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

Premtim Sahitaj[12], Jawan Kolanowski[3], Ariana Sahitaj[12], Veronika Solopova[12], Max Upravitelev[12], Daniel Röder[2], Iffat Maab[4], Junichi Yamagishi[4], Sebastian Möller[12], Vera Schmitt[12]

[1]Technische Universität Berlin, Quality and Usability Lab [2]DFKI, Berlin [3]Harz University of Applied Sciences [4]National Institute of Informatics, Tokyo

LREC 20 Palma 26

Motivation

The Problem

- Evidence in fact-check articles is locked in unstructured text
- Not reusable for automated verification systems
- Similar claims are re-verified redundantly across media

Our Approach

- Exploit in-article hyperlinks as natural evidence anchors
- Decontextualize anchor sentences into standalone premises via LLMs
- Build PrimeFacts: a reusable, structured evidence knowledge base

Research Questions

RQ1 Addressability

Can in-article hyperlinks reliably identify the core evidence supporting or refuting a claim?

RQ2 Portability

Does rewriting premises into stand-alone statements improve retrieval and automated verification?

RQ3 Robustness

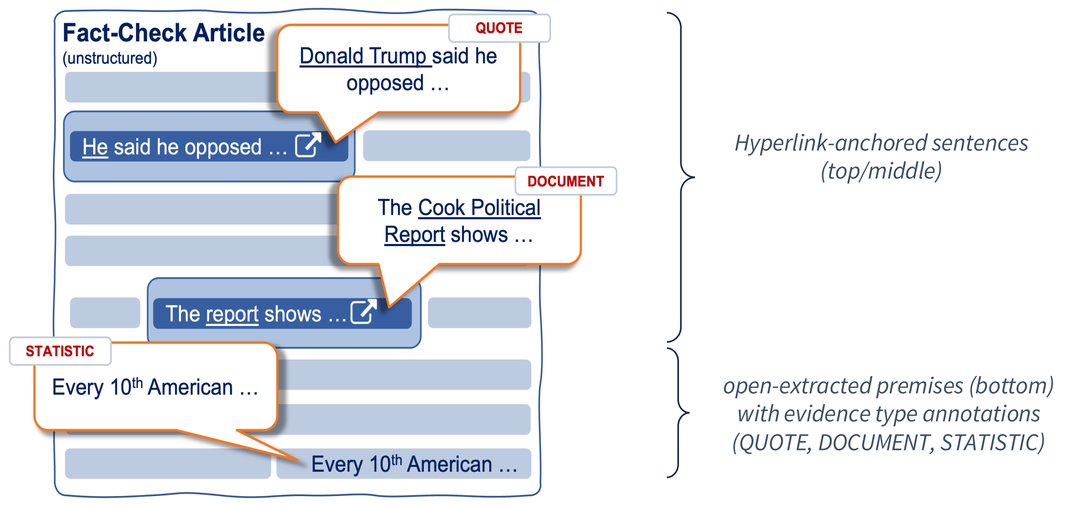
Are the findings consistent across different evaluation settings and task granularities?

PrimeFacts: Resource Overview

13,106	49,718	661
Fact-Check Articles	In-Article Anchors	Journalists

PolitiFact articles (up to Sep 2025) with claims, verdicts, sources. Cross-referenced via canonical URLs. Layout-aware normalization, sentence segmentation, label-leak filtering.

Figure 1: Illustrative Example



Extraction Framework: Three Evidence Modes

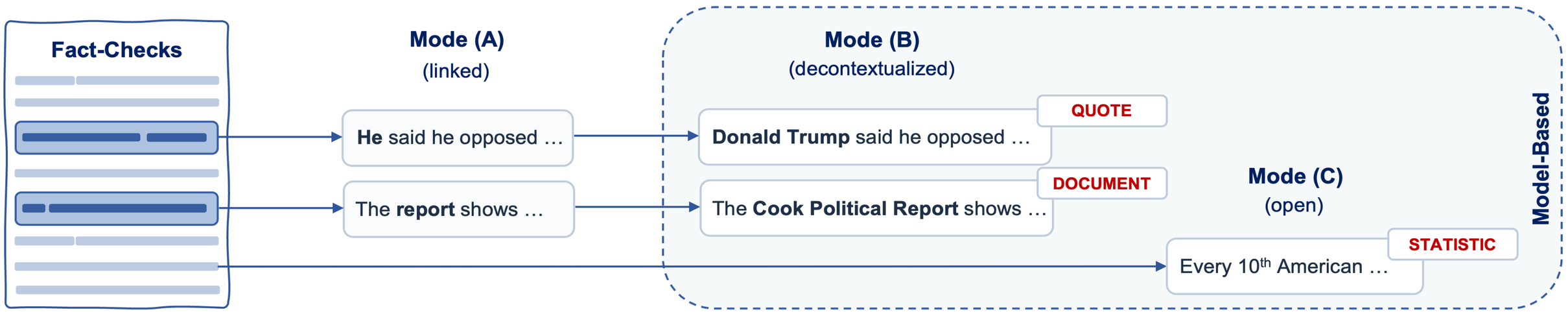
Mode A: Anchored
Verbatim hyperlink-anchored sentences retained as-is

