastings *et al.*, 2019]) and *CHEF*$_{rl}$(Bernoulli selection optimized with RL [Lei *et al.*, 2016]). We additionally use BERT[Devlin *et al.*, 2019] as a strong encoder baseline: larger encoders tended to overfit (low-resource setting) in our preliminary trials, so we use BERT as a competitive and standard control. We chunk long evidence to fit the model context (learning rate=$5e^{-5}$, weight decay=$1e^{-4}$, batch size=8).

| Setup | Dataset | F1 $\pm$ std | $\Delta$ | $p_{\text{boot}}$ | OR / McNemar($p$) | Balanced Acc. $\pm$ std | $\Delta$ | $p_{\text{boot}}$ | OR / McNemar($p$) |
|---|---|---|---|---|---|---|---|---|---|
| Vanilla | PHEMEplus | $0.5643 \pm 0.0091$ | – | – | – | $0.6072 \pm 0.0074$ | – | – | – |
| Oracle_SRE | | $0.5872 \pm 0.0127$ | +0.0229 | 0.124 | 1.52 / 0.0227 | $0.6117 \pm 0.0136$ | +0.0045 | 0.774 | 1.52 / 0.0227 |
| Oracle_SAE | | **0.6268 $\pm$ 0.0098** | **+0.0625** | **0.0000** | **2.13 / 0.0000** | **0.6558 $\pm$ 0.0135** | **+0.0486** | **0.0000** | **2.13 / 0.0000** |
| **Vanilla** | CovidFact | **0.7365 $\pm$ 0.0248** | – | – | – | **0.7559 $\pm$ 0.0240** | – | – | – |
| Ablation_SRE | | $0.7252 \pm 0.0161$ | -0.0113 | 0.549 | 0.96 / 0.5599 | $0.7533 \pm 0.0148$ | -0.0326 | 0.115 | 0.86 / 0.1153 |
| **Vanilla** | MMM-Fact | **0.7550 $\pm$ 0.0069** | – | – | – | **0.7451 $\pm$ 0.0065** | – | – | – |
| Ablation_SRE | | $0.6878 \pm 0.0064$ | -0.0672 | 0.0000 | 0.05 / 0.0000 | $0.7120 \pm 0.0127$ | -0.0377 | 0.0000 | 0.05 / 0.0000 |

Table 2: Claim-level veracity given oracle sub-claim labels (PHEMEPlus) and no sub-claim labels (CovidFact, MMM-Fact)§(3.4), to assess the effect of evidence alignment. Macro F1 and Balanced Accuracy with **paired** statistical analysis relative to the base model. $\Delta$ shows absolute performance change vs. Vanilla. Paired bootstrap $p$-values reports the two-sided probability of $\Delta$ crossing zero. 'OR / McNemar($p$)' report the paired odds ratio (OR) and McNemar($p$) for paired correctness differences. Best results in bold.

**Large Language Model.** For both claim and sub-claim veracity prediction our LLM of choice is Qwen3-14B reasoning model. We evaluate Qwen3-14B in zero-shot mode with a fixed prompt (temperature=0.3). We initially used Llama 3.1 8B-instruct and Phi4 Reasoning-14B but found both failed at structured input based prediction.

# 5 Results

## 5.1 Claim Veracity Results

As Table 2 shows, the benefits of decomposition depend heavily on how the evidence is aligned. On PHEMEPlus, decomposition offers a clear advantage when sub-claim evidence is available and aligned with sub-claims (SAE). We see statistically significant gains over the Vanilla baseline (+ 0.0625 F1 and +0.0486 balanced accuracy). However, simply repeating claim-level evidence (SRE) on the same dataset, without associated sub-claim evidence to match the sub-claims and their labels, produces negligible changes, suggesting that usefulness of decomposition is conditional upon evidence alignment. The limitations of SRE become even clearer on MMM-Fact and COVID-Fact where it fails to improve verification on either dataset. Vanilla slightly edges out on COVID-Fact, while on MMM-Fact, the drop from providing noisy sub-claims (without labels or evidence) is substantial (macro-F1 $\Delta : -0.0672$, balanced accuracy $\Delta : -0.0521$). These findings validate our core hypothesis that decomposition works when evidence is tightly coupled with valid sub-claim signals. Without such alignment, the process tends to introduce noise that harms the final inference, a phenomenon we analyze further in (§5.4). These findings highlight the need for proper evidence synthesis.

