Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Cross-Lingual Fact-Checking: Auditing Claims vs. Evidence

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-27 · cross-lingual fact checking · cross-lingual claim verification · multilingual fact checking · x-fact benchmark · xfever · concrete cross-lingual retrieval · translation errors in fact checking · source independence provenance · multilingual evidence retrieval

Original article: https://reliablecontext.com/articles/cross-lingual-fact-checking-verification-audit



In this article

1. Executive Summary
2. Introduction and Background
3. Task Boundaries and Evidence Conditions
4. The Anchor Studies and What Each Can Support
5. Auditing One Claim Through Five Stages
6. Experimental Design: Matched Budgets and Separate Evidence Objects
7. Failure Attribution: Translation, Transliteration, Coverage and Script
8. Source Independence and Provenance
9. Data Analysis and Evidence
10. Implications and Future Directions
11. Frequently Asked Questions (FAQs)
12. Conclusion

Executive Summary

Cross-lingual fact checking asks whether a system can judge a claim written in one language using evidence found in the same language or another one, and, for a researcher releasing a per-language result, which stage produced each error. This report separates five stages (claim detection and matching, evidence retrieval, passage support, source identity, verdict) and shows that the anchor papers measure different slices of them. **X-Fact** supplies **31,189** claims in **25 languages** (arxiv.org). Its abstract reports a best F-score of around 40% (aclanthology.org), which matches the in-domain test (41.9), not the zero-shot split, where the same paper's table gives **16.0** macro-F1 for that same evidence-based model with metadata (Attn-EA + Meta; 16.5 for the variant without metadata) (arxiv.org).

CONCRETE (Huang et al., 2022) reports a **2.23-point** absolute macro-F1 gain from cross-language retrieval (aclanthology.org). In its Table 2 the figure equals 19.83 minus 17.60, the score of the same mBERT reader with vanilla mDPR retrieval (aclanthology.org); against the paper's no-retrieval row (17.25) the gain is 2.58, and against the earlier published 16.7 it is 3.13 (derived arithmetic). Shcharbakova et al. (2025) evaluate three instruction-tuned LLMs on X-Fact and find, for few-shot prompting, the highest median for Latin-script languages across Llama 3.1 and Qwen 2.5 (aclanthology.org); their analysis repository states the pattern for all three models (github.com).

The design conclusions are practical. Language with more web text is not automatically easier: in Common Crawl crawl CC-MAIN-2026-43, English was the primary language of 42.56% of pages and Georgian of 0.047% (commoncrawl.github.io), yet the RANLP study found no clear resource-level pattern (aclanthology.org). Translation is not a neutral pre-processing step: it changed downstream F1 by up to 18 points in one cross-lingual open-retrieval QA benchmark (arxiv.org), yet cost under one accuracy point on a 600-pair professional-translation check (arxiv.org). Per-language samples are small enough that interval width matters, and the Wilson interval does not depend strongly on n or p (www.itl.nist.gov). Translations, syndications and fact-checks that cite one another are not independent sources, because a false value can spread through copying (lunadong.com).

The recommended audit holds the candidate pool and retrieval budget fixed, stores original text and translation as separate evidence objects, scores retrieval, support and verdict separately, and reports abstention and intervals per language and script. None of the cited results are pooled or independently reproduced here.

Introduction and Background

Cross-lingual claim verification is the task of assigning a veracity judgment to a claim when the supporting or refuting evidence may be in a different language from the claim. Three conditions are easy to conflate. In the **native** condition, claim and evidence are in the same original language. In the **translated** condition, one or both have been machine or human translated, as in XFEVER (arxiv.org). In the **cross-language** condition, a retriever must find evidence in languages other than the claim's, as in CONCRETE, where a claim in an arbitrary language is matched to passages that may be in any language (aclanthology.org).

The audience is researchers who build or compare multilingual fact-verification and knowledge-fusion systems and must decide whether a system really retrieves and uses cross-language evidence, and which language pairs or scripts need a separate error analysis. The decision concerns experimental design and the release of a defensible per-language result, not the choice of a translation product.

Two facts make an audit necessary. First, evidence quality is a dominant, under-reported variable. In the X-Fact paper's manual check of 100 Hindi examples, only 45% of search snippets held enough information to classify the claim (arxiv.org), while full web-page text would have sufficed for 83% (arxiv.org). Second, headline numbers come from different settings and are not a pooled benchmark. A search on this topic mixes verification, claim-detection and retrieval papers, which is why a stage-specific synthesis helps.

