From Articles to Premises: Building PrimeFacts, an Extraction Methodology and Resource for Fact-Checking Evidence

Premtim Sahitaj[1,2], **Jawan Kolanowski**[3], **Ariana Sahitaj**[1,2]
Veronika Solopova[1,2], **Max Upravitelev**[1,2], **Daniel Röder**[2], **Iffat Maab**[4]
Junichi Yamagishi[4], **Sebastian Möller**[1,2], **Vera Schmitt**[1,2]
[1]Technische Universität Berlin, Quality and Usability Lab, Berlin, Germany
[2]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany
[3]Harz University of Applied Sciences, Wernigerode, Germany
[4]National Institute of Informatics, Tokyo, Japan
{sahitaj, sebastian.moeller, vera.schmitt}@tu-berlin.de
{ariana.sahitaj, veronika.solopova}@tu-berlin.de, daniel.roeder@dfki.de
{maab, jyamagis}@nii.ac.jp, u37871@hs-harz.de

Abstract

Fact-checking articles encode rich supporting evidence and reasoning, yet this evidence remains largely inaccessible to automated verification systems due to unstructured presentation. We introduce `PrimeFacts`, a methodology and resource for extracting fine-grained evidence from full fact-checking articles. We compile 13,106 PolitiFact articles with claims, verdicts, and all referenced sources, and we identify 49,718 in-article hyperlinks as natural anchors to pinpoint key evidence. Our framework leverages large language models (LLMs) to rewrite these anchor sentences into stand-alone, context-independent premises and investigates the extraction of additional implicit evidence. In evaluations on cross-article evidence retrieval and claim verification, the extracted premises substantially improve performance. Decontextualized evidence yields higher retrievability, achieving up to a 30% relative gain in Mean Reciprocal Rank over verbatim sentences, and using the evidence for verdict prediction raises Macro-F_1 by 10-20 points over the baseline. These gains are consistent across different verdict granularities (2-class vs. 5-class) and model architectures. A qualitative analysis indicates that the decontextualized premises remain faithful to the original sources. Our work highlights the promise of reusing fact-checkers' evidence for automation and provides a large-scale resource of structured evidence from real-world fact-checks.

Keywords: Corpus Construction, Information Extraction, Large Language Models, Resource Evaluation

1. Introduction

Professional *fact-checking* has become a crucial process to counter misinformation and disinformation in politics and media. Journalistic fact-checkers investigate a check-worthy claim by gathering supporting and refuting evidence and then issuing a verdict on the claim's veracity (Graves, 2016; Jiang et al., 2020a). The results of this labor-intensive process are published as articles containing the claim, contextual background, evidence discussion, and a final verdict on truthfulness (Sahitaj et al., 2025). Standardization efforts like the ClaimReview schema have encouraged fact-checkers to explicitly mark the claim and verdict in each article, but other critical components, namely the *evidence* that led to the verdict and the *reasoning* connecting evidence to conclusion, are rarely structured or annotated due to the extra workload this would entail (Jiang et al., 2020a). As a result, rich supporting information in fact-check articles remains locked in unstructured text, limiting its reusability for automated verification systems and obscuring the transparency of how conclusions are reached for a given claim (Jiang et al., 2020a; Alhindi et al., 2018). Similar or rephrased claims tend to reappear across media, causing fact-checkers to spend effort redundantly on already verified information (Nakov et al., 2021; Panchendrarajan and Zubiaga, 2024). We argue that *evidence reuse* is a valuable extension of claim matching: once supporting

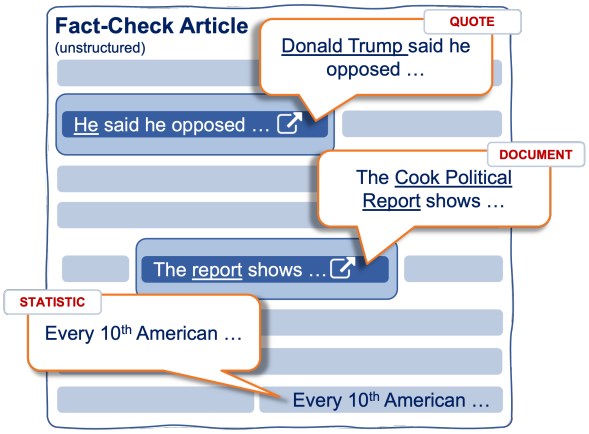

Figure 1: Example fragments from a fact-checking article. (Top and middle) hyperlink-anchored sentence *decontextualized* in *Mode B*. (Bottom) statistical claim identified by *open extraction* in *Mode C*.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7719–7731
11–16 May 2026.

material is extracted and normalized, the same evidence can be associated with multiple variants of a claim, enabling future claims to be verified using pre-existing evidence. Moreover, this would enable analyses of aspects such as source diversity and reasoning consistency. For instance, one could examine how premises are assembled to support or refute a target claim, or assess whether the evidence usage reveals informational bias in an article's reasoning (Wang et al., 2025a; Stab and Gurevych, 2017; Maab et al., 2024). However, manually annotating evidence at scale is impractical. Fact-check articles are often long, densely sourced, and rhetorically complex (Humprecht, 2020; Jiang et al., 2020b). Exhaustively identifying all relevant evidence sentences and rewriting them as standalone statements would require significant expert effort, leading to annotator fatigue, inconsistency, and prohibitive cost (Ostrowski et al., 2021; Schlichtkrull et al., 2023). This motivates exploring automated methods to unlock the evidence within fact-check articles.

These articles contain *in-text citations* that serve as natural anchors for evidence to enhance the transparency of the manual verification process (see Figure 1 for an illustrative example). Specifically, hyperlinks to primary sources such as data, reports and transcripts are extensively embedded within their writing (Cazzamatta, 2025a; Humprecht, 2020). Therefore, these reference links can be leveraged to automatically pinpoint the key supporting sentences within an article. By extracting and appropriately reformulating those anchor sentences, we aim to make the evidence both *addressable* (tied to a specific source reference) and *portable* (understandable outside the original context) for use in downstream automated retrieval and verification systems. In this work, we investigate the feasibility of automatically extracting evidence from fact-check articles by exploiting their cited sources and language models. We focus on three research questions:

- **RQ1 (Addressability):** To what extent can in-article hyperlinks serve as a reliable proxy for identifying the core evidence that supports or refutes a claim?
- **RQ2 (Portability):** Does rewriting premises into stand-alone, *decontextualized* statements improve their usefulness for retrieval and automated verification tasks?
- **RQ3 (Robustness):** Are the findings consistent across different evaluation settings and task granularities?