## 5.2 Ablation Results

Table 3 highlights a trade-off between evidence granularity and the quality of sub-claim labels. When using redundant claim-level evidence but no sub-claim veracity labels (Ablation SRE), the model relies little on sub-claim veracity. However, in the aligned setting where sub-claim evidence is provided (Ablation SAE), the absence of sub-claim veracity labels is detrimental. The model performs significantly worse than standard SAE and even falls below the SRE baseline. This indicates that granular, sub-claim-level, evidence is insufficient on its own. Without the signal from correct sub-claim veracity labels, the model struggles to synthesize the fine-grained information, leading to inference errors that coarser evidence setups seem to avoid.

## 5.3 Sub-claim Veracity Results

Table 4 presents the sub-claim veracity results. The zero-shot Qwen3-14B sets the state of the art with $56.94\%$ macro-F1. The performance gap between the GNN ($45.79\%$) and the encoder-based CHEF variants ($\approx 28\%$) underscores the necessity of structured modeling for this task. While encoders struggle with overfitting and evidence integration, the GNN successfully captures evidence interactions, though it still falls short of the LLM's reasoning capabilities. However, the low-resource nature of the dataset makes it much harder for trained encoders to perform well compared to much larger LLMs.

## 5.4 Effects of Noisy Sub-claim Predictions

Table 3 shows that replacing oracle sub-claim labels with predicted ones reduces performance. However, the severity of this drop depends on the evidence configuration. Under the aligned SAE setting, the degradation is moderate (Qwen $\text{SAE}_{\text{Noisy}}$: $\Delta - 0.0304$ F1). However, under SRE, the models collapse (dropping to $\approx 0.44$ F1), confirming that redundant claim-level evidence exacerbates error propagation from noisy sub-claim labels. This error propagation pattern aligns with findings from [Lyu *et al.*, 2025] in multi-hop verification, confirming that intermediate step errors compound through the pipeline. A deeper look at the per-model error analysis reveals distinct model biases. From Table 5 we can see that, during sub-claim prediction, when GNN commits (non-$U$), its accuracy on T/F ($\text{Acc}_v$) is comparable to Qwen (0.625 vs 0.646), but it abstains far more (35% vs 15.8%) and almost never detects refutations ($R_F$ 0.039 vs. 0.314). This gap is driven primarily by coverage on verifiable sub-claims (T/F). Qwen issues a polarity label for 85.7% of verified cases ($\text{Cov}_{ver}$=0.857), while GNN commits on only 66.7% (0.667). Moreover, Qwen's increased refutation sensitivity comes with modest precision ($P_F$ 0.276 vs. 0.154), indicating more frequent false positives for F alongside higher $R_F$. Thus, decomposition success depends not only on correctness when committing, but on the model's commit/abstain and refute/affirm bias. This distinction explains the downstream effects where in the presence of fine-grained evidence (SAE), the GNN's tendency to predict *Unverified* acts as a safety mechanism, but also misses crucial refutations. Ultimately, this analysis shows that sub-claim macro-F1 is a poor predictor of success and

