SEEK: Semantic Evidence Extraction via Adaptive ChunKing for Multilingual Fact-Checking

Anonymous ACL submission

Abstract

Multilingual fact verification requires evidence that is both relevant and sufficiently complete for reliable factuality prediction. However, existing systems often rely on search snippets, sentence-level evidence, or locally segmented passages, which can miss decisive context and produce fragmented evidence. To overcome these limitations, we propose SEEK, a *S*emantic *E*vidence *E*xtraction with Adaptive chun*K*ing framework that constructs coherent evidence chunks from full fact-checking articles by identifying semantic topic transitions and preserving local verification context. The constructed chunks are encoded with a multilingual encoder, and multilingual LLMs are then fine-tuned with LoRA adapters for veracity prediction. Experiments on X-FACT and RU22FACT show that SEEK improves Macro-F1 by up to **10**% over semantic chunking, **19**% over sentence chunking, and **20**% over search-snippet baselines. Evidence completeness and significance analyses further show that SEEK preserves richer verification context and enables more reliable multilingual fact-checking.

1 Introduction

The multilingual nature of today's online information ecosystem has made misinformation increasingly difficult to detect, contextualize, and verify (Panchendrarajan and Zubiaga, 2024). A single misleading claim may circulate across languages, platforms, and regional communities, often appearing in modified forms through translation, paraphrasing, or culturally specific framing (Quelle et al., 2025; Peng et al., 2025b). In such settings, fact verification is not limited to deciding whether a claim is true or false; it also requires identifying reliable evidence, preserving the meaning of the claim across languages, and reasoning over heterogeneous sources that may differ in structure, language, and level of detail (Guo et al., 2022; Gupta and Srikumar, 2021a).

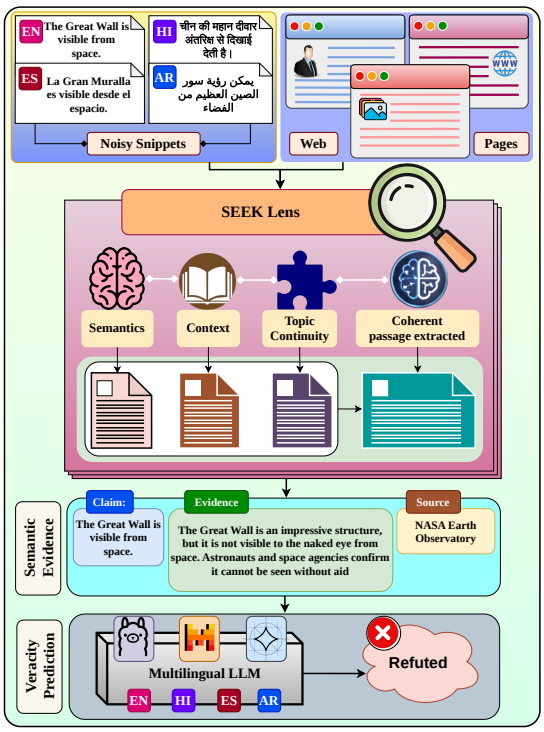

Figure 1: High-level illustration of the proposed multilingual fact verification framework. Noisy multilingual snippets and full web pages are transformed through a SEEK lens into coherent evidence, which is then used by a multilingual LLM for veracity prediction.

Human fact-checkers remain central to this process because they can interpret context, compare sources, and evaluate evidence with linguistic and domain-specific judgment (Nakov et al., 2021). However, manual fact-checking is difficult to scale against the volume and speed of online misinformation (Nakov et al., 2021; Nanhekhan et al., 2025). This has motivated automated fact verification, where NLP systems retrieve relevant evidence and predict claim veracity (Guo et al., 2022; Zheng et al., 2024a). Despite recent progress, system reliability depends strongly on evidence quality (Zheng et al., 2024a; Nanhekhan et al., 2025), especially

in multilingual settings where relevant information may be scattered across long fact-checking articles, news reports, official documents, or web pages (Zheng et al., 2024a; Gupta and Srikumar, 2021a; Cekinel et al., 2024; Peng et al., 2025b; Schlichtkrull et al., 2023). Such sources often mix background discussion, quoted claims, contextual explanations, and final verdicts within the same document (Augenstein et al., 2019; Schlichtkrull et al., 2023). As a result, short snippets or isolated sentences may miss the complete verification context, whereas full-document processing can introduce substantial irrelevant information (Zheng et al., 2024a; Schlichtkrull et al., 2023). This creates a mismatch between the retrieved evidence unit and the evidence span actually required for reliable claim verification (Zheng et al., 2024a; Schlichtkrull et al., 2023).

Existing evidence retrieval strategies often rely on fixed-size passages, sentence-level segmentation, or externally retrieved snippets (Chen et al., 2022; Schlichtkrull et al., 2023). While simple and efficient, these approaches may not align with the semantic structure of fact-checking documents (Zhang et al., 2023). Fixed windows can split reasoning across boundaries, whereas sentence-level retrieval may discard the surrounding context needed for correct interpretation (Zhang et al., 2023; Zheng et al., 2024a). This limitation is amplified in multilingual settings, where translation variation, code-mixing, and language-specific discourse patterns can weaken claim-evidence alignment (Panchendrarajan and Zubiaga, 2024; Peng et al., 2025b).

To address these limitations, we propose SEEK, a Semantic Evidence Extraction with adaptive chunKing framework for multilingual fact verification. Rather than relying on short snippets or fixed-length passages, SEEK constructs coherent evidence chunks from full web documents by detecting semantic topic shifts in a shared multilingual embedding space. These chunks preserve complete verification context while reducing irrelevant noise, and are then retrieved for veracity prediction using LoRA-tuned (Hu et al., 2022) multilingual large language models. The central idea is that reliable multilingual fact verification requires not only strong language models, but also better evidence granularity. Improving how evidence is segmented and retrieved can therefore enhance the reliability of downstream veracity prediction across multilingual settings.

Our contributions are as follows:

- We introduce SEEK, a multilingual evidence construction framework that combines contextual topic-shift detection, score smoothing, adaptive thresholding, and boundary overlap to generate verification-oriented evidence chunks.
- We combine SEEK with multilingual dense retrieval and LoRA-tuned multilingual LLMs, achieving strong veracity prediction performance across multilingual fact-checking benchmarks.
- We perform comprehensive evidence analysis on X-FACT and RU22FACT, including evidence completeness, similarity, and statistical significance studies across multilingual and generalization settings.
- We further conduct translation-based evaluation to analyze the effect of language normalization on evidence retrieval and veracity prediction in multilingual fact verification.

2 Related Work

2.1 General and Multilingual Fact-Checking

Automated fact-checking is typically framed as verifying a claim using either model-internal knowledge or external evidence. Benchmarks such as FEVER (Thorne et al., 2018) and LIAR (Wang, 2017) established evidence-based and fine-grained fact verification in English, while multilingual datasets such as X-FACT (Gupta and Srikumar, 2021b) and RU22FACT(Zeng et al., 2024) extend this task across languages, domains, and resource settings. These benchmarks show that multilingual fact verification depends not only on strong veracity models, but also on evidence that is relevant, complete, and linguistically usable for the target claim.

2.2 Evidence Extraction and Retrieval

Snippet- and sentence-level retrieval. Many fact-checking systems rely on search snippets, individual sentences, or short passages as evidence (Gupta and Srikumar, 2021a). These units are efficient and focused, but they often fragment the verification context. In multilingual settings, this issue is amplified because snippets may omit key background, distort named entities, or miss the link between the claim, investigation, and verdict (Gupta and

Srikumar, 2021a). Dense retrievers such as DPR and multilingual-e5 improve semantic matching between claims and evidence candidates (Karpukhin et al., 2020; Wang et al., 2024). Similarly, CONCRETE improves cross-lingual fact-checking by learning multilingual retrieval representations from trusted multilingual corpora (Huang et al., 2022). However, these retrieval methods do not directly resolve the granularity problem: the retrieved unit may still be too short to contain complete verification context or too broad to avoid irrelevant information.

Document chunking methodologies. Other approaches retrieve evidence from longer documents by splitting them into fixed-size, sentence-aware, or semantic chunks (Qu et al., 2025; Kiss, 2025). Fixed and sentence-based chunking are simple, but they can cut off decisive information when verification cues appear beyond the boundary (Qu et al., 2025). Semantic chunking improves coherence, yet local similarity-based boundaries can still be unstable in noisy multilingual fact-checking articles that mix claims, explanations, quotes, and verdicts (Qu et al., 2025; Kiss, 2025). This creates a key trade-off: shorter units reduce noise but risk context fragmentation, while longer units preserve context but may introduce irrelevant information (Qu et al., 2025).

2.3 Evidence Granularity for Multilingual Verification

Prior work has advanced multilingual fact-checking through better datasets, cross-lingual retrieval methods such as CONCRETE, translation-based normalization, and LLM-based verification (Huang et al., 2022; Peng et al., 2025a). However, the role of evidence granularity remains under-explored: retrieved text must not only be semantically related to the claim, but also complete enough to support factuality prediction (Viswanathan et al., 2025). This is especially important for multilingual verification, where systems must handle language variation, low-resource settings, noisy web pages, and dispersed evidence cues (Peng et al., 2025a; Viswanathan et al., 2025). This gap motivates an evidence construction strategy that operates beyond isolated snippets or fixed windows while still avoiding overly broad document segments. By aligning chunk boundaries with contextual topic shifts and retaining boundary overlap, the proposed approach preserves the continuity between the claim, supporting details, and verdict. In this way, it reduces context fragmentation in evidence retrieval and provides more complete verification evidence for multilingual fact-checking.

Datasets	No. of Classes	Language	Train	Val	Test	Total
MULTILINGUAL						
-X-FACT	7	25 languages	19,079	2,535	9,575	31,189
-RU22FACT	3	4 languages	11,217	1,600	3,216	16,033

Table 1: Details of multilingual datasets considered in our work.