The human ecosystem that produces labelled claims shapes source coverage. The Duke Reporters' Lab census counted 437 active fact-checking projects as of June 1, 2026, down from 441 at the end of 2025 (reporterslab.org), after a peak of 464 in 2024 (reporterslab.org). Every multilingual claim dataset inherits the language and regional coverage of such a population.

Scope notes:

- **Out of scope:** translation products, and claim-detection (check-worthiness) benchmark papers, which were not retrieved; that stage is defined but not benchmarked.
- **Replication:** as of October 2026, no independent successful replication of CONCRETE was found in the sources opened for this report.

Task Boundaries and Evidence Conditions

A fair comparison starts by naming which stage a dataset exercises and which it assumes away. FEVER (Thorne et al., 2018) defined the verify-against-evidence template. Its scoring counts a Supported or Refuted prediction as correct only if correct evidence is returned (aclanthology.org). The X-Fact line of work does not couple the stages: it scores the verdict alone.

Table 1 places the studies discussed here on the five stages. It states what each study scores and what it holds fixed, so that raw numbers are not equated across incompatible splits.

Study	Stages exercised	Claim language	Evidence and its language	Labels and metric	Headline result as stated
X-Fact (Gupta and Srikumar, 2021)	Verdict, with retrieved snippets as input	25 languages, native claims from fact-checking sites	Top five Google results per claim (huggingface.co); snippet language not stated explicitly in the paper	Seven labels; macro-F1	About 40% F-score, in-domain (aclanthology.org)
CONCRETE (Huang et al., 2022)	Cross-language evidence retrieval, then verdict	X-Fact claims; zero-shot split shares no language with training	Passages from one news publisher, in seven language folders (github.com)	X-Fact labels; macro-F1	2.23 points absolute, zero-shot (see CONCRETE section)
XFEVER (Chang et al., 2023)	Verdict over given evidence	FEVER claims machine-translated into five languages	Single-sentence evidence, also translated	Three labels; accuracy	Accuracy varies by language (see XFEVER section)
Shcharbakova et al. (2025)	Verdict by LLM from given claim and evidence	X-Fact seen and unseen languages	X-Fact evidence, claim-evidence pairs	Seven labels; macro-F1	Best 0.34 macro-F1 (aclanthology.org)
MultiClaim (Pikuliak et al., 2023) and **SemEval-2025 Task 7**	Claim matching: retrieve previously fact-checked claims	Multilingual social-media posts	Multilingual fact-checks	Success@10 (arxiv.org)	Translation to English won in 2023 (aclanthology.org); negligible difference in 2025 (arxiv.org)

Table 1 shows three structural facts:

- **No joint scoring.** No anchor study scores evidence and verdict jointly, so a verdict gain cannot be attributed to retrieval without an ablation such as CONCRETE's.
- **Different evidence kinds.** X-Fact snippets come from general web search, CONCRETE evidence from one news publisher's passages, and XFEVER evidence from a translated Wikipedia sentence.
- **Different metrics.** Even macro-F1 has two formulas in use that can differ by as much as 0.5 (arxiv.org).

Macro averaging is the unweighted mean of per-label scores and ignores label imbalance (scikit-learn.org), whereas micro averaging counts true positives globally (scikit-learn.org), so the formula must be stated.

The Anchor Studies and What Each Can Support

X-Fact: native claims, search-snippet evidence

X-Fact contains **31,189** fact-checks, labelled by expert fact-checkers (arxiv.org). The paper's Table 2 gives train 19,079 (13 languages), dev 2,535, in-domain test 3,826, out-of-domain test 2,368 and zero-shot test 3,381 (12 languages), which sum to the total (arxiv.org). The paper's prose says 13 languages form the zero-

shot set, which disagrees with its own table and with the Hugging Face card, which lists the zero-shot-only languages separately (huggingface.co); this report uses the table value of 12.

Five truthfulness grades plus Unverifiable and Other make seven labels. Source ratings were translated to English and merged semi-automatically, and the per-website mapping files are in the repository (github.com). The evidence per claim is described on the dataset card as the top five documents from a Google search on the claim statement (huggingface.co).