| Setup | F1 Macro ± std | Δ F1 | Balanced Accuracy ± std | Δ Acc |
|---|---|---|---|---|
| Oracle_SRE | $0.5872 \pm 0.0127$ | – | $0.6117 \pm 0.0136$ | – |
| Ablation_SRE | $0.5808 \pm 0.0064$ | $-0.0064$ | $0.6259 \pm 0.0052$ | $+0.0142$ |
| Oracle_SAE | $0.6268 \pm 0.0098$ | – | $0.6558 \pm 0.0132$ | – |
| Ablation_SAE | $0.5485 \pm 0.0052$ | $-0.0783$ | $0.6220 \pm 0.0065$ | $-0.0338$ |
| **Effect of Noisy sub-claim Predictions on Claim Verification (all Δ vs. SAE_oracle)** | | | | |
| **Oracle_SAE** | $\mathbf{0.6268 \pm 0.0098}$ | – | $\mathbf{0.6558 \pm 0.0132}$ | – |
| Qwen Noisy_SAE | $0.5964 \pm 0.0360$ | $-0.0304$ | $0.6425 \pm 0.0595$ | $-0.0133$ |
| Qwen Noisy_SRE | $0.4335 \pm 0.0439$ | $-0.1933$ | $0.4411 \pm 0.0512$ | $-0.2147$ |
| GNN Noisy_SAE | $0.5839 \pm 0.1202$ | $-0.0429$ | $0.5963 \pm 0.1361$ | $-0.0595$ |
| GNN Noisy_SRE | $0.4416 \pm 0.0489$ | $-0.1852$ | $0.4399 \pm 0.0609$ | $-0.2159$ |

Table 3: Claim-level veracity performance comparison under ablations and noisy sub-claim prediction inputs for PhemePLUS (§3.4). Ablation Δs are computed relative to the corresponding full model with sub-claim labels. That is, Ablation_SRE (no sub-claim labels, claim-level evidence) vs. Oracle_SRE (oracle sub-claim labels, claim-level evidence) and Ablation_SAE (no sub-claim labels, sub-claim level evidence) vs. Oracle_SAE (oracle sub-claim labels, sub-claim-level evidence). In the noisy setting, Δs are computed relative to **Oracle_SAE**, where noisy variants use predicted sub-claim labels and oracle uses the gold sub-claim labels. Noisy labels were obtained either via GNN or Qwen3-14B.

| Model | F1 Macro (%) |
|---|---|
| $\text{CHEF}_{latent}$ | $27.30 \pm 7.11$ |
| $\text{CHEF}_{rl}$ | $28.92 \pm 6.56$ |
| Bert-Chunk | $41.90 \pm 0.57$ |
| GNN | $45.79 \pm 1.25$ |
| Qwen3-14B | $56.94 \pm 0.92$ |

Table 4: Macro F1 scores (%) with standard deviation across 3 seeds for sub-claim-level veracity classification on PhemePLUS (§3.4).

| Model | U | F | $R_F$ | $P_F$ | $C_v$ | $A_{v\|\neg u}$ | $A_v$ |
|---|---|---|---|---|---|---|---|
| Qwen | 15.8 | 24.2 | .31 | .28 | .86 | .55 | .65 |
| GNN | 35.0 | 5.4 | .04 | .15 | .67 | .42 | .63 |

Table 5: Sub-claim-labeling profile (F/U) (§3.4). U and F are percentages. $R_F$ and $P_F$ are refutation recall and precision. $C_v$ is coverage/non-abstain rate on verified sub-claims (GT$\in \{T, F\}$). $A_{v|\neg u}$ counts abstentions as errors; $A_v$ is verified-subclaim accuracy. T prediction is similar: 60.0% for Qwen and 59.6% for GNN.

that error analysis, especially the balance between conservative abstention and aggressive detection determines whether decomposition can improve verification.

## 6 Discussion and Future Work

**The Alignment Bottleneck.** Our experiments clarify a boundary condition for decomposition-based verification: claim decomposition is not a universal performance booster, but rather depends on evidence granularity. Decomposition improves performance only when sub-claims are paired with sub-claim-aligned evidence (SAE setting). When evidence is repeated at the claim level (SRE), decomposition adds structure without resolving ambiguity and can degrade performance as seen on MMM-Fact and COVID-Fact, where the SRE setting failed to improve over Vanilla. This suggests that the primary real-world bottleneck is often *evidence synthesis*. Without reliable sub-claim–evidence alignment, decomposition offers limited value and may rather amplify noise.