Our contributions are four-fold: (i) We curate `PrimeFacts`, a resource of 13,106 PolitiFact fact-checks with structured article metadata and 49,718 in-article hyperlinks that serve as evidence anchors. (ii) We introduce a three-mode extraction framework: anchored verbatim sentences, LLM-based decontextualization into stand-alone premises, and open extraction of self-contained premises with source attributions. We additionally define a lightweight evidence-type ontology to characterize premise content. (iii) We propose a score for evaluating the faithfulness of decontextualizations, combining forward textual entailment with an asymmetric lexical-overlap penalty, and conduct a targeted human study to assess self-containment and evidence typing. (iv) We evaluate *evidence reuse* on two downstream tasks, cross-article retrieval and zero-shot claim verification, across six instruction-tuned LLMs and two verdict granularities.

2. Related Work

Automated fact-checking (AFC) research aims to verify claims by retrieving evidence and predicting verdicts, often producing a textual explanation (Guo et al., 2022). A persistent challenge in AFC is the shortage of training data that contains real-world claims paired with gold-standard evidence and detailed reasoning. Many existing datasets use artificially constructed claims or evidence from Wikipedia, e.g. FEVER (Thorne et al., 2018), and extensions (Schuster et al., 2021; Aly et al., 2021; Ma et al., 2024), or consider the referenced pages within fact-checking articles as evidence (Augenstein et al., 2019; Khan et al., 2022). Other datasets target fact-checking justifications by relying on heuristic extractions (Alhindi et al., 2018; Zeng and Gao, 2024) or are limited in scale due to manual annotation (Ostrowski et al., 2021; Wang et al., 2025b). These approaches do not isolate or quantify the evidence-bearing sentences that journalists integrate into the article's reasoning. Instead, they ingest entire referenced pages or treat fact-checks as the evidence. Reliance on whole fact-checking articles introduces noise and reduces system effectiveness (Samarinas et al., 2021; Xing et al., 2024; Deng et al., 2025; Sauchuk et al., 2022), while incomplete or inaccessible sources and link rot further undermine automated evidence retrieval (Cazzamatta, 2025a; Warren et al., 2025; Kavtaradze, 2024; Zhou et al., 2015; Klein et al., 2014). `PrimeFacts` addresses this gap by focusing on evidence units within fact-checking articles that journalists already anchor through in-text hyperlinks and end-of-article references. We promote both precise addressability via these anchored spans and enhanced portability by reformulating them into context-independent premises. Our open extraction approach shares the core insight of Chen et al. (2024) that decomposing text into atomic,

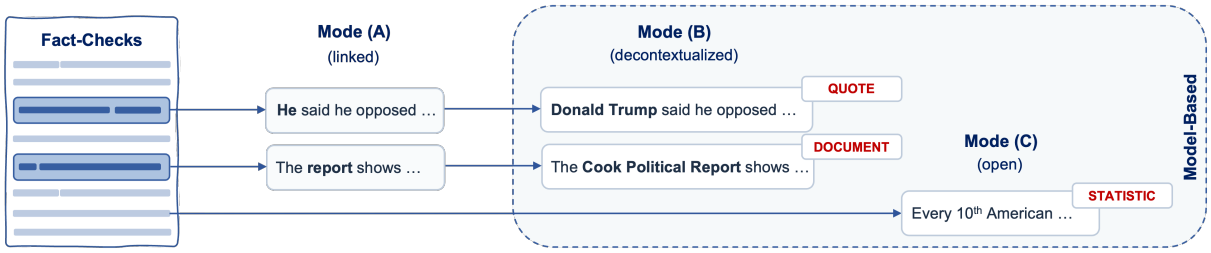

Figure 2: Extraction pipeline for transforming fact-checks into refined, decontextualized evidence.

self-contained propositions improves retrieval, but differs in scope and constraints. Chen et al. (2024) define propositions as minimal, self-contained factoids and exhaustively decompose passages so that all propositions together recover the full passage semantics, using a compact model distilled from GPT-4 outputs. In contrast, our approach performs selective, attribution-based extraction from fact-checking articles, bounded by the article's anchor count and targeting key evidence rather than every atomic fact. Recent work on maintaining editable knowledge bases (Li et al., 2025) further motivates the extraction of modular, updateable evidence units, a goal that our decontextualized premises directly support.

3. Data Collection

Each record in `PrimeFacts` corresponds to a single fact-checking article from the English-language PolitiFact[1] archive, collected from the chronological and topic-based indexes up to September 2025. We retain only fact-check articles targeting text-based claims, identified via URL and markup patterns, to preserve internal validity and consistent textual structure. The final dataset comprises 13,106 fact-check articles authored by 661 individual journalists, each containing at least one in-text hyperlink to an external source. For each article, we store a structured representation linking the canonical PolitiFact URL with its metadata, including crawl timestamp, editorial tags, claim information, author and speaker metadata, and a structured list of cited sources to ensure provenance and reproducibility. The released version of the resource[2] contains only derived metadata and annotations. Original fact-check article texts are not redistributed for copyright reasons. The resource is organized into three main components: (i) *Article metadata*, containing the article origin, verdict label, claim statement, and cited materials; (ii) *Entity metadata*, providing unified lookup profiles for authors and speakers to maintain consistent attribution across articles, and (iii) *Annotations*, linking each claim to its extracted anchor statements and aligning them with the corresponding article structure and metadata. All components are cross-referenced through their canonical URLs. A companion statistics file reports aggregate distributions to support stratified sampling and downstream evaluation. Beyond aggregating raw PolitiFact metadata, `PrimeFacts` contributes several processing steps that constitute a standalone resource: layout-aware text normalization derived from web browser rendering (Weichselbraun, 2021), sentence-level segmentation with stable letter identifiers, hyperlink extraction and cross-referencing with author-provided source lists, systematic label-leak filtering to prevent verdict contamination, and the decontextualized premise annotations as structured evidence. Together, these transformations turn unstructured fact-check articles into a queryable evidence base that supports retrieval-augmented verification and comparative analysis.