Evidence rights. The repository LICENSE grants broad rights to use, copy and modify "the Software" (github.com), and the Hugging Face dataset card metadata declares an MIT license (huggingface.co). In the pages reviewed, no statement addresses redistribution of the search snippets or the fact-checkers' original text separately, and the short paper has no ethics or limitations section. A release plan should record rights per evidence object rather than inherit a dataset-level license.

Results and caveats. All experiments use the mBERT encoder, results are averaged over four seeds (github.com), and snippets (not full pages) are the evidence. Evidence improved the in-domain score by 2.5 points, but on the zero-shot split the paper's text says the claim-only model slightly outperforms the evidence-based one, whereas its table scores are 14.7 (claim only), 16.7 (claim plus metadata), 16.5 (evidence model without metadata, Attn-EA) and 16.0 (evidence model with metadata, Attn-EA + Meta), so only the claim-plus-metadata variant is ahead (arxiv.org).

XFEVER: translated claims and evidence

XFEVER translates FEVER claims verifiable from single evidence sentences (arxiv.org) into five target languages (Spanish, French, Indonesian, Japanese, Chinese). Its premise is that labels are language invariant, while conceding that the machine-translated data contain errors. A 600-pair test subset was translated by professionals; on it, average differences between machine and professional translation were roughly 0.3 to 0.7 accuracy points (arxiv.org). This is a bounded finding: five language pairs, 600 items, one source domain.

Japanese is the lowest language in the mBERT row at 72.4 accuracy against 87.9 for English (arxiv.org). XFEVER therefore tests translated claims and translated evidence, not native claims; it cannot show how a system behaves on claims whose phrasing and named entities originate in the target language. An implementation of the paper is published in a public repository (github.com).

CONCRETE: cross-language retrieval

CONCRETE feeds a claim in any language to a retriever that returns passages in any language, and an mBERT reader predicts the label. The evidence is a crawl of one news publisher's passages, 347,557 in total (aclanthology.org), released in seven language folders (github.com). Translations for retriever training samples were pre-computed, not produced on the fly (github.com).

The 2.23-point denominator. The paper states the gain as an absolute macro-F1 gain over the best other models (aclanthology.org), and its text does not name the baseline row. In Table 2 the reader is mBERT in every row; CONCRETE scores 19.83 zero-shot and 40.53 in-domain, and the closest other row, vanilla mDPR, scores 17.60 and 36.79, so 19.83 minus 17.60 is 2.23; other zero-shot rows are no retrieval at 17.25, BM25 at 17.43, Google Search snippets at 16.02 and translate-everything-then-retrieve at 15.01, and previously

published figures are 16.7 and 16.0 (aclanthology.org). The gain is therefore 2.58 against no retrieval and 3.13 against the earlier 16.7 (derived arithmetic). In-domain, CONCRETE trails Google Search snippets (40.53 against 42.61), which the authors attribute to the reader exploiting patterns in search results (aclanthology.org).

Ablations relevant to this audit. Translating retrieved passages and retraining gave a Kendall correlation of -0.490 between performance change and claim-passage language distance (aclanthology.org). The most common remaining error was missing evidence in the passage collection (aclanthology.org). The authors state the system is far from a reliable cross-lingual fact checker (aclanthology.org); the public code and retrieved passages are available for inspection (github.com).

Replication status. The DisAI Replication Challenge page, run by a consortium including the Kempelen Institute of Intelligent Technologies, DFKI, the University of Copenhagen and CERTH (disai.eu), lists CONCRETE as a suggested paper (disai.eu). The page describes an opportunity for early-stage researchers (disai.eu), with about ten participants selected (disai.eu) and three months to replicate (disai.eu). It says the paper list is not exhaustive (disai.eu) and that participants submit a report on achieved results (disai.eu). A listing is a candidate nomination; it is not evidence that the result was reproduced.

LLM evaluation on X-Fact

Shcharbakova et al. (2025) evaluate three instruction-tuned LLMs (Llama 3.1 8B, Qwen 2.5 7B and Mistral Nemo 12B) with seven-shot prompting and LoRA fine-tuning. "Seen" means languages in the X-Fact training split and "unseen" the 12-language zero-shot split (aclanthology.org); it does not mean absence from LLM pretraining, and the authors treat official language support as a separate confound.