3 Datasets Details

In this section, we report the details of datasets used to evaluate our study. Since our work focuses on multilingual fact-checking, we consider two multilingual benchmarks: **X-FACT** (Gupta and Srikumar, 2021b) and **RU22FACT** (Zeng et al., 2024). These datasets allow us to evaluate the proposed framework across diverse languages, claims, and evidence sources. The statistical details of both datasets are summarized in Table 1. Additional dataset descriptions are provided in Appendix A.

4 Methodology

The overall fact-checking pipeline used in this work consists of four key components: (i) a web crawler, (ii) the SEEK chunking module, (iii) a multilingual dense retriever, and (iv) instruction-tuned large language models, namely LLaMA, Gemma, and Mistral. The individual components are described below:

Web crawler: In the X-FACT dataset (Gupta and Srikumar, 2021b), each claim is accompanied by five Google Search snippets and their corresponding source URLs. Since these snippets are often insufficient for reliable verification, we crawl the full content of each URL using **Crawl4AI** (UncleCode, 2024). The crawler removes boilerplate elements such as navigation bars and scripts and extracts the main textual content, yielding up to five documents per claim, with lengths ranging from hundreds to several thousand tokens. In contrast, each claim in RU22FACT is linked to a single source URL, which we retrieve and clean using the same pipeline.

Chunking Module: Due to the long and unstructured nature of crawled web pages, the extracted documents are divided into smaller textual units before evidence retrieval. Effective chunking should balance efficiency and completeness: very short chunks may miss the context needed for verifica-

tion, while very long chunks may introduce irrelevant content. Existing sentence-based or semantic chunking methods can still produce incomplete evidence when verification cues are spread across nearby but separated parts of a document. To address this, SEEK uses context-window topic-shift detection, smoothing, adaptive thresholding, and overlap-aware chunk construction to form coherent evidence passages.As shown in Appendix Figure 11, this helps preserve both the viral claim context and the later refuting evidence within the same retrieval unit.

Baseline chunking methods. We compare SEEK with two common document chunking strategies. Sentence-aware fixed-size chunking divides a document into consecutive groups of complete sentences under a fixed token budget of 512 tokens. This method preserves sentence boundaries and avoids cutting sentences in the middle, but it does not consider topic changes within the document. Therefore, it may either stop before the full verification context is captured or combine unrelated sentences in the same chunk. Semantic chunking (LangChain, 2024), in contrast, uses a multilingual sentence encoder to map each sentence s_i into an embedding $\mathbf{e}_i = f(s_i)$ and places a boundary when the cosine similarity between neighboring sentence embeddings falls below a threshold τ. Although this captures local semantic changes, it mainly relies on adjacent sentence similarity and can still produce incomplete or noisy evidence when the verification context spans multiple sentences.

SEEK. Although standard semantic chunking is more meaningful than fixed-size splitting, it can still be unstable for noisy multilingual web documents. A boundary decision based only on adjacent sentences is sensitive to local variations, translation style, and short sentence noise. To address this limitation, we propose a strategy, SEEK, that detects topic shifts using context-window comparison rather than only sentence-to-sentence similarity.

For each possible boundary position i, we define a left context window and a right context window:

$$L_i = \{s_{i-w+1}, \ldots, s_i\}, \quad R_i = \{s_{i+1}, \ldots, s_{i+w}\},$$

where w is the window size (set to $w = 3$ in our experiments). The embeddings of these two windows are computed by averaging sentence embeddings:

$$\mathbf{l}_i = \frac{1}{|L_i|} \sum_{s_j \in L_i} \mathbf{e}_j, \quad \mathbf{r}_i = \frac{1}{|R_i|} \sum_{s_j \in R_i} \mathbf{e}_j.$$

We then compute a semantic shift score:

$$\Delta_i = 1 - \cos(\mathbf{l}_i, \mathbf{r}_i).$$

A higher value of Δ_i indicates a stronger semantic change between the left and right contexts, suggesting a possible topic boundary. To reduce noisy fluctuations, the shift scores are smoothed:

$$\tilde{\Delta}_i = \frac{1}{k} \sum_{j=i-\lfloor k/2 \rfloor}^{i+\lfloor k/2 \rfloor} \Delta_j,$$

where k (set to $k = 3$ in our experiments) is the smoothing window size.

Instead of using a fixed threshold, we apply adaptive thresholding based on the distribution of shift scores within the document:

$$\tau = \text{Percentile}(\tilde{\Delta}, p),$$

where p controls the selectivity of boundary detection and is set to $p = 95\%$ in our experiments. A boundary is selected when:

$$b_i = \begin{cases} 1, & \text{if } \tilde{\Delta}_i \geq \tau, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, sentences between selected boundaries are grouped into chunks under a maximum token budget. We also include a small overlap between consecutive chunks to preserve boundary-level context. This produces chunks that are semantically coherent, token-efficient, and more suitable for dense evidence retrieval in multilingual fact-checking. An example visualization of the raw semantic shift scores, smoothing process, and adaptive boundary selection is provided in **Appendix B**.

Multilingual dense retriever: After generating semantically coherent chunks with SEEK, we retrieve claim-relevant evidence using a multilingual dense retrieval pipeline. Each claim and candidate chunk are encoded with **multilingual-e5-large-instruct** (Wang et al., 2024), which maps multilingual claims and evidence passages into a shared semantic embedding space.

Given a claim c and crawled source documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, each document d_j is segmented into a set of chunks:

$$\mathcal{P}_j = \{p_{j1}, p_{j2}, \ldots, p_{jm}\}.$$

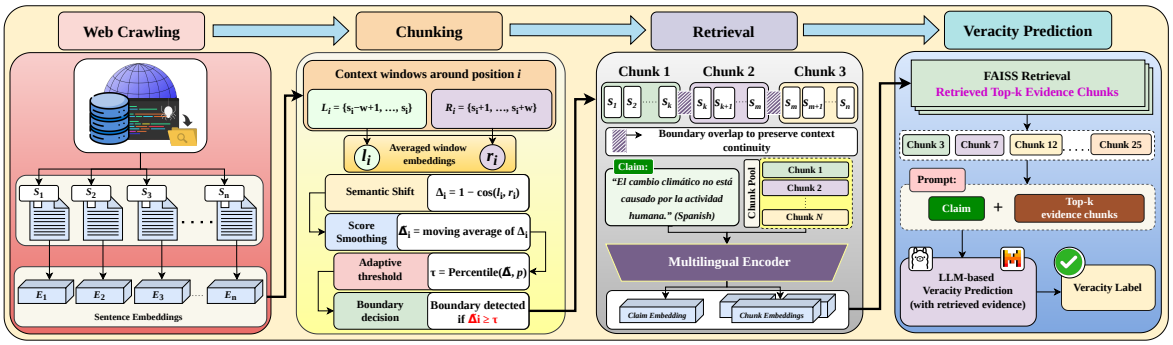

Figure 2: Overview of the SEEK and retrieval-based veracity prediction pipeline.

The claim and chunk embeddings are computed as:

$$\mathbf{h}_c = \text{Encoder}(c), \qquad \mathbf{h}_{p_{ji}} = \text{Encoder}(p_{ji}),$$

where both are produced using the same multilingual encoder. Following the instruction-tuned formulation of multilingual-e5, the claim is encoded with the retrieval instruction:

> *Instruct: Given a claim, retrieve relevant evidence from web documents that support or refute the claim.*

All chunks from the crawled documents are pooled into a global multilingual evidence space and indexed with FAISS (Douze et al., 2024). For each claim, cosine similarity is computed between the claim embedding and every candidate chunk:

$$s_{ji} = \cos(\mathbf{h}_c, \mathbf{h}_{p_{ji}}).$$

Instead of retrieving evidence separately from each document, we perform global retrieval across the complete chunk pool. We first retrieve the top-N candidates using FAISS, with $N = 20$:

$$\mathcal{R} = \text{TopN}\left(\cos(\mathbf{h}_c, \mathbf{h}_{p_{ji}})\right).$$

These candidates are then re-ranked using the same multilingual bi-encoder, and the final top-K chunks are selected, with $K = 5$:

$$\mathcal{E}^* = \text{TopK}(\mathcal{R}).$$

Here, $\mathcal{E}^*$ denotes the final evidence set passed to the downstream veracity prediction model. The selected evidence chunks are concatenated with the claim for veracity prediction. Unlike sentence-level retrieval, which may return isolated or incomplete evidence, SEEK retrieves semantically coherent chunks that preserve topic continuity and verification context while reducing irrelevant noise.

Instruction-tuning of large language models: After retrieval, the top-K ranked evidence chunks $\mathcal{E}^* = \{p_1^*, p_2^*, \ldots, p_K^*\}$ are used for downstream veracity prediction. For each claim, the input is formed by concatenating the task instruction, claim, and retrieved evidence:

$$\text{Input} = \mathcal{I} \oplus \mathcal{C} \oplus p_1^* \oplus p_2^* \oplus \cdots \oplus p_K^*,$$

where $\mathcal{I}$ denotes the dataset-specific instruction, $\mathcal{C}$ is the claim, and $\oplus$ denotes textual concatenation. The instructions for X-FACT and RU22FACT are provided in Appendix C. For X-FACT, evidence is retrieved from up to five associated web sources, while for RU22FACT, it is retrieved from the corresponding source article.

We formulate veracity prediction as an instruction-following causal language modeling task (Zhang et al., 2025), where the model generates the factuality label conditioned on the claim and evidence. We fine-tune multilingual LLMs, including **LLaMA**, **Gemma**, and **Mistral**, using **LoRA** (Hu et al., 2022). The base model is frozen, and only the low-rank adaptation parameters and output language modeling head are updated. Given the target label sequence $y_{1:T}$, the training loss is:

$$\mathcal{L} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, \text{Input}).$$

This trains the model to generate the correct veracity label using the claim and the retrieved evidence, where SEEK provides complete verification context for prediction.

5 Experimental Setup

This section describes the baselines, implementation settings, and evaluation protocols used to assess both veracity prediction performance and retrieval quality in multilingual fact-checking.