4. Methodology

Before describing the framework, we clarify the key terms used throughout this paper. An *anchor* is an in-article hyperlink that marks a sentence citing an external source. A *source* is the external document or page linked by an anchor (e.g., a government report or dataset). *Evidence* refers broadly to any information supporting or refuting a claim. A *premise* is a decontextualized, self-contained evidence statement derived from the article, suitable for reuse outside its original context. We use these terms consistently hereafter. We propose a framework for extracting fine-grained evidence from fact-check articles and for evaluating its usefulness in downstream verification settings. Figure 2 gives an overview. The framework includes three evidence modes: hyperlink-anchored sentences, their decontextualized variants, and premises obtained through open extraction. We evaluate the resulting evidence representations in cross-article retrieval and claim verification, and we further assess their faithfulness to the original article content.

[1] https://www.politifact.com

[2] https://huggingface.co/datasets/xplainlp/prime-facts

4.1. Evidence Extraction

Each article is segmented into distinct sentence-like units u, each assigned a stable identifier ι. For downstream experiments, we remove sentences that explicitly state the verdict to prevent label leakage, ensuring that models see only evidence, not conclusions (Alhindi et al., 2018; Glockner et al., 2022). We then identify anchor sentences that contain at least one hyperlink to an external source. These in-text hyperlinks typically mark factual assertions supported by external evidence (Cazzamatta, 2025a). The rationale for treating hyperlinked sentences as high-quality evidence markers is grounded in journalistic practice: fact-checkers deliberately embed links to primary sources such as government data, official records, prior reporting, and expert statements to substantiate their analysis and enhance transparency (Humprecht, 2020). This editorial convention makes hyperlinks natural proxies for evidence-bearing content. We cross-reference each hyperlink with the article's author-provided source list and retain only those pointing to sources that the author has explicitly listed by name in the article's reference section, to ensure we capture only high-quality anchors. We only consider articles with at least one anchor. This leaves us with 13,106 articles and a final set of 49,718 anchors. Using the identified anchors, we derive two sets of evidence, corresponding to Mode A and its decontextualized variation Mode B. Mode C extracts decontextualized premises without anchors. Each mode produces a collection of candidate premises per article with a unique identifier ι linking it back to the article text for provenance.

4.1.1. Mode A: Anchored Evidence

Mode A uses each anchor sentence verbatim as an evidence unit. This yields a set of factual statements directly grounded in the fact-checker's cited sources. For example, if an article states, "In 2020, the city's homicide rate was the lowest on record" with a hyperlink to a police report, we retain that sentence unchanged as evidence. By design, these anchor-based evidence units are high-precision, but this does not guarantee high recall. However, using them as-is can limit reusability, since anchor sentences may contain pronouns or context-dependent references that are not self-explanatory outside the article context. Mode A primarily addresses *addressability* (RQ1) and serves as a baseline for our extraction framework.

4.1.2. Mode B: Decontextualization

Mode B aims to improve the *portability* of evidence by rewriting each Mode A sentence into a standalone premise. We employ an LLM to *decontextualize* each anchor sentence, i.e. to make implicit references explicit so that the sentence is understandable out of context (Choi et al., 2021). For example, a Mode A statement *"He took office in 2019"* might be rewritten as *"Volodymyr Zelenskyy took office as President of Ukraine in May 2019"*, resolving the pronoun and adding context for clarity. Concretely, for each anchor sentence the LLM receives the full letter-segmented article together with the claim statement and the target letter identifier pointing to the anchor sentence. This provides the model with sufficient surrounding context to resolve coreferences and implicit references. We use a zero-shot prompting strategy with structured JSON output guided by a Pydantic schema (see Appendix A), which constrains the model to return a single decontextualized sentence along with an evidence-type category in one joint generation step. The prompt explicitly instructs the model to preserve the original meaning and factual content while only integrating details necessary for standalone interpretation. This follows best practices from recent work on minimality in decontextualization (Gunjal and Durrett, 2024). Each decontextualized premise is output together with its corresponding identifier ι for traceability to the source unit u. Mode B produces one decontextualized premise for each Mode A evidence unit. This allows evidence from one fact-check to be more readily understood and reused in verifying other claims, addressing RQ2. Building on prior work investigating evidence operationalization in the fact-checking newsroom by Cazzamatta (2025c) and consistent with comparative evidence that fact-checkers routinely add background and context to qualify and make verdicts interpretable (Cazzamatta, 2025b), we define a novel ontology of evidence types and assign types to each decontextualized premise: `QUOTE` as attributed statement by a person or organization, `STATISTIC` as numeric fact from an official dataset or series, `DOCUMENT` as findings of an authoritative record (e.g., law, ruling, report, prior fact-check), `CONTEXT` as background attribution or qualification needed to interpret premises, and `OTHER` if none of the above fits. These labels do not affect the extraction framework, but help characterize what kinds of information the premises convey.

4.1.3. Mode C: Open Extraction

Mode C explores a more open-ended extraction using LLM generation. Instead of relying only on explicit hyperlinks, we prompt an LLM to process the entire fact-checking article and directly output a set of self-contained premises. Prompt details are provided in Appendix A. The purpose of this approach is to capture any central premises in the article that the journalist may have implied and not explicitly linked. We constrain this process to maintain align-

ment with the article: the model is instructed to produce at most n premises, where n equals the number of Mode A anchors in that article, to ensure a fair comparison. For each generated premise, the LLM must cite the supporting unit identifier ι from which it drew the information. Comparing Mode C to the anchor-driven modes thus reveals whether important premises were not considered by only looking at hyperlinks. This serves as a stress test for our framework, since any premises found in Mode C should ideally overlap with or complement Mode A/B if the anchor-based approach is comprehensive.