Mode B: Decontextualized
LLM rewrites anchors into standalone, context-independent premises

Mode C: Open Extraction
LLM extracts self-contained premises from full article text without anchor constraints

Example: A: "He took office in 2019" → B: "Volodymyr Zelenskyy took office as President of Ukraine in May 2019"

Figure 2: Extraction Pipeline



Evaluation Framework

Faithfulness (DFS)
Forward entailment penalized by lexical overlap

Retrievability
BM25 cross-article retrieval using claim as query

Verification Utility
Zero-shot Macro-F_1 (2-class and 5-class)

$$DFS(D) = (1/|D|) \sum E(p,s) \cdot (1 - O(p,s))$$

E = forward textual entailment (NLI), O = asymmetric lexical overlap. Rewards substantive rewrites; penalizes near-copies.

Results: Cross-Article Retrieval

Model / Mode	MRR@10	nDCG@3	Recall@10
Baseline A	0.43	0.26	0.21
Qwen3-8B B	0.47	0.30	0.25
Qwen3-8B C	**0.88**	**0.67**	**0.57**
Qwen3-235B B	0.58	0.40	0.35
Qwen3-235B C	**0.81**	**0.62**	**0.54**
Llama-3.3-70B B	0.59	0.41	0.36
Llama-3.3-70B C	**0.84**	**0.63**	**0.53**

▲ Mode A → B → C: strictly increasing retrieval. Decontextualization yields up to 30% relative MRR gain. Mode C reaches MRR@10 of 0.84–0.88.

Results: Claim Verification

Model / Mode	2-Class F_1	5-Class F_1	Coverage
Baseline (majority)	0.39	0.10	—
Qwen3-8B A	0.58	0.21	0.82
Qwen3-8B B	0.63	0.22	0.64
Qwen3-8B C	**0.73**	**0.25**	**0.72**
Llama-3.3-70B A	0.68	0.28	0.93
Llama-3.3-70B B	0.74	0.35	0.78
Llama-3.3-70B C	**0.81**	**0.42**	**0.86**

▲ Evidence raises Macro-F_1 by 10–20 points over baseline. Decontextualized premises consistently outperform verbatim anchors. Best: Llama-3.3-70B Mode C (0.81 / 0.42).

Faithfulness & Human Evaluation

DFS Faithfulness

- Smaller models: high entailment, low DFS → mostly trivial copies
- Larger models (Qwen3-235B, Llama-70B): better balance, substantive rewrites
- Mode B more faithful than Mode C (anchor-driven grounding)

Human Evaluation

87%	F_1=0.86	0
Self-containment agreement	Evidence type classification	Factual errors detected

Key Takeaways

Hyperlinks are reliable evidence markers
In-text citations provide a strong, scalable signal for locating evidence-bearing content (RQ1).

Decontextualization improves portability
Stand-alone premises improve retrieval (+30% MRR) and verification (+10–20pt F_1) (RQ2).

Robust across settings
Consistent gains across 6 LLMs, 2 verdict granularities, and all evaluation dimensions (RQ3).

Hybrid strategy recommended
Use Mode B as faithful foundation, supplement with non-redundant Mode C for broader coverage.

Acknowledgments & References

Funded by BMFTR (news-polygraph 03RU2U151C, FAR-REASONING 16IS23068). Supported by JST CREST Grants (JPMJCR20D3), Japan.

Selected References

[1] Thorne et al. (2018). FEVER: A Large-scale Dataset for Fact Extraction and VERification. NAACL.
[2] Choi et al. (2021). Decontextualization: Making Sentences Stand-Alone. TACL.
[3] Chen et al. (2024). Dense X Retrieval: What Retrieval Granularity Should We Use? EMNLP.
[4] Gunjal & Durrett (2024). Molecular Facts: Desiderata for Decontextualization. EMNLP Findings.
[5] Schlichtkrull et al. (2023). AVeriTeC: Real-World Claim Verification. NeurIPS.