5.1 Baselines

We compare SEEK with several retrieval and evidence construction baselines.

Google Search Snippets: Following the original X-FACT setup, this baseline uses the provided Google Search snippets as evidence.

CONCRETE Retriever: We compare against CONCRETE (Huang et al., 2022), a multilingual dense retriever trained with the Cross-lingual Inverse Cloze Task (X-ICT). It retrieves evidence passages using multilingual dense similarity matching.

Sentence Chunking: This baseline splits each crawled webpage into consecutive sentence groups under a fixed token budget. This strategy is described in detail in Section 4.

Semantic Chunking: This baseline groups neighboring sentences into chunks based on embedding similarity. This strategy is also described in detail in Section 4.

LLM-Generated Evidence: For RU22FACT, we also compare with the LLM-generated evidence setting from the original RU22FACT work (Zeng et al., 2024), where evidence is generated by an LLM instead of retrieved from crawled web pages.

5.2 Implementation Details

We perform multilingual dense retrieval using **intfloat/multilingual-e5-large-instruct**, which encodes claims and evidence chunks into a shared multilingual semantic space. The resulting chunk embeddings are indexed with FAISS to enable efficient dense similarity search. After retrieval, the top-ranked evidence chunks are concatenated with the input claim and provided to instruction-tuned large language models for veracity prediction. We fine-tune **LLaMA**, **Gemma**, and **Mistral** using Low-Rank Adaptation (LoRA) under a causal language modeling objective. Further details on these models are provided in Appendix D. All experiments are implemented using **LLaMAFactory** (Zheng et al., 2024b) and conducted on a single NVIDIA A100 80 GB PCIe GPU.

5.3 Translation-based Retrieval Analysis

To study the effect of multilingual variation on retrieval, we translate all X-FACT and RU22FACT claims and crawled webpages into English before applying the same SEEK pipeline. The translated documents are chunked into semantically coherent passages, and retrieval is performed using **sentence-transformers/all-MiniLM-L6-v2** (Transformers, 2021). The retrieved chunks are then combined with the translated claim for veracity prediction. This setup tests whether retrieval improves when claim-evidence matching is performed in a unified English semantic space rather than a multilingual one.

Setting	Model	ID	OOD	ZS
Search Snippets	LLaMA	0.40 (±1.07)	0.26 (±0.25)	0.22 (±1.04)
	Gemma	0.43 (±1.63)	0.28 (±1.00)	0.20 (±1.36)
	Mistral	0.46 (±1.05)	0.30 (±0.45)	0.21 (±0.70)
CONCRETE	LLaMA	0.33 (±3.90)	0.19 (±0.18)	0.18 (±1.19)
	Gemma	0.35 (±1.60)	0.21 (±1.42)	0.15 (±0.54)
	Mistral	0.41 (±0.71)	0.22 (±0.46)	0.18 (±0.37)
Sentence Chunking	LLaMA	0.65 (±0.40)	0.31 (±1.31)	0.22 (±0.91)
	Gemma	0.61 (±0.30)	0.35 (±1.63)	0.23 (±1.16)
	Mistral	0.63 (±0.83)	0.37 (±0.95)	0.25 (±1.48)
Semantic Chunking	LLaMA	0.63 (±0.99)	0.33 (±0.65)	0.23 (±1.40)
	Gemma	0.61 (±0.63)	0.37 (±0.44)	0.23 (±1.17)
	Mistral	0.60 (±0.28)	0.37 (±0.37)	0.21 (±0.16)
SEEK	LLaMA	**0.67** (±0.51)	**0.41** (±1.48)	**0.30** (±0.62)
	Gemma	**0.64** (±2.54)	**0.39** (±0.28)	**0.24** (±0.40)
	Mistral	**0.66** (±0.55)	**0.39** (±0.57)	**0.27** (±1.79)

Table 2: Macro-F1 (mean ± std) on X-Fact across In-Domain (ID), Out-of-Domain (OOD), and Zero-Shot (ZS) settings.

5.4 Evaluation Protocol

For X-FACT, we report Macro-F1 under the original benchmark settings: in-domain (ID), out-of-domain (OOD), and zero-shot (ZS), with split details provided in Appendix A. For RU22FACT, we report Macro-F1 on the official test split. In addition to veracity prediction, we evaluate retrieval quality using claim-evidence semantic similarity, evidence completeness analysis, and qualitative retrieval analysis. To ensure fair comparison, we use the same fine-tuning, retrieval, decoding, and evaluation settings across all chunking strategies; detailed hyperparameters are provided in Appendix D.1.

6 Results and Discussion

We report veracity prediction results using Macro-F1 as the main evaluation metric. The results are summarized in Tables 2 and 3.

6.1 Findings on X-FACT

Table 2 reports Macro-F1 results on X-FACT across in-domain (ID), out-of-domain (OOD), and

Setting	Model	Macro-F1
LLM	LLaMA	0.73 (±0.47)
	Gemma	0.70 (±2.53)
	Mistral	0.72 (±0.71)
Sentence Chunking	LLaMA	0.71 (±0.92)
	Gemma	0.71 (±1.85)
	Mistral	0.72 (±1.02)
Semantic Chunking	LLaMA	0.79 (±0.54)
	Gemma	0.78 (±0.59)
	Mistral	0.79 (±0.67)
SEEK	LLaMA	**0.89** (±0.37)
	Gemma	**0.90** (±0.42)
	Mistral	**0.90** (±0.37)

Table 3: Macro-F1 on RU22FACT across four settings: LLM, sentence chunking, semantic chunking, and SEEK.

Models	ID	OOD	ZS
LLaMA	0.65 (±2.35)	**0.41** (±0.69)	0.36 (±3.54)
Gemma	0.65 (±2.12)	0.39 (±0.66)	**0.37** (±1.97)
Mistral	**0.66** (±1.06)	0.40 (±0.63)	0.37 (±1.02)

Table 4: Translation-based Macro-F1 performance of SEEK on X-FACT using different multilingual LLMs. Results are reported as Macro-F1 (mean ± std) across random seeds.

zero-shot (ZS) settings. SEEK consistently outperforms search-snippet evidence and the CONCRETE retriever, showing that veracity prediction benefits from more complete and semantically coherent evidence. The strongest results are obtained by **LLaMA with SEEK**, achieving **0.67** Macro-F1 in ID, **0.41** in OOD, and **0.30** in ZS. This corresponds to gains of up to **+0.21** over the best search-snippet baselines and up to **+0.26** over CONCRETE. These gains are especially important in OOD and ZS settings, where domain and language shifts make evidence retrieval more challenging. Across LLaMA, Gemma, and Mistral, SEEK yields stable improvements over sentence-level and standard semantic chunking baselines. This indicates that the gains arise mainly from improved evidence construction rather than a particular LLM backbone. Overall, coherent document-level chunks provide more reliable multilingual verification signals than snippets or isolated retrieval units.

6.2 Findings on RU22FACT

Table 3 reports Macro-F1 results on the official RU22FACT test split. The results show that evidence quality strongly affects factual verification across different retrieval settings. The LLM-only setting is competitive, with LLaMA reaching **0.73** and Mistral reaching **0.72**, while sentence chunking provides only marginal gains. Semantic chunking improves performance, with LLaMA and Mistral both reaching **0.79**. However, SEEK gives the strongest results, with Gemma and Mistral achieving **0.90** and LLaMA reaching **0.89**. This gives a gain of **+0.17** over the best LLM-only result and **+0.11** over the best semantic chunking result. Overall, these results show that SEEK consistently improves fact verification by providing more complete and coherent evidence, enabling different LLMs to make more reliable predictions.

Models	Macro-F1
LLaMA	0.90 (±0.76)
Gemma	0.89 (±1.34)
Mistral	0.90 (±1.04)

Table 5: Translation-based Macro-F1 performance of SEEK on RU22FACT.

6.3 Findings in Translation-based Retrieval

We further evaluate translation-based retrieval with SEEK, where claims and crawled documents are translated into English before retrieval and prediction. As shown in Table 4, this setting performs strongly on X-FACT, with the best Macro-F1 scores of **0.66** in ID, **0.41** in OOD, and **0.37** in ZS. On RU22FACT, Table 5 shows that all models achieve similarly high performance, with LLaMA and Mistral reaching **0.90** and Gemma reaching **0.89**. These results suggest that translation-based retrieval is complementary to SEEK: translation reduces linguistic variation, while SEEK preserves complete verification context.

6.4 Evaluation of Retrieval Quality

Beyond veracity prediction, we examine whether the retrieved evidence is useful for fact verification. Since gold evidence annotations are unavailable, we compare sentence chunking, semantic chunking, and SEEK using proxy analyses.

Claim-Evidence Semantic Similarity: We compute cosine similarity between each claim and its retrieved evidence in the multilingual embedding space. Higher scores indicate stronger claim–evidence alignment, and Figure 3 shows the resulting distributions across chunking strategies.

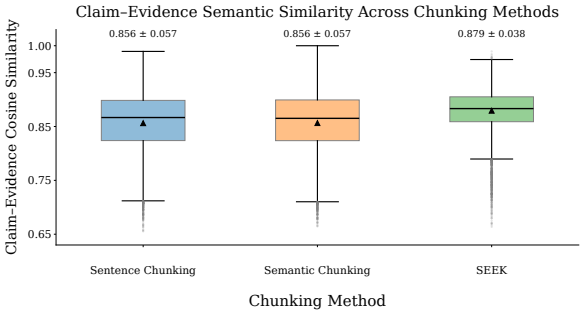

Figure 3: Claim-evidence semantic similarity across different chunking strategies.

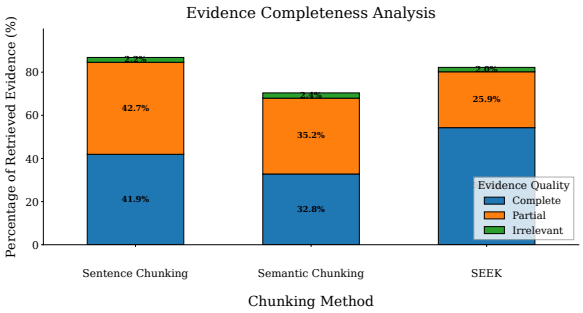

Figure 4: Evidence completeness analysis on X-FACT using an LLM-based evaluator.