Absolute scores are low. The best is 0.34 macro-F1, for Norwegian with fine-tuned Qwen 2.5, an unseen language (aclanthology.org). Latin-script languages have the highest median few-shot macro-F1 across Llama 3.1 and Qwen 2.5 (aclanthology.org), South Asian unseen languages fall below 0.13 (aclanthology.org), and the authors note that script difficulty cannot be separated from tokenizer inefficiency (aclanthology.org). The supplementary repository holds analysis results, not a full pipeline (github.com).

Auditing One Claim Through Five Stages

The audit unit is a claim traced through five stages, each with its own input, output and error type. The example is illustrative.

(Hypothetical Example) An Indonesian-language post states that a national statistics agency published a specific figure for a quantity in a given year.

- **Stage 1, claim detection and matching.** Decide whether the post contains a checkable claim and whether a previously fact-checked claim already matches it. Matching is a retrieval task with its own limits: the 2025 shared task counted success if at least one relevant claim was in the top 10 (arxiv.org).
- **Stage 2, evidence retrieval.** Query a fixed multilingual pool and record the language, script, URL and observation date of every candidate. Failure type: **retrieval**, when no relevant passage is in the top K.

- **Stage 3, passage support.** Label each retrieved passage as supporting, refuting or insufficient, in the original language and, separately, in translation. Failure types: **translation**, when the two labels differ; **verification**, when the original-language label is wrong.
- **Stage 4, source identity.** Resolve each passage to its originating publisher. Failure type: **source coverage**, when the pool lacks any independent originator for the claim's country or topic. This report's category overlaps with the most common remaining error in CONCRETE's analysis, supporting or refuting evidence missing from the passage collection, which was mostly seen for country-specific claims.
- **Stage 5, verdict.** Aggregate support into a label or abstain. FEVER-style scoring credits the verdict only if the correct evidence was returned (see Task Boundaries).

If the agency's own Indonesian release is in the pool but ranked below K, the failure belongs to Stage 2 even if the verdict is wrong. If it is retrieved and correct but a translated copy renders the figure with a different thousands separator, the failure belongs to Stage 3. If only a syndicated copy is retrieved, Stage 4 resolves it to the same originator. The X-Fact manual check described in the Introduction suggests that evidence insufficiency can account for errors that would otherwise be charged to the verdict model.

Experimental Design: Matched Budgets and Separate Evidence Objects

Paired fixtures

Build one fixture per claim-evidence pair. Preserve the **original text** and every **translation** as separate evidence objects, each with its own source URL, original language, script, translation method and version, and observation date. Do not overwrite the original with its translation: gains from translated training data have been observed only on translated test sets, not original ones (arxiv.org). The PROV-O vocabulary supplies the relations: an entity may be digital or conceptual (www.w3.org), an activity such as translation occurs over a period of time (www.w3.org), and a derivation is a transformation of one entity into another (www.w3.org). PROV-O is a W3C Recommendation dated 30 April 2013 (www.w3.org).

Record language and script as separate fields. In HTML and XML a language tag indicates the language of content (www.w3.org). BCP 47 script subtags indicate the script or writing system (www.rfc-editor.org), as in "sr-Latn" for Serbian in Latin script (www.rfc-editor.org), and are omitted when they add no distinguishing value (www.rfc-editor.org). A language's writing system can occasionally use more than one script (www.unicode.org), and script is not the same as Unicode block (www.unicode.org). W3C guidance is to keep tags as short as possible (www.w3.org), so the fixture should store the script explicitly even when the tag omits it. Fact-check search interfaces accept a language code but may ignore region (developers.google.com).

For the claim itself, schema.org defines ClaimReview as a fact-checking review of claims made in some creative work (schema.org), and it separates the claim summary (schema.org) from the item reviewed (schema.org); Google's documentation distinguishes the author of the claim from the author of the fact check (developers.google.com). The only language field a ClaimReview inherits is the content language property (schema.org). The documentation page was last updated on 2026-09-08 (developers.google.com).

Fixed pool and fixed budget

To compare native and translated evidence fairly, hold the candidate pool and the retrieval budget constant:

- **Native arm:** retrieve in the claim's language from the pool, top K passages.
- **Translated arm:** translate claim or pool into a pivot language, then retrieve with the same K over the same pool documents.
- **Cross-language arm:** retrieve from the full multilingual pool with the same K.
- **Fixed generation settings:** same reader, same prompt language, same label set.