4.2. Evidence Reuse

4.2.1. Faithfulness

Decontextualization rewrites a source sentence into a stand-alone premise that preserves the original factual content while resolving contextual ambiguity (Choi et al., 2021). Prior work in abstractive summarization has shown that Natural Language Inference (NLI) models, which determine if a premise entails, contradicts, or is neutral towards a source, correlate higher with human judgments of factual consistency than standard measures such as `ROUGE` or `BERTScore` (Maynez et al., 2020). We adapt this idea and evaluate the *faithfulness* of a decontextualized premise p to its source sentence s as a NLI problem, using *forward* textual entailment. Forward direction captures the desideratum that a more specific, context-completed rewrite should imply the key facts of the original sentence, while the reverse direction does not need to hold. For example, the premise *"The unemployment rate doubled in 2016, according to the Bureau of Labor Statistics."* correctly entails the more general source statement *"The rate doubled in 2016"*. The reverse is not necessarily true, making this a directional check of information preservation. However, NLI models can be susceptible to lexical-overlap (Naik et al., 2018). A premise that simply copies, minimally edits, or truncates s can receive a high entailment score despite failing the decontextualization objective. To address this, we introduce the **Decontextualization Faithfulness Score (DFS)**, a composite measure that balances factual entailment with an explicit lexical overlap penalty across a dataset $D \subseteq P \times S$, where P and S are the sets of generated premises and source sentences

$$\mathrm{DFS}(D) = \frac{1}{|D|} \sum_{(p,s)\in D} \mathrm{E}(p,s)\,\big(1 - \mathrm{O}(p,s)\big),$$

$$\mathrm{O}(p,s) = \frac{\big|\mathrm{t}(p) \cap \mathrm{t}(s)\big|}{\big|\mathrm{t}(p)\big|},$$

where $\mathrm{E}(p,s) \in [0,1]$ is the probability that p entails s, $\mathrm{O}(p,s) \in [0,1]$ is the asymmetric lexical overlap of p covered by s, and $\mathrm{t}(\cdot)$ denotes the (multi)set of tokens. The token overlap is normalized by the length of the generated premise $|\,\mathrm{t}(p)|$, which correctly penalizes premises that are simple excerpts of the source while rewarding the addition of new, contextualizing tokens that improve portability without undermining factual consistency with its source. We compute DFS for both Mode B and Mode C across model outputs to evaluate the faithfulness of the decontextualizations.

4.2.2. Retrievability

Retrievability operationalizes portability: if a premise is self-contained enough, then, given a claim as a query, standard retrieval methods should locate it more reliably in a corpus of prior fact-checks. To assess how well the resulting premises serve as a reusable knowledge base for claim and evidence matching (Panchendrarajan and Zubiaga, 2024), we simulate the task of retrieving relevant evidence for a new claim by searching a database of verified facts from prior fact-checks. Specifically, for each mode, we compile all premises from the collected articles into a search index. We treat each article's claim statement as a query and attempt to retrieve candidate evidence from the indexed premises of all fact-checks. We compute ranked lists over the mode-specific indexes and score effectiveness with standard information retrieval metrics. By comparing retrieval performance across modes, we aim to quantify the benefit of evidence decontextualization and open extraction on evidence matching. We interpret relative gains from Mode A→B/C as evidence that decontextualization improves cross-article portability.

4.2.3. Verification Utility

We investigate whether surfaced premises enable zero-shot claim verification. For each claim, a model receives the premises for that article and the label schema, and must output (i) a verdict from the allowed set and (ii) a brief justification that cites used premises via identifiers ι (Guo et al., 2022). Cited IDs make decisions traceable and let us quantify evidence use (Jolly et al., 2022). Alongside verdict accuracy, we report *citation coverage* as the fraction of presented premises that are cited: with S_{given} the shown premise IDs and $S_{\text{cited}} \subseteq S_{\text{given}}$ those mentioned, $C = |S_{\text{cited}}|/|S_{\text{given}}|$. Coverage contrasts evidence modes as more informative premises should improve task performance while citing fewer items. Accordingly, we treat coverage as diagnostic rather than a target, since correctness is our primary objective.

5. Experiments

5.1. Setup

We evaluate our approach on the collected and processed corpus of 13,106 PolitiFact fact-check articles (Section 3). Each fact-check instance provides a query claim and its extracted evidence set per mode. Because our method does not require model fine-tuning, we do not partition the data into training splits. Instead, all evidence extraction and verification experiments are conducted in a zero-shot inference setting with LLMs, respectively. We compare six publicly available state-of-the-art instruction-tuned LLMs of varying scale and architecture: Qwen3 at 8B, 14B, and 32B dense and 235B mixture of experts (MoE) with 22B active during inference (Yang et al., 2025), Llama 3.3 at 70B dense (Meta AI, 2024), and Llama 4 Scout MoE with 109B total and 17B active (Meta, 2025).

5.2. Automatic Evaluation

We operationalize the three research questions through complementary evaluations: retrieval performance measures portability (RQ2) by testing whether decontextualized premises are more discoverable across articles; verification performance measures whether portable evidence also improves downstream task accuracy; and faithfulness analysis ensures that LLM-based rewriting preserves factual content. By evaluating across six LLMs of varying scale and architecture and two verdict granularities (binary and five-class), we assess robustness (RQ3). We acknowledge that model-size differences confound model-specific and knowledge-driven effects and discuss this in Limitations.

5.2.1. Retrieval Performance

Each claim statement is used as a query to retrieve evidence sentences from an index of all extracted premises (Section 4), simulating cross-article evidence reuse. We use BM25 to construct an efficient retrieval index (Lù, 2024). We measure standard ranking metrics: Mean Reciprocal Rank at 10 (MRR@10), normalized Discounted Cumulative Gain (nDCG) at 3 and 10, and Recall (R) at 1, 3, 10, treating the premises from the claim's own fact-check as the relevant gold truth set. Higher MRR and nDCG indicate that the relevant evidence is ranked near the top, while higher Recall@k indicates more of the gold premises are retrieved within the top k results.