**Sensitivity to Error Profiles.** Moving from oracle to predicted sub-claim labels shows upstream noise affects decomposition, but the *type* of labeling error matters more than overall accuracy. Under noise, SRE collapses, while SAE can remain comparatively stable depending on the predictor's label bias. We observe that the GNN behaves conservatively, frequently predicting *Unverified*; in a granular-evidence setting this can be beneficial, by avoiding incorrect polar signals. By contrast, Qwen is biased towards refutations and risks more false positives. Thus,stabilizing decomposition pipelines may require optimizing the sub-claim predictor's *label bias* rather than maximizing macro-F1 alone.

**Future Directions: Human-Centric and Real-Time Verification.** Beyond algorithmic accuracy, the ultimate utility of decomposition lies in its interpretability. A key avenue for future work is evaluating whether sub-claim-level rationales genuinely assist human fact-checkers (e.g., journalists) in identifying missing evidence or isolating disputed claims, rather than simply increasing cognitive load. Finally, our current setup uses temporally bounded but static evidence snapshots. In breaking news scenarios, such as those on social media, information evolves rapidly. Future systems must move towards *just-in-time* verification, utilizing timeline summarization to capture evolving narratives and explicitly representing the status of sub-claims as "unresolved" until sufficient evidence becomes available.

## 7 Conclusion

We investigate the role of structured sub-claim decomposition in complex claim verification. By leveraging a real-world dataset with human-annotated sub-claim evidence spans, we show that decomposition is *not* a universal remedy: it requires granular, sub-claim-aligned evidence and reliable veracity labels to be effective. However, **our results reveal that when these ideal conditions are met, sub-claim decomposition leads to significant performance gains over vanilla methods that treat claims monolithically.**

By contrast, when evidence is limited to the claim level

or when sub-claim labels are noisy, the benefits of decomposition diminish or even become detrimental to performance, as shown with our findings on MMM-Fact and Covid-Fact. Our ablation experiments further confirm a critical interaction between evidence granularity and label reliability: providing granular evidence without reliable labels can confuse the verification model. Furthermore, downstream robustness depends heavily on the *label bias* of the verification model (e.g., the safety of abstaining via an *Unverified* label versus the risk of committing to incorrect polarity). Ultimately, we conclude that the primary bottleneck for deployment in real-world settings is not decomposition itself, but the capability for precise evidence synthesis and alignment.

Future research should prioritize developing robust sub-claim predictors that optimize for label bias, specifically favoring calibrated abstention over incorrect polarity. Additionally, critical advancements are needed in evidence synthesis to produce sub-claim-aligned spans, particularly under temporal constraints. Ultimately, the goal is to design verifiers that can leverage structured decomposition while remaining resilient to noisy labeling inherent in real-world settings.

## Acknowledgements

This work was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant ref EP/V030302/1) and the Alan Turing Institute (grant ref EP/N510129/1). This work was also supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK (KP0016) as a Keystone project lead by Maria Liakata. F. Ruggeri is partially supported by the project European Commission's NextGeneration EU programme, PNRR – M4C2 – Investimento 1.3, Partenariato Esteso, PE00000013 - "FAIR - Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI" and by the European Union's Justice Programme under Grant Agreement No. 101087342 for the project "Principles Of Law In National and European VAT". We also thank Queen Mary University of London Comp-Research Team for their infrastructure support.

## References

[Atanasova *et al.*, 2020] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating fact checking explanations. In *Proc. of the 58th Annual Meeting of the ACL*, pages 7352–7364, Online, July 2020. ACL.

[Bastings *et al.*, 2019] Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *Proc. of the 57th Annual Meeting of the ACL*, pages 2963–2977, Florence, Italy, July 2019. ACL.

[Bilal *et al.*, 2024] Iman Munire Bilal, Preslav Nakov, Rob Procter, and Maria Liakata. Generating zero-shot abstractive explanations for rumour verification, 2024.