Report **recall@K** per language, the fraction of relevant items retrieved in the top K (web.archive.org) (evaluating fixed low cutoffs such as 10 or 30 (web.archive.org)), and **complete-evidence recall**, the fraction of claims for which all required passages are in the top K. The second matters because multi-evidence tasks are hard to satisfy fully.

Multilingual fact checking evidence retrieval needs explicit measurement of language bias. A language-fair retriever should yield equivalent ranking lists for semantically identical queries in different languages, and lexical and neural rankers should be compared on that criterion rather than on pooled recall. A benchmark of monolingual retrieval cannot by itself validate cross-language retrieval, so the retrieval arm needs its own cross-language evaluation.

Scoring support, verdict and abstention

Score support and verdict separately, and include an explicit abstention outcome for cases where original-language evidence cannot be checked, for example because a page is blocked or the language is outside the verifier's competence. Selective classification formalizes the trade, abstaining when in doubt while keeping coverage as high as possible (arxiv.org). Report coverage with every accuracy figure.

Table 2 lists the per-language, per-script and per-evidence-language columns a defensible release should contain. It specifies definitions and denominators and holds no empirical values, because none of the anchor studies publishes this matrix.

Column	Definition	Denominator	Interval or caveat
Claim language and script	Original language and writing system of the claim	Claims with that language	Report scripts separately: script and tokenizer cost are confounded
Evidence language	Language of each retrieved passage, original or translated	Passages	Never merge original and translation rows
n claims	Number of claims in the cell	n/a	Small n gives wide intervals
Recall@K	Relevant items in top K over relevant items	Claims with at least one relevant item	Wilson interval (www.itl.nist.gov)
Complete-evidence recall	Claims with all required passages in top K	Claims needing multiple passages	Wilson interval
Support accuracy	Correct support label per passage	Retrieved passages with gold support labels	Cluster by claim, not by passage

Column	Definition	Denominator	Interval or caveat
Verdict macro-F1	Unweighted mean of per-class F1	Claims	State the formula (scikit-learn.org); bootstrap resampling for intervals
Abstention rate and coverage	Share of claims left unanswered, and answered share	Claims	Report accuracy at that coverage
Source-independent count	Originators after resolving translations and syndication	Passages	See Table 3

The matrix makes small cells visible. Underpowered experiments are common in NLP (aclanthology.org). For proportions, the Wilson interval does not depend strongly on n or p (www.itl.nist.gov), and its lower limit cannot be negative (www.itl.nist.gov). **(Hypothetical Example)** Computed in this report: 30 correct out of 50 claims (60%) has a 95% Wilson interval of 0.462 to 0.724; the same 60% on 200 claims has 0.531 to 0.665; on 1,000 claims, 0.569 to 0.630. With 0 errors in 40 trials the Wilson upper bound is 0.088, whereas the Wald interval degenerates to zero width. Per-language totals in multilingual benchmarks are modest, and any language-by-label-by-evidence cell is far smaller.

Failure Attribution: Translation, Transliteration, Coverage and Script

Every error in the five-stage trace belongs to one of four origins, each with its own diagnostic:

- **Retrieval:** the needed passage is absent from the top K; diagnostic is recall@K and complete-evidence recall per language.
- **Translation:** the support label differs between the original and its translation; diagnostic is label agreement across the paired evidence objects.
- **Source coverage:** the pool holds no independent originator for the claim; diagnostic is the source-independent count in Table 3.
- **Verification:** the label is wrong on correct, original-language evidence; diagnostic is support accuracy and macro-F1.

Translation drift

The evidence on whether to translate is mixed and depends on task, language and translation system. For retrieval, translation can dominate outcomes: in XOR-TyDi QA, human-translated queries gave 38.2 F1 in one setting against 32.9 and 20.5 with two machine translation systems (arxiv.org). In claim matching, English embeddings on machine-translated input were the best option in 2023 (Table 1), while the 2025 shared task found the difference negligible (Table 1), which its organisers read as improvement in multilingual models.

Entity transliteration

Named entities are the sharpest translation hazard because a retriever matches on them. In XOR-TyDi QA, a Japanese question about a racehorse was translated as being about almonds, and almond articles were retrieved (arxiv.org). A fixture should therefore store the original-script entity string, the transliteration or

translation, and a resolved identifier, and measure retrieval with and without the translated form.