Evidence Completeness Analysis: Because similarity alone does not ensure verification sufficiency, we use a locally deployed LLaMA-based evaluator to label each retrieved evidence instance as *Complete*, *Partial*, or *Irrelevant*. The prompt is given in Appendix E. As shown in Figures 4 and 5, SEEK retrieves more complete evidence than both baselines. Complete evidence reaches 54.3% on X-FACT, compared with 41.9% for sentence chunking and 32.8% for semantic chunking. On RU22FACT, SEEK achieves 76.0%, improving over sentence and semantic chunking by 29.0 and 15.3 percentage points, respectively, respectively. These results indicate that the proposed chunking strategy better preserves verification-ready context while reducing incomplete evidence retrieval.

Qualitative and Significance Analysis: We further provide a qualitative comparison in Figure 11, focusing on semantic continuity, investigation flow, and whether the retrieved evidence contains final supporting or refuting statements. Additional examples are included in Appendix B. McNemar's test further confirms that SEEK yields statistically significant gains over other methods, particularly on RU22FACT and X-FACT ID. Significance tests are reported in Appendix F.

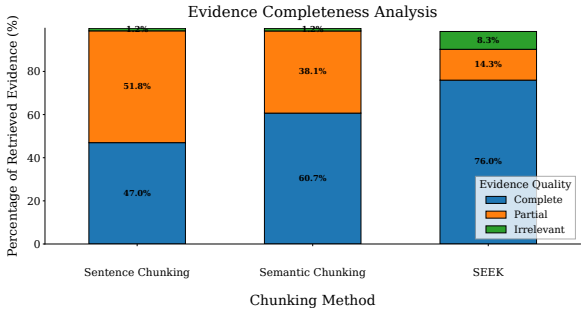

Figure 5: Evidence completeness analysis on RU22FACT using an LLM-based evaluator.

7 Conclusion

In this work, we proposed **SEEK**, a semantic evidence extraction framework for multilingual fact-checking. SEEK moves beyond short snippets, isolated sentences, and fixed-size chunks by building coherent evidence passages from full web documents. It detects topic shifts and keeps boundary context, allowing the retrieved evidence to better preserve the claim, supporting details, and final verification signal. Experiments on X-FACT and RU22FACT show that SEEK improves veracity prediction across in-domain, out-of-domain, and zero-shot settings. The retrieval analysis further shows that SEEK retrieves evidence that is not only semantically relevant, but also more complete for verification. Overall, our findings highlight that reliable multilingual fact-checking depends strongly on evidence construction. By providing focused and context-rich evidence chunks, SEEK helps downstream LLMs make more reliable factuality predictions and reduces the limitations of retrieval based on snippets, sentences, and standard semantic chunks.

8 Limitations

SEEK depends on the quality of crawled web documents; incomplete or noisy pages can still lead to missing evidence. Since the method detects semantic topic shifts, it may also miss cases where verification cues are spread across distant document sections. Our experiments are limited to X-FACT and RU22FACT, so evaluation on more languages, domains, and real-time fact-checking settings is needed. Finally, human-annotated evidence completeness labels would provide a stronger evaluation signal, which we leave for future work.

References

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.

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2024. Cross-lingual learning vs. low-resource fine-tuning: A case study with fact-checking in Turkish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4127–4142, Torino, Italia. ELRA and ICCL.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. GERE: Generative evidence retrieval for fact verification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2184–2189, New York, NY, USA. Association for Computing Machinery.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. arXiv preprint.

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

Ashim Gupta and Vivek Srikumar. 2021a. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Ashim Gupta and Vivek Srikumar. 2021b. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR 2022*.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

C. Kiss. 2025. Max–min semantic chunking of documents for rag application. *Discover Computing*, 28(117).

LangChain. 2024. Langchain semantic chunking. `https://python.langchain.com/docs/how_to/semantic_chunker/`. Accessed: 2024-07-04.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Kevin Nanhekhan, V. Venktesh, Erik Martin, Henrik Vatndal, Vinay Setty, and Avishek Anand. 2025. Flashcheck: Exploration of efficient evidence retrieval for fast fact-checking. In *Advances in Information Retrieval*, volume 15575 of *Lecture Notes in Computer Science*, pages 385–399, Cham. Springer.

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.

Q. Peng et al. 2025a. Semeval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podrouzek, Matúš Mesarčík, Jaroslav Kopčan, and Anders Søgaard. 2025b. SemEval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2498–2511, Vienna, Austria. Association for Computational Linguistics.

R. Qu et al. 2025. Is semantic chunking worth the computational cost? *Findings of the Association for Computational Linguistics: NAACL 2025*.

Dorian Quelle, Calvin Yixiang Cheng, Alexandre Bovet, and Scott A. Hale. 2025. Lost in translation: using

global fact-checks to measure multilingual misinformation prevalence, spread, and evolution. *EPJ Data Science*, 14(1):22.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. Updated version of NAACL2018 paper. https://doi.org/10.48550/arXiv.1803.05355.

Sentence Transformers. 2021. sentence-transformers/all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2026-05-17.

UncleCode. 2024. Crawl4ai: Open-source llm friendly web crawler & scraper. https://github.com/unclecode/crawl4ai. GitHub repository. Please use the commit hash you're working with.

V. Viswanathan et al. 2025. Claimiq at checkthat! 2025: Comparing prompted and fine-tuned language models for verifying numerical claims.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

William Y. Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 422–426. Association for Computational Linguistics.

Yirong Zeng, Xiao Ding, Yi Zhao, Xiangyu Li, Jie Zhang, Chao Yao, Ting Liu, and Bing Qin. 2024. Ru22fact: Optimizing evidence for multilingual explainable fact-checking on russia-ukraine conflict. *arXiv preprint arXiv:2403.16662*.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. From relevance to utility: Evidence retrieval with feedback for fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6373–6384, Singapore. Association for Computational Linguistics.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2025. Instruction tuning for large language models: A survey.

Liwen Zheng, Chaozhuo Li, Xi Zhang, Yu-Ming Shang, Feiran Huang, and Haoran Jia. 2024a. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9274–9281, Bangkok, Thailand. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Additional Details on Datasets

This appendix provides additional details on the datasets used in our study, including their label distributions and evaluation splits.

A.1 X-FACT

X-FACT is a multilingual fact-checking dataset containing 31,189 claims collected from 85 fact-checking websites across 25 languages and 11 language families (Gupta and Srikumar, 2021b). Each claim is annotated with a factuality label and includes metadata such as language, source URL, claim date, and review date.

The dataset contains seven labels: *True, Mostly True, Partially True/Misleading, Mostly False, False, Complicated/Hard to categorize*, and *Other*. As shown in Table 6.

For evaluation, X-FACT provides three test settings: *In-domain*, where claims come from sources and languages seen during training; *Out-of-domain*, where languages are seen but sources are unseen; and *Zero-shot*, where both languages and sources are unseen during training. Table 7 shows the language-wise distribution, highlighting that several languages appear only in the zero-shot split, making this setting useful for evaluating cross-lingual generalization.

Class	Train	Dev	Indomain	OOD	Zeroshot
False	7,515	977	1,464	1,109	1,471
PT/M	4,359	553	858	769	772
True	4,080	551	851	329	332
MT	1,380	194	300	0	73
MF	848	114	159	0	415
C/H	540	80	106	150	264
Other	357	66	88	11	54
Total	**19,079**	**2,535**	**3,826**	**2,368**	**3,381**

Table 6: Label-wise distribution of the X-FACT dataset across training, development, in-domain, out-of-domain, and zero-shot evaluation splits. PT/M = Partially True/Misleading, MT = Mostly True, MF = Mostly False, and C/H = Complicated/Hard to categorize.

Language	Indomain	Dev	OOD	Train	Zeroshot	Total
Portuguese	1121	747	472	5601	0	7941
Indonesian	448	297	647	2231	0	3623
Arabic	314	209	0	1567	0	2090
Georgian	307	203	0	1529	0	2039
Polish	267	176	0	1325	0	1768
Turkish	169	111	613	858	0	1751
Italian	190	125	255	943	0	1513
Hindi	170	112	381	845	0	1508
Tamil	221	145	0	1097	0	1463
Spanish	211	139	0	1049	0	1399
German	142	95	0	712	0	949
Romanian	140	93	0	698	0	931
Serbian	126	83	0	624	0	833
Albanian	0	0	0	0	666	666
Bengali	0	0	0	0	568	568
Russian	0	0	0	0	442	442
Norwegian	0	0	0	0	406	406
Persian	0	0	0	0	281	281
Azerbaijani	0	0	0	0	253	253
French	0	0	0	0	198	198
Dutch	0	0	0	0	166	166
Gujarati	0	0	0	0	129	129
Punjabi	0	0	0	0	105	105
Marathi	0	0	0	0	87	87
Sinhala	0	0	0	0	80	80
Total	**3826**	**2535**	**2368**	**19079**	**3381**	**31189**

Table 7: Language-wise distribution of the X-FACT dataset across evaluation splits. Languages appearing only in the zero-shot split are unseen during training and are used to evaluate cross-lingual transfer.

A.2 RU22FACT

RU22FACT is a multilingual explainable fact-checking dataset focused on the 2022 Russia–Ukraine conflict (Zeng et al., 2024). It contains 16,033 claims in four languages: English, Chinese, Russian, and Ukrainian. Each sample includes a claim, optimized evidence, a reference explanation, metadata, and one of three labels: *Supported*, *Refuted*, or *Not Enough Information* (NEI).

Table 8 shows the label distribution, where *Supported* is the largest class. Table 9 reports the language-wise distribution, with English forming the largest portion of the dataset.