We first examine the effectiveness of evidence retrieval across different extraction modes. Table 1 compares decontextualized premises from Mode B against the baseline using verbatim sentences from Mode A. Decontextualization yields substantial gains in all metrics for every model. For instance, Llama-3.3-70B achieves an MRR@10 of 0.59 with Mode B, compared to 0.43 for the baseline, which is a 37% improvement. Recall@10 improves from 0.21 to 0.36, meaning the self-contained premises allow 71% more of the relevant evidence to be retrieved in the top-10 results. We observe consistent improvements at rank-3 as well (nDCG@3 from 0.26 to 0.41). These results confirm that making evidence sentences context-independent greatly increases their *portability* and discoverability by lexical-matching methods, addressing RQ2. Table 1 also shows retrieval results for Mode C with LLM-generated premises without anchor cues. Mode C outperforms the baseline by a wide margin. Across models, MRR@10 ranges from 0.78 to 0.88, indicating that a large share of queries have a relevant premise at rank 1. For example, Qwen3-8B and Qwen3-14B reach MRR@10 scores of 0.88 and 0.81, respectively. Recall@10 more than doubles relative to the baseline, reaching up to 0.57, with all models at 0.48 or higher, meaning that Mode C premises capture a larger portion of the self-selected evidence per claim. The retrieval scores for Mode C suggest that the LLMs extract evidence that is more directly aligned with the claim than anchored sentences from Mode A. This highlights a potential trade-off. While Mode C yields high recall and may surface additional relevant premises beyond explicit hyperlink anchors, it may also benefit from more direct lexical overlap with the claim wording. We return

	MRR	nDCG		Recall		
Model/Mode	10	3	10	1	3	10
Baseline A	0.43	0.26	0.23	0.10	0.15	0.21
Qwen3-8B						
B (decontext.)	0.47	0.30	0.27	0.11	0.18	0.25
C (open)	0.88	0.67	0.62	0.30	0.46	0.57
Qwen3-14B						
B (decontext.)	0.46	0.28	0.25	0.11	0.17	0.23
C (open)	0.81	0.59	0.54	0.27	0.40	0.50
Qwen3-32B						
B (decontext.)	0.50	0.32	0.29	0.12	0.20	0.27
C (open)	0.78	0.56	0.51	0.24	0.38	0.48
Qwen3-235B						
B (decontext.)	0.58	0.40	0.37	0.15	0.25	0.35
C (open)	0.81	0.62	0.57	0.26	0.41	0.54
Llama-4-Scout						
B (decontext.)	0.49	0.31	0.28	0.12	0.19	0.27
C (open)	0.80	0.59	0.54	0.26	0.40	0.50
Llama-3.3-70B						
B (decontext.)	0.59	0.41	0.38	0.15	0.25	0.36
C (open)	0.84	0.63	0.57	0.28	0.43	0.53

Table 1: Retrieval results across models.

	5-class		2-class	
Labels	**Count**	**%**	**Count**	**%**
true	1,513	11.5%	—	—
mostly-true	2,283	17.4%	3,794	35.6%
half-true	2,443	18.6%	—	—
mostly-false	2,425	18.5%	6,860	64.4%
false	4,442	33.9%	—	—

Table 2: Distribution of data for five- and two-class settings.

to this point in Section 6. Overall, the trend from Mode A $\rightarrow$ B $\rightarrow$ C is one of strictly increasing retrieval effectiveness, demonstrating the benefit of evidence decontextualization and open extraction.

5.2.2. Verification Performance

We evaluate label prediction using Macro-F_1 for both a binary setting, omitting `half-true` in the binary collapse, and a fine-grained five-class setting. Macro-F_1 is appropriate due to class imbalance, see Table 2, and the need to reward balanced performance across all verdict categories. We also quantify the evidence usage in each model's explanation by computing the citation coverage. Table 3 reports Macro-F_1 scores for each model and evidence mode, for both the binary and five-class verdict prediction tasks. Several clear patterns emerge. First, providing any evidence (Mode A) dramatically improves performance over the majority-class baseline which achieves Macro-F_1 of 0.39 for binary and 0.10 for five-class. Even the raw anchor sentences enable Macro-F_1 in the 0.57-0.68 range (binary) depending on the model, confirming that journalist-provided reference sentences capture relevant factual information needed to judge veracity. This supports RQ1.

Second, decontextualizing the evidence consistently boosts accuracy over Mode A. All models see an absolute Macro-F_1 gain of 4-7 points in the binary setting and up to 8 points in the five-class setting when using self-contained premises from Mode B. For example, Qwen3-32B improves from 0.59 to 0.66 (binary) and 0.27 to 0.28 (five-class). The largest jump is for Llama-3.3-70B, rising to 0.74 (binary) and 0.35 (five-class) in Mode B, a relative improvement of $\sim$8 and $\sim$27 percent, respectively. This indicates that evidence portability (RQ2) is not only beneficial for retrieval, but also aids the model in understanding and applying the evidence to the claim. By reducing ambiguity through operations such as resolving pronouns or making implicit context explicit, decontextualized premises make it easier for the verification model to connect facts to the claim.

Third, Mode C yields the highest verification performance across the board. Qwen3-235B and Llama-3.3-70B both reach binary Macro-F_1 scores of 0.81, and Llama-3.3-70B achieves the top five-class score of 0.42. Relative to Mode B, this corresponds to an additional gain of 12 points for Qwen3-235B in the binary setting and 7 points for Llama-3.3-70B in the five-class setting. Notably, the improvements from Mode B to Mode C are smaller than from Mode A to Mode B, suggesting diminishing returns and possible overlap between anchor-based and open-extracted evidence. On further investigation, we find that, on average, about 25% of the source references surfaced in Mode C overlap with anchor-based evidence from Mode A. The trend holds across all model sizes and for both binary and fine-grained tasks, supporting the robustness hypothesis in RQ3. We also observe that larger models tend to perform better overall. For instance, Llama-3.3-70B outperforms the smaller Qwen3 models in each mode, which is expected given its greater parametric knowledge.