Script and tokenization

Tokenizers can encode the same text at very different lengths in different languages, which is one reason script and cost effects are confounded. In the RANLP study, for the fine-tuned models, invalid outputs (empty responses or labels not given in English, including answers in the input language) ranged from 2 examples for Qwen 2.5 on the test subset to 20.05% of the test subset (20.14% of the zero-shot subset) for Mistral Nemo (aclanthology.org). Invalid-output rates must be reported alongside macro-F1, since a model that answers in the wrong language is scored wrong without being evaluated on the claim.

More web text is not automatically easier

Web text volume is very unequal. Common Crawl identifies page language with CLD2 (commoncrawl.github.io) and only HTML pages are detected (commoncrawl.github.io). In crawl CC-MAIN-2026-43, the primary-language share was 42.56% for English (commoncrawl.github.io), 1.89% for Polish (commoncrawl.github.io), 1.06% for Indonesian (commoncrawl.github.io), 0.67% for Arabic (commoncrawl.github.io), 0.34% for Norwegian (commoncrawl.github.io), 0.21% for Hindi (commoncrawl.github.io) and 0.047% for Georgian (commoncrawl.github.io). On Wikipedia, English has 7,255,635 articles (meta.wikimedia.org), Indonesian 802,536 (meta.wikimedia.org) and Georgian 200,802 (meta.wikimedia.org); the page is a live table updated several times a day (meta.wikimedia.org).

Set these against the RANLP results (an observation, not a statistical test). The best seen languages, Polish and Indonesian, reach about 0.31 and hold 1.89% and 1.06% of the crawl; the best unseen language, Norwegian, holds 0.34% yet reached the study's top score of 0.34 (seen and unseen subsets differ, so these scores are not strictly comparable); Arabic (0.67%) and Georgian (0.047%) were poorest among seen languages (aclanthology.org). The resource-level analysis found no clear pattern (aclanthology.org). Within a benchmark, language is also nested in source: X-Fact's out-of-domain split is defined by a different fact-checking source (arxiv.org), so per-language difficulty is confounded with site, label mapping and topic. A language with more web text cannot be assumed easier without controlling these variables.

Source Independence and Provenance

Cross-language evidence multiplies duplicates. An agency release, its English version, a newspaper summary, a syndicated translation and a fact-check that cites the summary are five documents but one upstream report. Truth discovery integrates multi-source information by estimating source reliability (arxiv.org), usually assuming sources make observations independently. Copying breaks that assumption, because a false value can be spread through copying (lunadong.com), and sources that share many rare values are likely to be copying from each other (lunadong.com). Detection approaches include probabilistic models of copying over update histories (lunadong.com).

Professional norms point the same way. The International Fact-Checking Network code requires sources in enough detail for readers to replicate the work (ifcncodeofprinciples.poynter.org), checking key elements against more than one named source (ifcncodeofprinciples.poynter.org), and the best available primary

sources (ifcncodeofprinciples.poynter.org). Applications for signatory status are assessed by independent assessors against 31 criteria (ifcncodeofprinciples.poynter.org).

Table 3 shows how an audit should resolve a set of documents to originators before counting support. The example is hypothetical.

Document (Hypothetical Example)	Language	Relation	Originating publisher	Counts as independent?
Agency data release	Indonesian	Primary source	The agency	Yes (originator 1)
Agency English version	English	Translation of the release	The agency	No, same originator
Newspaper article quoting the release	Indonesian	Derived from the release	The agency	No, derived
Syndicated Spanish version of the article	Spanish	Translation of a derived document	The agency	No, derived twice
Portuguese fact-check citing the article	Portuguese	Derived, analytic	The agency (for the figure)	No for the figure; the check itself is separate work
University survey with its own measurement	English	Separate measurement	The university	Yes (originator 2)

Six documents resolve to two independent originators. A verifier that counted six supporting passages would overstate corroboration threefold. The PROV derivation relation (www.w3.org) is the formal basis for the resolution, and a stored originating-publisher field makes the count reproducible. Support from a fact-check should be recorded as support for the fact-checker's conclusion, not as independent measurement of the underlying figure.

Data Analysis and Evidence

The quantitative evidence cannot be pooled, but it can be tabulated by setting and re-expressed on common footing where the arithmetic is valid. Published figures are reported as stated by their sources; derived quantities are marked.