Class	Train	Dev	Test	Total
Supported	6,836	1,079	2,166	10,081
Refuted	3,299	450	902	4,651
NEI	1,082	71	148	1,301
Total	**11,217**	**1,600**	**3,216**	**16,033**

Table 8: Label-wise distribution of the RU22FACT dataset across training, development, and test splits. NEI denotes Not Enough Information.

Language	Train	Dev	Test	Total
English	6,082	867	1,741	8,690
Chinese	1,055	152	305	1,512
Russian	1,621	231	465	3,399
Ukrainian	2,458	350	704	3,512
Total	**11,217**	**1,600**	**3,216**	**16,033**

Table 9: Language-wise distribution of the RU22FACT dataset across training, development, and test splits.

B Qualitative Analysis of Chunk Formation

To qualitatively illustrate the proposed Semantic Evidence Extraction with adaptive chunKing (SEEK) method, we use a real Hindi fact-checking example consisting of a viral claim and its corresponding source article, as shown in Figure 7. The claim states that the woman inspector in the viral image is posted in the same area where her father works as a rickshaw driver. The source article investigates this claim and concludes that the viral story is misleading. Figure 6 shows the raw semantic shift scores, their smoothed version, and the adaptive threshold used by SEEK to identify meaningful topic boundaries.

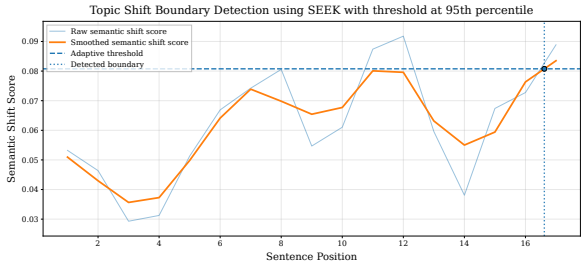

Figure 6: Example of topic-shift boundary detection using SEEK on a Hindi fact-checking document. The smoothed semantic shift curve reduces local sentence-level noise, while the adaptive threshold selects meaningful chunk boundaries.

Figure 8 shows the chunks produced by the proposed SEEK method. The first chunk preserves the complete fact-checking flow, including the viral claim, evidence investigation, corrective clarification, and final refutation, while the second chunk mainly contains trailing unrelated webpage content. In contrast, Figure 9 shows that semantic chunking splits the verification context across two chunks: the first chunk contains the claim and early investigation, whereas the second chunk contains the final clarification but also mixes it with unrelated navigation content. Similarly, Figure 10 shows

Example Hindi Claim and Source Document

Claim

वायरल तस्वीर में दिख रही महिला दरोगा उसी क्षेत्र में तैनात हैं जहाँ उनके पिता रिक्शा चलाते हैं।

Source Document

फैक्ट चेक: रिक्शा चालक की बेटी के इंस्पेक्टर बनने की मार्मिक कहानी की ये है सच्चाई\n सोशल मीडिया पर एक महिला इंस्पेक्टर की तस्वीर वायरल हो रही है, जिसके साथ दावा किया जा रहा है कि यह महिला उसी इलाके में दरोगा नियुक्त की गई है, जहां उनके पिता रिक्शा चलाते हैं. इस दावे में कितनी सच्चाई है....जानने के लिए पढ़िए पूरी खबर.\n वायरल तस्वीर में दिख रही महिला दरोगा उसी क्षेत्र में तैनात हैं जहां उनके पिता रिक्शा चलाते हैं.\n वायरल तस्वीर इंस्पेक्टर प्रियंका नेगी की है, जो बिलासपुर में अपने घर से 200 किलोमीटर दूर तैनात हैं. उनके पिता रिक्शा नहीं चलाते, बल्कि उनके परिवार के पास कई दुकानें हैं.\n * (अपडेटेड 08 जुलाई 2020, 7:01 PM IST)\n\n\nसोशल मीडिया पर एक महिला पुलिस अधिकारी की तस्वीर वायरल हो रही है. तस्वीर में दिख रहा है कि पुलिस यूनिफॉर्म में एक महिला घर के दरवाजे पर खड़ी मुस्कुरा रही है. तस्वीर के साथ दावा किया जा रहा है कि यह महिला उसी इलाके में दरोगा नियुक्त की गई है, जहां उनके पिता रिक्शा चलाते हैं. \nइंडिया टुडे एंटी फेक न्यूज़ वॉर रूम (AFWA) ने पाया कि जिन महिला पुलिस अधिकारी की तस्वीर वायरल है, वे हिमाचल प्रदेश के बिलासपुर पुलिस थाने में तैनात हैं, लेकिन उनकी इस तस्वीर के साथ जोड़कर प्रसारित हो रही मार्मिक कहानी में कोई सच्चाई नहीं है.\nफेसबुक यूजर ने एक महिला पुलिस अधिकारी की फोटो पोस्ट करते हुए लिखा, \"रिक्शा चालक की बेटी उसी क्षेत्र की दरोगा बनी, जिस क्षेत्र में उसके पिता रिक्शा चलाते हैं. बधाई तो बनती है.\" इस पोस्ट के जवाब में कई लोगों ने बधाइयां दी हैं. इस पोस्ट का आर्काइव्ड वर्जन देखा जा सकता है.\nफेसबुक पर हो रही है.\n**इंस्पेक्टर प्रियंका नेगी की है तस्वीर**\nसबसे पहले हमने फोटो को रिवर्स सर्च किया और फेसबुक पर के पेज पर 8 मई को यही फोटो अपलोड की गई थी.\nप्रियंका नेगी के बारे में हमने इंटरनेट पर सर्च किया तो सामने आईं. इन खबरों के मुताबिक, प्रियंका नेगी हिमाचल प्रदेश महिला कबड्डी टीम की कैप्टन रह चुकी हैं. नेगी भारतीय कबड्डी टीम का हिस्सा भी रह चुकी हैं, जिसने 2018 के एशियन गेम में दूसरा स्थान जीता था. बतौर इंस्पेक्टर वे हिमाचल के सदर बिलासपुर पुलिस थाने में नियुक्त हैं. हालांकि, नेगी से जुड़ी किसी भी खबर में उनके पिता के रिक्शा चालक होने का ज़िक्र नहीं है.\nनेगी सिरमौर की रहने वाली हैं जो कि बिलासपुर से 200 किलोमीटर दूर स्थित है. हमने नेगी से संपर्क किया तो उन्होंने बताया कि उन्हें पता है कि उनका फोटो गलत दावों के साथ वायरल किया जा रहा है. \nउन्होंने कहा, \"मैंने देखा था और मुझे बहुत बुरा भी लगा कि मेरे परिवार के बारे में लोग ऐसे लिख रहे हैं. फोटो तो मेरी है, लेकिन उसके साथ मेरे पापा के रिक्शा चलाने का जो दावा किया गया है, वह सही नहीं है. मेरे परिवार के पास कई दुकाने हैं और यही हमारा खानदानी बिजनेस रहा है.\"\nकुल मिलाकर यह कहा जा सकता है कि सोशल मीडिया पर जिस तरीके से इंस्पेक्टर प्रियंका नेगी के बारे में लिखा जा रहा है, वह पूरी तरह से सही नहीं है. नेगी एक पुलिस इंस्पेक्टर जरूर हैं, लेकिन वे अपने घर सिरमौर से काफी दूर, बिलासपुर में पोस्टेड हैं. नेगी कबड्डी में भी देश का नाम ऊंचा कर चुकी हैं.\nक्या आपको लगता है कोई मैसेज झूठा ? सच जानने के लिए उसे हमारे नंबर पर भेजें.आप हमें पर ईमेल भी कर सकते हैं \n [सबसे तेज़ ख़बरों के लिए आजतक ऐप] [मंच पर महाभारत... केंद्र डांस फेस्टिवल में नृत्य नाटिका कर्ण का मार्मिक मंचन, टुल्लू मुर्मू के अभिनय ने जीता दिल]

Key Finding: The viral claim is false/misleading. The officer is Priyanka Negi, but her father is not a rickshaw driver and she is posted far from her home region.

Figure 7: Example Hindi claim and source document.

that sentence chunking fragments the fact-checking flow into separate sentence-budgeted chunks, pushing the decisive refuting evidence into the second chunk and reducing evidence completeness within a single retrieved chunk. This example demonstrates that SEEK better separates the core verification evidence from less relevant trailing content.

C Dataset-Specific Instructions

We use dataset-specific task instructions to guide the language model for veracity prediction. The instruction is prepended to the claim and retrieved evidence chunks, and the exact prompts used for X-FACT and RU22FACT are reported in Table 10.

D Model Details

Table 11 summarizes the encoder and decoder models used in our framework. The encoder models are used for multilingual claim–evidence representation and retrieval, while the instruction-tuned large language models are used for downstream veracity prediction.

We use multilingual-e5-large as the retrieval encoder because our task involves fact-checking across multiple languages. This encoder provides a shared multilingual representation space, allowing claims and evidence chunks within each language to be encoded consistently for dense retrieval. FAISS is used to perform efficient nearest-neighbour search over the global chunk pool.