[Brodersen *et al.*, 2010] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th Int. Conf. on Pattern Recognition*, pages 3121–3124, 2010.

[Chen *et al.*, 2022] Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. Generating literal and implied subquestions to fact-check complex claims. In *Proc. of the 2022 Conf. on EMNLP*, pages 3495–3516, Abu Dhabi, United Arab Emirates, December 2022. ACL.

[Chen *et al.*, 2024] Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild. In *Proc. of the 2024 Conf. of the NAACL: Human Language Technologies*, pages 3569–3587, Mexico City, Mexico, June 2024. ACL.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conf. of the NAACL: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota, June 2019. ACL.

[Dougrez-Lewis *et al.*, 2022] John Dougrez-Lewis, Elena Kochkina, Miguel Arana-Catania, Tom Sherborne, Yulan Lei, and Maria Liakata. PHEMEplus: Enriching social media rumour verification with external evidence. In *Proc. of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 1–13, Dublin, Ireland, May 2022. ACL.

[Dougrez-Lewis *et al.*, 2025] John Dougrez-Lewis, Mahmud Elahi Akhter, Federico Ruggeri, Sebastian Löbbers, Yulan He, and Maria Liakata. Assessing the reasoning capabilities of llms in the context of evidence-based claim verification, 2025.

[Ge *et al.*, 2025] Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented llms. *ArXiv*, abs/2505.17762, 2025.

[Glockner *et al.*, 2022] Max Glockner, Yufang Hou, and Iryna Gurevych. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proc. of the 2022 Conf. on EMNLP*, pages 5916–5936, Abu Dhabi, United Arab Emirates, December 2022. ACL.

[Gong *et al.*, 2024] Haisong Gong, Huanhuan Ma, Qiang Liu, Shu Wu, and Liang Wang. Navigating the noisy crowd: Finding key information for claim verification. *CoRR*, abs/2407.12425, 2024.

[Guo *et al.*, 2022] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the ACL*, 10:178–206, 2022.

[Hu *et al.*, 2022] Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proc. of the 2022 Conf. of the NAACL: Human Language Technologies*, pages 3362–3376, Seattle, United States, July 2022. ACL.

[Hu *et al.*, 2024] Qisheng Hu, Quanyu Long, and Wenya Wang. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance?, 2024.

[Huang *et al.*, 2024] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying

Song, and Denny Zhou. Large language models cannot self-correct reasoning yet, 2024.

[Jolly *et al.*, 2022] Shailza Jolly, Pepa Atanasova, and Isabelle Augenstein. Generating fluent fact checking explanations with unsupervised post-editing. *Information*, 13(10):500, 2022.

[Kamoi *et al.*, 2023] Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in Wikipedia. In *Proc. of the 2023 Conf. on EMNLP*, pages 7561–7583, Singapore, December 2023. ACL.

[Kochkina and Liakata, 2020] Elena Kochkina and Maria Liakata. Estimating predictive uncertainty for rumour verification models. In *Proc. of the 58th Annual Meeting of the ACL*, pages 6964–6981, Online, July 2020. ACL.

[Lei *et al.*, 2016] Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proc. of the 2016 Conf. on EMNLP*, pages 107–117, Austin, Texas, November 2016. ACL.

[Li *et al.*, 2024] Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the ACL: NAACL 2024*, pages 163–181, Mexico City, Mexico, June 2024. ACL.

[Liu *et al.*, 2024] Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *Findings of the ACL: ACL 2024*, pages 15556–15583, Bangkok, Thailand, August 2024. ACL.

[Liu *et al.*, 2025] Xinyu Liu, Junhao Yu, Jing Zhao, and Zhouhan Li. Bidev: Bidirectional decomposition for complex claim verification, 2025.

[Lyu *et al.*, 2025] Xiucheng Lyu, Chengyu Cao, Mingwei Sun, and Ruifeng Xu. Mitigating error propagation in multi-hop fact verification with logic reasoning. *International Journal of Machine Learning and Cybernetics*, 2025.