	Two Class		Five Class	
Model/Mode	**Macro-F1**	**Coverage**	**Macro-F1**	**Coverage**
Baseline	0.39	—	0.10	—
Qwen3-8B				
A (linked)	0.58	0.82	0.21	0.75
B (decontext.)	0.63	0.64	0.22	0.63
C (open)	0.73	0.72	0.25	0.71
Qwen3-14B				
A (linked)	0.57	0.78	0.23	0.74
B (decontext.)	0.61	0.65	0.27	0.63
C (open)	0.72	0.77	0.34	0.74
Qwen3-32B				
A (linked)	0.59	0.71	0.27	0.71
B (decontext.)	0.66	0.64	0.28	0.64
C (open)	0.76	0.75	0.33	0.75
Llama-4-Scout				
A (linked)	0.65	0.74	0.27	0.74
B (decontext.)	0.70	0.65	0.27	0.65
C (open)	0.76	0.79	0.34	0.78
Llama-3.3-70B				
A (linked)	0.68	0.93	0.28	0.93
B (decontext.)	0.74	0.78	0.35	0.77
C (open)	**0.81**	0.86	**0.42**	0.85
Qwen3-235B				
A (linked)	0.62	0.82	0.26	0.83
B (decontext.)	0.69	0.70	0.30	0.46
C (open)	**0.81**	0.81	0.39	0.79

Table 3: Results across extraction modes (A-C) and models. Bold values indicate best Macro-F_1 per setting.

5.2.3. Evidence Faithfulness

As defined in Section 4.2.1, we quantify faithfulness with the *Decontextualization Faithfulness Score* (DFS), which combines forward textual entailment E with an explicit penalty for lexical copy-overlap. In all experiments, E is estimated by a standard DeBERTa-Large cross-encoder fine-tuned on SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). Table 4 reports mean forward entailment (E) and DFS for *Mode B* between anchored source sentences and their decontextualized premises and *Mode C* between the referenced source sentences and open-extracted premises. *Caveat:* DFS can underestimate quality when the original source sentence is already self-contained and well decontextualized, because high lexical overlap O depresses the score even if E is strong. Across models, three patterns emerge. First, in Mode B, smaller and mid-sized models achieve very high entailment with low DFS, indicating mostly minimal edits (e.g., Qwen3-14B has $E{=}0.91$ but DFS$=0.03$, and Qwen3-8B $E{=}0.81$ with DFS$=0.06$), whereas Qwen3-235B and Llama-3.3-70B strike a better balance with higher DFS (0.19 and 0.21) at moderate E (0.76 and 0.67), suggesting more substantive, portable rewrites rather than near-copies. Second, in Mode C, as expected for more abstractive generation, E decreases across models while DFS rises for some configurations, indicating non-trivial reformulations. Qwen3-235B attains the strongest DFS in Mode C (0.16), followed by Qwen3-8B (0.11) and Llama-3.3-70B (0.09), reflecting premises that are less verbatim yet still sufficiently supported to aid downstream use. Notably, Mode B outputs appear more faithful to their references than Mode C, plausibly because the two-step, anchor-driven pipeline with the anchor selection followed by constrained decontextualization biases rewrites toward the cited source, whereas open extraction has more freedom to abstract and synthesize. Third, forward entailment alone tends to overestimate trivial copy-edits, while DFS differentiates portable decontextualizations from near-verbatim text. Models with higher DFS in Mode B (Qwen3-235B, Llama-3.3-70B) also yield strong retrieval and verification results (Table 1, Table 3), aligning faithfulness with utility, and although Mode C is more abstractive (lower E), its DFS indicates that many generated premises remain useful as complementary element to anchor-driven evidence.

5.3. Human Study

To complement the automatic metrics, we conducted a manual annotation study of the evidence to evaluate extraction utility. We randomly sampled 100 premises each from Mode B and Mode C outputs. The sample covered premises extracted by the strongest overall model, Qwen3-235B, to evaluate best-case outputs. Each article contributed at most one sampled item to diversify topics. Two annotators with a background in fact-checking independently labeled all items, using annotation guidelines, and were tasked with assessing (a) whether the statement is self-contained and interpretable without surrounding context, and (b) the evidence type assigned to the statement: *Document*, *Statistic*, *Quote*, or *Context*. Question (a) was rated on an ordinal scale from incomplete (1) to complete (3). For Mode B (a), the results show an observed agreement rate of 0.87 and a Krippendorff's alpha of 0.255 due to both annotators agreeing on a majority of cases to be self-contained (3). For Mode B (b), we measure an observed agreement of 0.58 and a Krippendorff's alpha of 0.441 due to significant disagreements on the `CONTEXT` label. For Mode C (a), the results show an observed agreement rate of 0.835 and a Krippendorff's alpha of 0.474. For Mode C (b), we measure an observed agreement of 0.67 and a Krippendorff's alpha of 0.561. After resolving evidence-type annotation disagreements through discussion, with the final label restricted to one of the two original annotator choices, Qwen3-235B achieves a Macro-F_1 of 0.859 for Mode B and 0.857 for Mode C. Furthermore, we did not identify label leakage or factual inconsistencies in either Mode B or Mode C within the annotated samples.

6. Discussion

Our results show that fact-checking articles contain reusable evidence that can be systematically unlocked. In-text hyperlinks provide a strong and scalable signal for locating evidence-bearing statements, and decontextualizing these statements into stand-alone premises consistently improves both retrieval and verification. This suggests that fact-checkers' sourcing practices can be repurposed to build structured evidence resources for automated fact-checking. In this sense, `PrimeFacts` can be interpreted as an intermediate layer between doc-

	Mode B		Mode C	
Model	E	DFS	E	DFS
Qwen3-8B	0.81	0.06	0.50	0.11
Qwen3-14B	**0.91**	0.03	**0.65**	0.06
Qwen3-32B	0.86	0.09	0.60	0.07
Qwen3-235B	0.76	0.19	0.50	**0.16**
Llama-4-Scout	0.69	0.07	0.42	0.05
Llama-3.3-70B	0.67	**0.21**	0.39	0.09

Table 4: Results for Mean Forward Entailment and DFS for Mode B and Mode C. Bold values indicate best performance per metric.

ument retrieval and claim verification. Instead of reasoning directly over long source documents, verification models operate on compact, decontextualized premises that explicitly encode the factual content of the evidence.

The comparison between Modes B and C reveals a clear trade-off. Mode B provides grounded, source-linked premises that remain close to the journalist's explicitly cited evidence, while Mode C often surfaces additional relevant premises beyond hyperlink anchors and achieves the strongest downstream performance. At the same time, open extraction can introduce redundancy or produce statements with high lexical overlap to the claim itself. On average, about 25% of the source references surfaced in Mode C overlap with anchor-based evidence from Mode A, indicating partial but nontrivial complementarity rather than simple duplication. In practice, this suggests a hybrid strategy: use Mode B as a faithful foundation and supplement it with non-redundant Mode C premises to improve coverage.