For veracity prediction, we evaluate three

Example Chunks from SEEK

Chunk	Actual Chunk Excerpt	Actual Chunk in Original Language
Chunk 1	This chunk captures the complete fact-checking flow. It first states the viral social-media claim that a woman police inspector is posted in the same area where her father allegedly works as a rickshaw driver. It then presents the fact-checking investigation by AFWA, including reverse image search, identification of the officer as Inspector Priyanka Negi, and background details about her posting in Bilaspur. The chunk also includes direct clarification from Priyanka Negi, where she confirms that the image is hers but rejects the claim about her father being a rickshaw driver. Thus, this chunk preserves the main claim, supporting investigation, corrective evidence, and final refuting conclusion.	# फैक्ट चेक: रिक्शा चालक की बेटी के इंस्पेक्टर बनने की मार्मिक कहानी की ये है सच्चाई\n सोशल मीडिया पर एक महिला इंस्पेक्टर की तस्वीर वायरल हो रही है, जिसके साथ दावा किया जा रहा है कि यह महिला उसी इलाके में दरोगा नियुक्त की गई है, जहां उनके पिता रिक्शा चलाते हैं. इस दावे में कितनी सच्चाई है....जानने के लिए पढ़िए पूरी खबर.\n वायरल तस्वीर में दिख रही महिला दरोगा उसी क्षेत्र में तैनात हैं जहां उनके पिता रिक्शा चलाते हैं.\n वायरल तस्वीर इंस्पेक्टर प्रियंका नेगी की है, जो बिलासपुर में अपने घर से 200 किलोमीटर दूर तैनात हैं. उनके पिता रिक्शा नहीं चलाते, बल्कि उनके परिवार के पास कई दुकानें हैं.\n * (अपडेटेड 08 जुलाई 2020, 7:01 PM IST)\n\n\n\nसोशल मीडिया पर एक महिला पुलिस अधिकारी की तस्वीर वायरल हो रही है. तस्वीर में दिख रहा है कि पुलिस यूनिफॉर्म में एक महिला घर के दरवाजे पर खड़ी मुस्कुरा रही है. तस्वीर के साथ दावा किया जा रहा है कि यह महिला उसी इलाके में दरोगा नियुक्त की गई है, जहां उनके पिता रिक्शा चलाते हैं. \nइंडिया टुडे एंटी फेक न्यूज़ वॉर रूम (AFWA) ने पाया कि जिन महिला पुलिस अधिकारी की तस्वीर वायरल है, वे हिमाचल प्रदेश के बिलासपुर पुलिस थाने में तैनात हैं, लेकिन उनकी इस तस्वीर के साथ जोड़कर प्रसारित हो रही मार्मिक कहानी में कोई सच्चाई नहीं है.\nफेसबुक यूजर ने एक महिला पुलिस अधिकारी की फोटो पोस्ट करते हुए लिखा, \"रिक्शा चालक की बेटी उसी क्षेत्र की दरोगा बनी, जिस क्षेत्र में उसके पिता रिक्शा चलाते हैं. बधाई तो बनती है.\" इस पोस्ट के जवाब में कई लोगों ने बधाइयां दी हैं. इस पोस्ट का आर्काइव्ड वर्जन देखा जा सकता है.\nफेसबुक पर हो रही है.\n**इंस्पेक्टर प्रियंका नेगी की है तस्वीर**\nसबसे पहले हमने फोटो को रिवर्स सर्च किया और फेसबुक पर के पेज पर 8 मई को यही फोटो अपलोड की गई थी.\nप्रियंका नेगी के बारे में हमने इंटरनेट पर सर्च किया तो सामने आईं. इन खबरों के मुताबिक, प्रियंका नेगी हिमाचल प्रदेश महिला कबड्डी टीम की कैप्टन रह चुकी हैं. नेगी भारतीय कबड्डी टीम का हिस्सा भी रह चुकी हैं, जिसने 2018 के एशियन गेम में दूसरा स्थान जीता था. बतौर इंस्पेक्टर वे हिमाचल के सदर बिलासपुर पुलिस थाने में नियुक्त हैं. हालांकि, नेगी से जुड़ी किसी भी खबर में उनके पिता के रिक्शा चालक होने का ज़िक्र नहीं है.\nनेगी सिरमौर की रहने वाली हैं जो कि बिलासपुर से 200 किलोमीटर दूर स्थित है. हमने नेगी से संपर्क किया तो उन्होंने बताया कि उन्हें पता है कि उनका फोटो गलत दावों के साथ वायरल किया जा रहा है. \nउन्होंने कहा, \"मैंने देखा था और मुझे बहुत बुरा भी लगा कि मेरे परिवार के बारे में लोग ऐसे लिख रहे हैं. फोटो तो मेरी है, लेकिन उसके साथ मेरे पापा के रिक्शा चलाने का जो दावा किया गया है, वह सही नहीं है. मेरे परिवार के पास कई दुकाने हैं और यही हमारा खानदानी बिजनेस रहा है.\"\nकुल मिलाकर यह कहा जा सकता है कि सोशल मीडिया पर जिस तरीके से इंस्पेक्टर प्रियंका नेगी के बारे में लिखा जा रहा है, वह पूरी तरह से सही नहीं है. नेगी एक पुलिस इंस्पेक्टर जरूर हैं, लेकिन वे अपने घर सिरमौर से काफी दूर, बिलासपुर में पोस्टेड हैं.
Chunk 2	The chunk mainly contains trailing webpage content, including a statement about Negi's kabaddi achievements and unrelated article/navigation text.	नेगी कबड्डी में भी देश का नाम ऊंचा कर चुकी हैं.\nक्या आपको लगता है कोई मैसेज झूठा ? सच जानने के लिए उसे हमारे नंबर पर भेजें.आप हमें पर ईमेल भी कर सकते हैं \n [सबसे तेज़ ख़बरों के लिए आजतक ऐप] [मंच पर महाभारत... केंद्र डांस फेस्टिवल में नृत्य नाटिका कर्ण का मार्मिक मंचन, दुल्लू मुर्मू के अभिनय ने जीता दिल]

Figure 8: Example chunks produced by the proposed Semantic Evidence Extraction with Adaptive chunKing(SEEK) method on a Hindi fact-checking document. The figure reports English summaries and original Hindi chunks

instruction-tuned LLM families: LLaMA, Gemma, and Mistral. These models are selected because they are strong open-source multilingual instruction-following models with different architectures and training strategies. Evaluating multiple model families allows us to examine whether the proposed chunking strategy consistently improves downstream fact verification performance, rather than being effective only for a single model.

D.1 Hyperparameters

All veracity prediction models are fine-tuned for 3 epochs using LoRA adaptation. Unless otherwise specified, the LoRA rank is set to 8. The same retrieval configuration, decoding settings, and evaluation protocol are used across all experiments to ensure a fair comparison between different chunking strategies.

E Prompt for Evidence Completeness Evaluation

We use the prompt shown in Table 13 to evaluate whether the retrieved evidence contains sufficient verification context.

F Significance Test Results

We report all McNemar significance test results in this appendix to assess whether the gains of SEEK over different evidence construction baselines are statistically reliable. Table 14 compares SEEK with semantic chunking, Table 15 compares it with sentence chunking, Table 16 compares it with the CONCRETE baseline, Table 17 compares it with Google Search snippets, and Table 18 compares it with the LLM-only baseline. Across all tables, B>O denotes examples correctly predicted only by the baseline, whereas O>B denotes examples correctly predicted only by SEEK. Net Gain is computed as O>B minus B>O, and statistical significance is reported at $\alpha = 0.05$.

Example Chunks from Semantic Chunking

Chunk	Actual Chunk Excerpt	Actual Chunk in Original Language
Chunk 1	This chunk contains the beginning of the fact-checking article. It introduces the viral claim that a woman inspector is posted in the same area where her father allegedly works as a rickshaw driver. It also gives the initial background of the viral social-media post and mentions that the image was investigated through reverse image search. However, this chunk stops before the full verification is completed. It includes the claim and early investigation details, but it does not contain the final clarification from Priyanka Negi or the complete refutation of the viral claim.	# फैक्ट चेक: रिक्शा चालक की बेटी के इंस्पेक्टर बनने की मार्मिक कहानी की ये है सच्चाई\n सोशल मीडिया पर एक महिला इंस्पेक्टर की तस्वीर वायरल हो रही है, जिसके साथ दावा किया जा रहा है कि यह महिला उसी इलाके में दरोगा नियुक्त की गई है, जहां उनके पिता रिक्शा चलाते हैं. इस दावे में कितनी सच्चाई है....जानने के लिए पढ़िए पूरी खबर.\n वायरल तस्वीर में दिख रही महिला दरोगा उसी क्षेत्र में तैनात हैं जहां उनके पिता रिक्शा चलाते हैं.\n वायरल तस्वीर इंस्पेक्टर प्रियंका नेगी की है, जो बिलासपुर में अपने घर से 200 किलोमीटर दूर तैनात हैं. उनके पिता रिक्शा नहीं चलाते, बल्कि उनके परिवार के पास कई दुकानें हैं.\n * (अपडेटेड 08 जुलाई 2020, 7:01 PM IST)\n\n\nसोशल मीडिया पर एक महिला पुलिस अधिकारी की तस्वीर वायरल हो रही है. तस्वीर में दिख रहा है कि पुलिस यूनिफॉर्म में एक महिला घर के दरवाजे पर खड़ी मुस्कुरा रही है. तस्वीर के साथ दावा किया जा रहा है कि यह महिला उसी इलाके में दरोगा नियुक्त की गई है, जहां उनके पिता रिक्शा चलाते हैं. \nइंडिया टुडे एंटी फेक न्यूज़ वॉर रूम (AFWA) ने पाया कि जिन महिला पुलिस अधिकारी की तस्वीर वायरल है, वे हिमाचल प्रदेश के बिलासपुर पुलिस थाने में तैनात हैं, लेकिन उनकी इस तस्वीर के साथ जोड़कर प्रसारित हो रही मार्मिक कहानी में कोई सच्चाई नहीं है.\nफेसबुक यूजर ने एक महिला पुलिस अधिकारी की फोटो पोस्ट करते हुए लिखा, \"रिक्शा चालक की बेटी उसी क्षेत्र की दरोगा बनी, जिस क्षेत्र में उसके पिता रिक्शा चलाते हैं.
Chunk 2	This chunk continues the remaining part of the fact-checking article. It includes the corrective evidence, where Priyanka Negi clarifies that the woman in the image is her, but the claim about her father being a rickshaw driver is false. It also explains her actual posting details and confirms that her father is not associated with the viral claim. However, this chunk also includes trailing unrelated webpage content, such as navigation text and a separate statement about Negi's kabaddi achievements. Thus, semantic chunking preserves the later verification context but mixes it with extra irrelevant content.	बधाई तो बनती है.\" इस पोस्ट के जवाब में कई लोगों ने बधाइयां दी हैं. इस पोस्ट का आर्काइव्ड वर्जन देखा जा सकता है.\nफेसबुक पर हो रही है.\n**इंस्पेक्टर प्रियंका नेगी की है तस्वीर**\nसबसे पहले हमने फोटो को रिवर्स सर्च किया और फेसबुक पर के पेज पर 8 मई को यही फोटो अपलोड की गई थी.\nप्रियंका नेगी के बारे में हमने इंटरनेट पर सर्च किया तो सामने आई. इन खबरों के मुताबिक, प्रियंका नेगी हिमाचल प्रदेश महिला कबड्डी टीम की कैप्टन रह चुकी हैं. नेगी भारतीय कबड्डी टीम का हिस्सा भी रह चुकी हैं, जिसने 2018 के एशियन गेम में दूसरा स्थान जीता था. बतौर इंस्पेक्टर वे हिमाचल के सदर बिलासपुर पुलिस थाने में नियुक्त हैं. हालांकि, नेगी से जुड़ी किसी भी खबर में उनके पिता के रिक्शा चालक होने का ज़िक्र नहीं है.\nनेगी सिरमौर की रहने वाली हैं जो कि बिलासपुर से 200 किलोमीटर दूर स्थित है. हमने नेगी से संपर्क किया तो उन्होंने बताया कि उन्हें पता है कि उनका फोटो गलत दावों के साथ वायरल किया जा रहा है. \nउन्होंने कहा, \"मैंने देखा था और मुझे बहुत बुरा भी लगा कि मेरे परिवार के बारे में लोग ऐसे लिख रहे हैं. फोटो तो मेरी है, लेकिन उसके साथ मेरे पापा के रिक्शा चलाने का जो दावा किया गया है, वह सही नहीं है. मेरे परिवार के पास कई दुकाने हैं और यही हमारा खानदानी बिजनेस रहा है.\"\nकुल मिलाकर यह कहा जा सकता है कि सोशल मीडिया पर जिस तरीके से इंस्पेक्टर प्रियंका नेगी के बारे में लिखा जा रहा है, वह पूरी तरह से सही नहीं है. नेगी एक पुलिस इंस्पेक्टर जरूर हैं, लेकिन वे अपने घर सिरमौर से काफी दूर, बिलासपुर में पोस्टेड हैं. नेगी कबड्डी में भी देश का नाम ऊंचा कर चुकी हैं.\nक्या आपको लगता है कोई मैसेज झूठा ? सच जानने के लिए उसे हमारे नंबर पर भेजें.आप हमें पर ईमेल भी कर सकते हैं \n [सबसे तेज़ ख़बरों के लिए आजतक ऐप] [मंच पर महाभारत... केंद्र डांस फेस्टिवल में नृत्य नाटिका कर्ण का मार्मिक मंचन, टुल्लू मुर्मू के अभिनय ने जीता दिल]