[Ma *et al.*, 2025] Jiatong Ma, Linmei Hu, Rang Li, and Wenbo Fu. Local: Logical and causal fact-checking with llm-based multi-agents. In *Proc. of the ACM on Web Conference 2025*, page 1614–1625, New York, NY, USA, 2025. ACM.

[Min *et al.*, 2023] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FactScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proc. of the 2023 Conf. on EMNLP*, pages 12076–12100, Singapore, December 2023. ACL.

[Nakov *et al.*, 2021] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. Automated fact-checking for assisting human fact-checkers. In *Proc. of the 30th IJCAI*, pages 4551–4558. IJCAI, 2021.

[Pan *et al.*, 2023] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. In *Proc. of the 61st Annual Meeting of the ACL*, pages 6981–7004, Toronto, Canada, July 2023. ACL.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, ACL '02, pages 311–318, USA, 2002. Association for Computational Linguistics.

[Quelle and Bovet, 2023] Dorian Quelle and Alexandre Bovet. The perils and promises of fact-checking with large language models. In *Proc. of the 2023 ACM Conf. on Fairness, Accountability, and Transparency*, 2023.

[Saakyan *et al.*, 2021] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proc. of the 59th ACL and the 11th IJCNLP*, pages 2116–2129, Online, August 2021. Association for Computational Linguistics.

[Schlichtkrull *et al.*, 2023a] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. In *Advances in NeurIPS 36*, 2023.

[Schlichtkrull *et al.*, 2023b] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. *ArXiv*, abs/2305.13117, 2023.

[Sundriyal *et al.*, 2022] Megha Sundriyal, Parantak Rani, Sriparna Saha, Pushpak Bhattacharyya, and Tanmoy Chakraborty. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In *Proc. of the 2022 Conf. on EMNLP*, pages 7701–7715, Abu Dhabi, United Arab Emirates, December 2022. ACL.

[Tang *et al.*, 2024] Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proc. of the 2024 Conf. on EMNLP*, pages 8818–8847, Miami, Florida, USA, November 2024. ACL.

[Wang and Shu, 2023] Haoran Wang and Kai Shu. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the ACL: EMNLP 2023*, pages 6288–6304, Singapore, December 2023. ACL.

[Wang, 2024] Yuxia et al. Wang. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the ACL: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA, November 2024. ACL.

[Wanner *et al.*, 2024] Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. A closer look at claim decomposition. In *Proc. of the 13th Joint Conf. on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico, June 2024. ACL.

[Xu *et al.*, 2025] Wenyan Xu, Dawei Xiang, Tianqi Ding, and Weihai Lu. Mmm-fact: A multimodal, multi-domain fact-checking dataset with multi-level retrieval difficulty, 2025.

[Yang *et al.*, 2024] Sohee Yang, Jonghyeon Kim, Joel Jang, Seonghyeon Ye, Hyunji Lee, and Minjoon Seo. Improving probability-based prompt selection through unified evaluation and analysis. *Transactions of the ACL*, 12:664–680, 2024.

[Yang *et al.*, 2025] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. Accessed: 2026-01-15.

[Zhang and Gao, 2023] Xuan Zhang and Wei Gao. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proc. of the 13th IJCNLP and the 3rd AACL*, pages 996–1011, Bali, Indonesia, November 2023. ACL.

[Zhang *et al.*, 2019] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675, 2019.

[Zhang *et al.*, 2024] Zhihao Zhang, Yixing Fan, Ruqing Zhang, and Jiafeng Guo. A claim decomposition benchmark for long-form answer verification, 2024.

[Zhao *et al.*, 2024] Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. PACAR: Automated fact-checking with planning and customized action reasoning using large language models. In *Proc. of the 2024 LREC-COLING*, pages 12564–12573, Torino, Italia, May 2024. ELRA and ICCL.

[Zubiaga *et al.*, 2016] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29, 2016.

[Zubiaga *et al.*, 2018] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2):1–36, 2018.