We also observe that decontextualized premises often support stronger predictions while requiring fewer cited items in the generated justifications. This suggests that self-contained premises may be individually more informative for decision-making, although citation coverage should be interpreted cautiously because it also reflects model selection behavior. One explanation is that decontextualization removes discourse dependencies that would otherwise require models to reconstruct context from surrounding text. By converting evidence into self-contained premises, the reasoning task becomes closer to structured factual inference than long-context interpretation. Faithfulness analyses further indicate that stronger models tend to produce supported rewrites rather than verbatim copies, which aligns with their improved downstream verification utility.

These trends were consistent across model families and across both binary and five-class verdict settings, indicating that the gains stem from the evidence representation rather than from a single model architecture. Overall, PrimeFacts shows that evidence extracted from professional fact-checks can support retrieval-augmented and semi-automated verification workflows, although human oversight remains important when generated premises are used in high-stakes settings.

7. Conclusion

We presented PrimeFacts, a methodology and resource for transforming full-length fact-checking articles into a reusable evidence resource, and demonstrated its value for automated misinformation detection. Our framework leverages fact-checkers' own sourcing practices by using hyperlink-anchored evidence, decontextualizing these statements into stand-alone premises, and investigating whether similar evidence can also be extracted without relying on anchors. This yields structured evidence representations that are suitable for downstream retrieval and verification. Our findings support the core assumptions of the paper. First, in-article hyperlinks provide a strong and scalable signal for identifying evidence-bearing content. Second, rewriting anchored evidence into decontextualized premises substantially improves both cross-article retrieval and verdict prediction. Third, these improvements remain consistent across different verdict granularities and model architectures. Together, these results show that evidence extracted from professional fact-checks can serve as an effective intermediate representation between long-form journalistic articles and automated claim verification.

By introducing the PrimeFacts resource and extraction methodology, we aim to support future research on retrieval-augmented fact-checking, evidence reuse, and transparent decision support. More broadly, our work suggests that professional fact-checks are not only useful as final verdicts, but also as rich repositories of structured evidence that can support more transparent, reusable, and effective verification systems.

Ethical Considerations

Our target is to develop PrimeFacts as a structured knowledge base for intelligent decision-support systems in fact-checking and related applications. While it enables automated evidence retrieval and verdict prediction, these functions are designed to assist rather than replace human judgment, particularly in high-stakes or politically sensitive contexts. Collaborative human oversight remains essential to interpret nuance, context, and evolving facts. The evidence extracted in PrimeFacts reflects how fact-checkers present and justify information within their articles. Both the selection and presentation of evidence may encode subtle biases from the authors, such as framing, emphasis, or omission of counterpoints, which our extraction pipeline may in turn reproduce. Similarly, while fact-checking organizations are reputable and adhere to editorial standards, their verdicts and accompanying justifications are not free from subjective interpretation and editorial policies. These judgments can be influenced by institutional perspectives, available sources, or political context. Analyses or systems built on this resource should therefore explicitly account for such potential biases. The dataset is derived from copyrighted fact-checking articles. We publicly release only derived

metadata and annotations. Original fact-check article texts are not redistributed for copyright reasons.

Limitations

Our work has some limitations that suggest avenues for future work. One primary limitation is the reliance on explicit in-text citations (hyperlinks) to identify evidence. While PolitiFact articles are richly linked, some fact-checks or segments rely on implicit evidence or general knowledge that is not captured by a specific hyperlink. Our pipeline would miss such uncited yet important premises. In domains or languages where fact-checkers provide fewer references, a hybrid extraction strategy, combining the anchor-based method with additional open extraction, may be necessary to achieve high recall. Another limitation lies in the scope of the extracted evidence. We isolate individual supporting sentences but do not explicitly capture the logical structure or multi-hop reasoning that a fact-checker might apply across an article. For example, an article might piece together two separate facts to reach a conclusion, but our current method would list these facts separately without representing their inferential connection. This could limit the usefulness of the evidence in tasks requiring joint reasoning. Future extensions should link premises into argumentative chains and label their roles, enabling a closer mirroring of human reasoning steps. The use of large language models for evidence rewriting and generation introduces additional considerations. Although we took measures to preserve faithfulness, such as constrained prompting and post-hoc entailment checks, LLMs can occasionally produce subtly altered or extraneous details. In our manual evaluation we did not observe major factual errors, but there remains a risk of hallucination, especially as we push the models to be more abstractive across long contexts. Users of the `PrimeFacts` framework should treat decontextualized premises as suggestions to be compared against the original anchor or reference statements. Furthermore, our evaluation is conducted exclusively on PolitiFact, a single English-language fact-checking platform with a particular editorial style and sourcing convention. While Mode C is platform-agnostic by design and Mode B requires only that articles contain hyperlinks, we have not yet validated our pipeline on other platforms (e.g., Full Fact, Snopes) or languages. Cross-platform and multilingual evaluation is planned as future work. We also note that our operationalization of portability through retrieval performance and robustness through multi-model, multi-granularity evaluation may not capture all aspects of these concepts. Model-size differences in our LLM comparisons introduce a confound, since larger models have both more parametric knowledge and better instruction-following ability; disentangling these factors is an open question. Finally, we have not tested robustness to adversarial noise or contradictory evidence in the premise set, which would be a valuable stress test for future work.

Acknowledgments

This work is funded by the German Federal Ministry for Research, Technology and Aeronautics (BMFTR) in the scope of the projects *newspolygraph* (03RU2U151C) and *FAR-REASONING* (16IS23068). This work is supported by JST CREST Grants (JPMJCR20D3), Japan.

References

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)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

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.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Regina Cazzamatta. 2025a. Building a cross-border european information network: Hyperlink connections among fact-checking organizations. *Media and Communication*, 13.

Regina Cazzamatta. 2025b. Decoding Correction Strategies: How Fact-Checkers Uncover Falsehoods Across Countries. *Journalism Studies*, 26(7):777–799.