Figure 9: Example chunks produced by the semantic chunking method on a Hindi fact-checking document. The figure reports English summaries and original Hindi chunks

Example Chunks from Sentence Chunking

Chunk	Actual Chunk Excerpt	Actual Chunk in Original Language
Chunk 1	This chunk contains the opening part of the fact-checking article. It introduces the viral claim that the woman inspector is posted in the same area where her father allegedly works as a rickshaw driver. It also includes the initial verification steps, such as reverse image search and background information about the viral post. However, this chunk ends before the full clarification is completed. It captures the claim and early investigation, but the final refuting evidence is pushed into the next chunk.	# फैक्ट चेक: रिक्शा चालक की बेटी के इंस्पेक्टर बनने की मार्मिक कहानी की ये है सच्चाई\n सोशल मीडिया पर एक महिला इंस्पेक्टर की तस्वीर वायरल हो रही है, जिसके साथ दावा किया जा रहा है कि यह महिला उसी इलाके में दरोगा नियुक्त की गई है, जहां उनके पिता रिक्शा चलाते हैं. इस दावे में कितनी सच्चाई है....जानने के लिए पढ़िए पूरी खबर.\n वायरल तस्वीर में दिख रही महिला दरोगा उसी क्षेत्र में तैनात हैं जहां उनके पिता रिक्शा चलाते हैं.\n वायरल तस्वीर इंस्पेक्टर प्रियंका नेगी की है, जो बिलासपुर में अपने घर से 200 किलोमीटर दूर तैनात हैं. उनके पिता रिक्शा नहीं चलाते, बल्कि उनके परिवार के पास कई दुकानें हैं.\n * (अपडेटेड 08 जुलाई 2020, 7:01 PM IST)\n\n\nसोशल मीडिया पर एक महिला पुलिस अधिकारी की तस्वीर वायरल हो रही है. तस्वीर में दिख रहा है कि पुलिस यूनिफॉर्म में एक महिला घर के दरवाजे पर खड़ी मुस्कुरा रही है. तस्वीर के साथ दावा किया जा रहा है कि यह महिला उसी इलाके में दरोगा नियुक्त की गई है, जहां उनके पिता रिक्शा चलाते हैं. \nइंडिया टुडे एंटी फेक न्यूज़ वॉर रूम (AFWA) ने पाया कि जिन महिला पुलिस अधिकारी की तस्वीर वायरल है, वे हिमाचल प्रदेश के बिलासपुर पुलिस थाने में तैनात हैं, लेकिन उनकी इस तस्वीर के साथ जोड़कर प्रसारित हो रही मार्मिक कहानी में कोई सच्चाई नहीं है.\nफेसबुक यूजर ने एक महिला पुलिस अधिकारी की फोटो पोस्ट करते हुए लिखा, \"रिक्शा चालक की बेटी उसी क्षेत्र की दरोगा बनी, जिस क्षेत्र में उसके पिता रिक्शा चलाते हैं. बधाई तो बनती है.\" इस पोस्ट के जवाब में कई लोगों ने बधाइयां दी हैं. इस पोस्ट का आर्काइव्ड वर्जन देखा जा सकता है.\nफेसबुक पर हो रही है.\n**इंस्पेक्टर प्रियंका नेगी की है तस्वीर**\nसबसे पहले हमने फोटो को रिवर्स सर्च किया और फेसबुक पर के पेज पर 8 मई को यही फोटो अपलोड की गई थी.\nप्रियंका नेगी के बारे में हमने इंटरनेट पर सर्च किया तो सामने आई.
Chunk 2	This chunk contains the later verification details. It includes Priyanka Negi's clarification that the image is hers, but the claim about her father being a rickshaw driver is false. It also mentions her actual posting details and confirms that the viral story is misleading. However, sentence chunking splits the fact-checking flow across two chunks and also includes some unrelated trailing webpage content. Thus, the verification context is fragmented rather than preserved in one coherent chunk.	इन खबरों के मुताबिक, प्रियंका नेगी हिमाचल प्रदेश महिला कबड्डी टीम की कैप्टन रह चुकी हैं. नेगी भारतीय कबड्डी टीम का हिस्सा भी रह चुकी हैं, जिसने 2018 के एशियन गेम में दूसरा स्थान जीता था. बतौर इंस्पेक्टर वे हिमाचल के सदर बिलासपुर पुलिस थाने में नियुक्त हैं. हालांकि, नेगी से जुड़ी किसी भी खबर में उनके पिता के रिक्शा चालक होने का ज़िक्र नहीं है.\nनेगी सिरमौर की रहने वाली हैं जो कि बिलासपुर से 200 किलोमीटर दूर स्थित है. हमने नेगी से संपर्क किया तो उन्होंने बताया कि उन्हें पता है कि उनका फोटो गलत दावों के साथ वायरल किया जा रहा है. \nउन्होंने कहा, \"मैंने देखा था और मुझे बहुत बुरा भी लगा कि मेरे परिवार के बारे में लोग ऐसे लिख रहे हैं. फोटो तो मेरी है, लेकिन उसके साथ मेरे पापा के रिक्शा चलाने का जो दावा किया गया है, वह सही नहीं है. मेरे परिवार के पास कई दुकाने हैं और यही हमारा खानदानी बिजनेस रहा है.\"\nकुल मिलाकर यह कहा जा सकता है कि सोशल मीडिया पर जिस तरीके से इंस्पेक्टर प्रियंका नेगी के बारे में लिखा जा रहा है, वह पूरी तरह से सही नहीं है. नेगी एक पुलिस इंस्पेक्टर जरूर हैं, लेकिन वे अपने घर सिरमौर से काफी दूर, बिलासपुर में पोस्टेड हैं. नेगी कबड्डी में भी देश का नाम ऊंचा कर चुकी हैं.\nक्या आपको लगता है कोई मैसेज झूठा ? सच जानने के लिए उसे हमारे नंबर पर भेजें.आप हमें पर ईमेल भी कर सकते हैं \n [सबसे तेज़ ख़बरों के लिए आजतक ऐप] [मंच पर महाभारत... केंद्र डांस फेस्टिवल में नृत्य नाटिका कर्ण का मार्मिक मंचन, टुल्लू मुर्मू के अभिनय ने जीता दिल]

Figure 10: Example chunks produced by the sentence chunking method on a Hindi fact-checking document. The figure reports English summaries and original Hindi chunks

Dataset	Instruction
X-FACT	Classify the given {language} claim into one of the seven categories: *TRUE, MOSTLY-TRUE, PARTLY-TRUE/MISLEADING, FALSE, MOSTLY-FALSE, COMPLICATED/HARD-TO-CATEGORISE, OTHER*, based on the provided evidence. The label definitions are: TRUE = fully supported by the evidence; MOSTLY-TRUE = mostly supported but with minor inaccuracies; PARTLY-TRUE/MISLEADING = partially supported but with significant omissions; FALSE = clearly contradicted or unsupported; MOSTLY-FALSE = largely incorrect with only a small element of truth; COMPLICATED/HARD-TO-CATEGORISE = too complex for a straightforward label; OTHER = does not fit the above categories. Provide exactly one label.
RU22FACT	Classify the given {language} claim into one of three categories: *SUPPORTED, REFUTED, NEI*, based on the provided evidence. The label definitions are: SUPPORTED = the evidence supports the claim; REFUTED = the evidence contradicts the claim; NEI = not enough information to verify. Provide exactly one label.