Regina Cazzamatta. 2025c. Redefining objectivity: Exploring types of evidence by fact-checkers in four European countries. *European Journal of Communication*, 40(2):144–164.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Dong Yu, and Hongming Zhang. 2024. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making Sentences Stand-Alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Xingyu Deng, Xi Wang, and Mark Stevenson. 2025. The next phase of scientific fact-checking: advanced evidence retrieval from complex structured academic papers. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 436–448.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lucas Graves. 2016. *Deciding What's True: The Rise of Political Fact-Checking in American Journalism*. Columbia University Press, New York.

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*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Edda Humprecht. 2020. How Do They Debunk "Fake News"? A Cross-National Comparison of Transparency in Fact Checks. *Digital Journalism*, 8(3):310–327.

Shan Jiang, Simon Baumgartner, Abe Ittycheriah, and Cong Yu. 2020a. Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles. In *Proceedings of The Web Conference 2020*, WWW '20, pages 1592–1603, New York, NY, USA. Association for Computing Machinery.

Shan Jiang, Simon Baumgartner, Abe Ittycheriah, and Cong Yu. 2020b. Factoring fact-checks: Structured information extraction from fact-checking articles. In *Proceedings of The Web Conference 2020*, pages 1592–1603.

Shailza Jolly, Pepa Atanasova, and Isabelle Augenstein. 2022. Generating Fluent Fact Checking Explanations with Unsupervised Post-Editing. *Information*, 13(10):500.

Lasha Kavtaradze. 2024. Challenges of automating fact-checking: A technographic case study. *Emerging Media*, 2(2):236–258.

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new Dataset for Claim Entailment and Inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304, Dublin, Ireland. Association for Computational Linguistics.

Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly context not found: one in five articles suffers from reference rot. *PloS one*, 9(12):e115253.

Belinda Z. Li, Emmy Liu, Alexis Ross, Abbas Zeitoun, Graham Neubig, and Jacob Andreas. 2025. Language modeling with editable external knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3070–3090. Association for Computational Linguistics.

Xing Han Lù. 2024. BM25S: Orders of magnitude faster lexical search via eager sparse scoring.

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.

Iffat Maab, Edison Marrese-Taylor, Sebastian Padó, and Yutaka Matsuo. 2024. Media bias detection across families of language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4083–4098, Mexico City, Mexico. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and

Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Meta. 2025. Llama 4 scout (17b, 16 experts) — instruct version. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct. Accessed: 2025-10-13.

Meta AI. 2024. Llama 3.3 70B Instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed: 2025-10-13.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress Test Evaluation for Natural Language Inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mucahid Kutlu, and Yavuz Selim Kartal. 2021. Overview of the CLEF–2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-Hop Fact Checking of Political Claims.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim Detection for Automated Fact-checking: A Survey on Monolingual, Multilingual and Cross-Lingual Research. *Natural Language Processing Journal*, 7:100066.

Premtim Sahitaj, Iffat Maab, Junichi Yamagishi, Jawan Kolanowski, Sebastian Möller, and Vera Schmitt. 2025. Towards Automated Fact-Checking of Real-World Claims: Exploring Task Formulation and Assessment with LLMs.

Chris Samarinas, Wynne Hsu, and Mong-Li Lee. 2021. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91.

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

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

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.

Xiaoou Wang, Elena Cabrio, and Serena Villata. 2025a. Safe: Structured argumentation for fact-checking with explanations. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 11114–11118. International Joint Conferences on Artificial Intelligence Organization. Demo Track.

Xiaoou Wang, Elena Cabrio, and Serena Villata. 2025b. When automated fact-checking meets argumentation: Unveiling fake news through argumentative evidence. *Argument and Computation*, 16.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Albert Weichselbraun. 2021. Inscriptis - a python-based HTML to text conversion library optimized

for knowledge extraction from the web. *Journal of Open Source Software*, 6(66):3557.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Rui Xing, Timothy Baldwin, and Jey Han Lau. 2024. Evaluating transparency of machine generated fact checking explanations. *arXiv e-prints*, pages arXiv–2406.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Fengzhu Zeng and Wei Gao. 2024. JustiLM: Few-shot Justification Generation for Explainable Fact-Checking of Real-world Claims. *Transactions of the Association for Computational Linguistics*, 12:334–354.

Ke Zhou, Claire Grover, Martin Klein, and Richard Tobin. 2015. No more 404s: predicting referenced link rot in scholarly articles for pro-active archiving. In *Proceedings of the 15th ACM/IEEE-CS joint conference on digital libraries*, pages 233–236.

A. Appendix

A.1. Mode B: Decontextualization Prompt

For each anchor sentence, the LLM receives a system prompt followed by a user message. The system prompt instructs the model to produce a single decontextualized sentence with an evidence-type category via structured JSON output:

> **System:** "You are a careful scientific editor. Produce ONE decontextualized sentence that stands alone, explicitly preserving or adding entities, numbers, dates that make the sentence clear even when read outside of the article. Assign a category label using exactly one of: QUOTE, STATISTIC, DOCUMENT, CONTEXT, OTHER. [...category guide with definitions and examples...] Return JSON only."
>
> **User:** "Claim: {claim} \n Article (labeled): {segmented_article} \n Target letter: {letter} \n Target sentence: {target_sentence} \n Return JSON only."

The JSON schema constrains the output to three fields: `letter` (the segment identifier), `decontextualized` (the rewritten sentence), and `category` (one of the five evidence types).

A.2. Mode C: Open Extraction Prompt

Mode C receives the full article and extracts multiple premises at once. The system prompt specifies a bounded range of premises and uses the same category guide:

> **System:** "You are a careful scientific editor. Extract {min}–{max} non-redundant key premises from the labeled article. For each premise, provide: (a) exactly ONE letter anchor from the article that supports it; (b) ONE decontextualized sentence that stands alone; and (c) a category using exactly one of: QUOTE, STATISTIC, DOCUMENT, CONTEXT, OTHER. [...category guide...] Output JSON only."
>
> **User:** "Claim: {claim} \n Article (labeled): {segmented_article} \n Return JSON only."

The output schema constrains the response to a list of premise objects, each with `letter`, `decontextualized`, and `category` fields. The maximum list length is set to the number of Mode A anchors for that article.