Table 10: Dataset-specific task instructions used for veracity prediction.

Component	Model Family	Model	Size
Retrieval encoder	Multilingual E5	multilingual-e5-large-instruct	560M
Veracity prediction	LLaMA	LLaMA-3.1-8B-Instruct	8B
Veracity prediction	Gemma	Gemma-2-9B-Instruct	9B
Veracity prediction	Mistral	Mistral-7B-Instruct-v0.3	7B

Table 11: Details of encoder and decoder models used in our experiments.

Hyperparameter	Value
Fine-tuning method	LoRA
Number of epochs	3
LoRA rank	8
Retrieved candidates	20
Final evidence chunks	5
Retrieval method	FAISS dense retrieval
Evaluation metric	Macro-F1

Table 12: Hyperparameter settings used across experiments.

You are evaluating retrieved evidence for a fact-checking task.

Claim: {claim}

Retrieved Evidence: {evidence}

Question: Does the retrieved evidence contain enough information to verify whether the claim is true or false?

Choose exactly one label:

Complete: The evidence contains sufficient information to verify the claim.
Partial: The evidence is related to the claim but incomplete.
Irrelevant: The evidence is unrelated, noisy, empty, blocked webpage text, or does not help verify the claim.

Return only one word: *Complete*, *Partial*, or *Irrelevant*.

Table 13: Prompt used for evidence completeness evaluation.

Dataset	Split	Model	B>O	O>B	Net Gain	McNemar χ^2	p-value	Significance
RU22FACT	Test	Gemma	31	136	+105	64.77	8.88×10^{-16}	✓
	Test	LLaMA	41	147	+106	58.64	1.89×10^{-14}	✓
	Test	Mistral	41	144	+103	56.24	6.42×10^{-14}	✓
X-FACT	ID	Gemma	306	487	+181	40.86	1.64×10^{-10}	✓
	ID	LLaMA	255	410	+155	35.66	2.35×10^{-9}	✓
	ID	Mistral	290	497	+207	53.92	2.09×10^{-13}	✓
	OOD	Gemma	187	329	+142	38.53	5.39×10^{-10}	✓
	OOD	LLaMA	171	298	+127	33.85	5.95×10^{-9}	✓
	OOD	Mistral	198	229	+31	2.11	1.47×10^{-1}	×
	ZS	Gemma	303	371	+68	6.66	9.86×10^{-3}	✓
	ZS	LLaMA	273	394	+121	21.59	3.38×10^{-6}	✓
	ZS	Mistral	322	389	+67	6.13	1.33×10^{-2}	✓

Table 14: McNemar significance test comparing our SEEK method against standard semantic chunking. B>O denotes examples correctly predicted only by the baseline, while O>B denotes examples correctly predicted only by our method. Net Gain is computed as O>B minus B>O. Significance is reported at $\alpha = 0.05$.

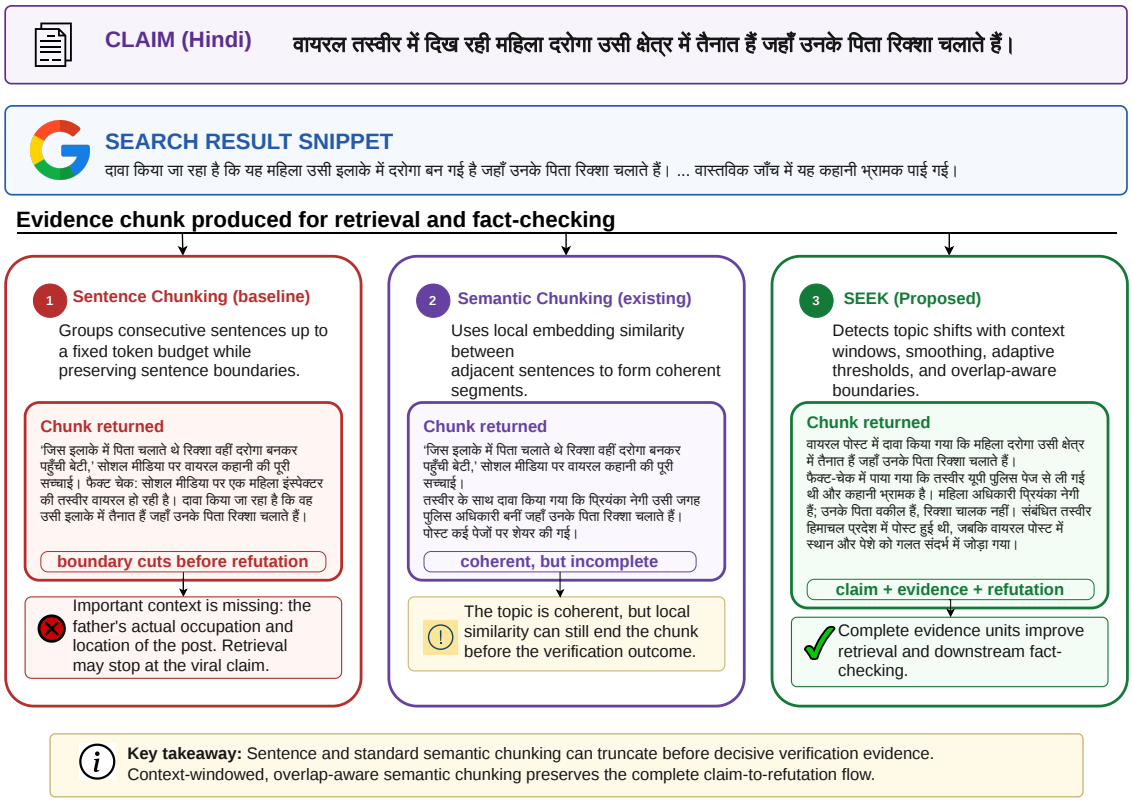

Figure 11: Illustration of the chunking strategies used for evidence retrieval. Sentence chunking groups consecutive sentences under a token budget but may truncate the evidence before the full verification context is reached. Existing semantic chunking uses embedding-based similarity to form coherent chunks, but can still miss later refuting evidence in noisy web documents. Our SEEK preserves a longer and more complete evidence unit by detecting topic shifts using contextual windows, smoothing, adaptive thresholding, and boundary overlap.

Dataset	Split	Model	B>O	O>B	Net Gain	McNemar χ^2	p-value	Significance
RU22FACT	Test	Gemma	33	283	+250	196.21	$< 10^{-16}$	✓
	Test	LLaMA	37	297	+260	200.84	$< 10^{-16}$	✓
	Test	Mistral	32	294	+262	208.96	$< 10^{-16}$	✓
X-FACT	ID	Gemma	310	465	+155	30.60	3.17×10^{-8}	✓
	ID	LLaMA	273	433	+160	35.81	2.18×10^{-9}	✓
	ID	Mistral	283	408	+125	22.25	2.39×10^{-6}	✓
	OOD	Gemma	212	344	+132	30.87	2.77×10^{-8}	✓
	OOD	LLaMA	208	241	+33	2.28	1.31×10^{-1}	×
	OOD	Mistral	158	297	+139	41.85	9.83×10^{-11}	✓
	ZS	Gemma	364	367	+3	0.01	9.41×10^{-1}	×
	ZS	LLaMA	259	402	+143	30.51	3.33×10^{-8}	✓
	ZS	Mistral	347	465	+118	16.86	4.03×10^{-5}	✓

Table 15: McNemar significance test comparing our SEEK method against sentence chunking.

Dataset	Split	Model	B>O	O>B	Net Gain	McNemar χ^2	p-value	Significance
X-FACT	ID	Gemma	294	1225	+931	569.39	$< 10^{-16}$	✓
	ID	LLaMA	260	1084	+824	503.97	$< 10^{-16}$	✓
	ID	Mistral	290	1186	+896	542.70	$< 10^{-16}$	✓
	OOD	Gemma	235	563	+328	134.00	$< 10^{-16}$	✓
	OOD	LLaMA	199	568	+369	176.56	$< 10^{-16}$	✓
	OOD	Mistral	180	588	+408	215.69	$< 10^{-16}$	✓
	ZS	Gemma	357	734	+377	129.58	$< 10^{-16}$	✓
	ZS	LLaMA	337	721	+384	138.65	$< 10^{-16}$	✓
	ZS	Mistral	365	669	+304	88.79	$< 10^{-16}$	✓

Table 16: McNemar significance test comparing our SEEK method against the concrete baseline.

Dataset	Split	Model	B>O	O>B	Net Gain	McNemar χ^2	p-value	Significance
X-FACT	ID	Gemma	356	1052	+696	343.06	$< 10^{-16}$	✓
	ID	LLaMA	320	1145	+825	463.46	$< 10^{-16}$	✓
	ID	Mistral	359	1031	+672	323.91	$< 10^{-16}$	✓
	OOD	Gemma	227	503	+276	103.60	$< 10^{-16}$	✓
	OOD	LLaMA	246	522	+276	98.47	$< 10^{-16}$	✓
	OOD	Mistral	184	466	+282	121.48	$< 10^{-16}$	✓
	ZS	Gemma	417	515	+98	10.10	1.49×10^{-3}	✓
	ZS	LLaMA	349	646	+297	88.06	$< 10^{-16}$	✓
	ZS	Mistral	364	622	+258	66.99	2.22×10^{-16}	✓

Table 17: McNemar significance test comparing our SEEK method against the Google snippet baseline.

Dataset	Split	Model	B>O	O>B	Net Gain	McNemar χ^2	p-value	Significance
RU22FACT	Test	Gemma	46	246	+200	135.62	$< 10^{-16}$	✓
	Test	LLaMA	44	261	+217	152.97	$< 10^{-16}$	✓
	Test	Mistral	41	253	+212	151.43	$< 10^{-16}$	✓

Table 18: McNemar significance test comparing our SEEK method against the LLM-based baseline.