Published figures, as stated.

- **AVeriTeC:** 4,568 real-world claims from 50 organizations (arxiv.org).
- **Common Crawl, CC-MAIN-2026-43:** English 42.56% of pages, Georgian 0.047% (commoncrawl.github.io).
- **Wikipedia:** English 7,255,635 articles against 200,802 for Georgian (meta.wikimedia.org), as of the access date of 2026-10-27, because the table updates several times a day (meta.wikimedia.org).
- **Fact-checking projects:** 437 active as of June 1, 2026, against 341 in the 2021 census (reporterslab.org).

- **Evidence sufficiency in X-Fact (100 Hindi examples):** 45% of snippets sufficient against 83% for full pages (arxiv.org).

Derived quantities (arithmetic on the figures cited above, not new measurements).

- **X-Fact in-domain to zero-shot gap:** the best in-domain model (Attn-EA + Meta) falls from 41.9 to 16.0 macro-F1 on the zero-shot split, a drop of 25.9 points, or to about 38% of its in-domain score; Attn-EA without metadata scores 16.5 there.
- **CONCRETE gain, three denominators:** 2.23 points over vanilla mDPR (17.60), 2.58 over no retrieval (17.25) and 3.13 over the earlier published 16.7. Relative to the mDPR row, 2.23 is a 12.7% relative improvement.
- **Web share ratios:** English to Georgian is about 900 to 1 in Common Crawl pages and about 36 to 1 in Wikipedia articles; Indonesian to Norwegian is about 3.1 to 1 in pages.
- **Fact-checker supply:** the 2026 count is 96 projects (28%) above the 2021 census but 27 (5.8%) below the 2024 peak of 464.

Interval width (computed in this report). For a 60% accuracy, the 95% Wilson interval has these widths:

- **n = 50:** 0.262 (0.462 to 0.724).
- **n = 200:** 0.134 (0.531 to 0.665).
- **n = 1,000:** 0.061 (0.569 to 0.630).

Quadrupling n roughly halves the width. A language cell with 50 claims therefore cannot reliably distinguish 0.50 from 0.70.

What the numbers do and do not show. The 2.23-point CONCRETE gain is one setting, with one reader and one evidence collection. The XFEVER 0.3 to 0.7 point translation gap is one 600-pair subset. Variance reporting is scarce: X-Fact averages over four seeds (github.com). Every number above needs its setting beside it, and the audit should recompute baselines rather than quote them.

Implications and Future Directions

- **For experimental design.** Release a per-language, per-script, per-evidence-language table (Table 2) with intervals, and report retrieval and verdict separately. Treat any language-level claim from fewer than a few hundred items as exploratory. Prefer fixed pools over live web search for comparisons, because the X-Fact snippets depend on a search engine at a point in time, and record observation dates for every retrieved page. Re-run translation arms whenever the translation system changes.
- **For replication.** CONCRETE's public code, retrieved passages and pre-computed translations (github.com) make replication feasible, but the challenge listing is only a candidate nomination (disai.eu). A replication report should state the baseline row against which any gain is computed, as this report does.
- **For source coverage.** With 437 active fact-checking projects worldwide in 2026 (reporterslab.org) against 341 in the 2021 census (reporterslab.org), and an earlier report counting 443 (reporterslab.org) (year totals are revised retroactively), the pool of originators is finite and unevenly spread across

languages. The census also recorded 10 new projects in 2025 against more than 30 that stopped posting (reporterslab.org). Coverage gaps by country are a retrieval-independent ceiling on recall.

- **For provenance record keeping.** The audit above is a provenance problem as much as a modelling one: every value needs a verbatim source, an observation date and a declared reason when evidence is missing. Reliable Context, the publisher of this report, is a system of record built around those properties: one canonical record per real thing, a verbatim cited source and observation date behind every value, unknowns declared as not found, blocked or not applicable with a reason, and any past state of the record reproducible. It runs inside a customer's perimeter. It is not a vector database, a RAG framework, a data catalog or a verification model, and this report makes no claim that it improves any metric discussed here. Teams that need retrieval infrastructure or a verification model should choose those tools directly.
- **Open questions.** No source reviewed reports a controlled native-versus-translated evidence experiment at matched retrieval budgets across many languages for a native-claim dataset; this is the gap the proposed audit targets. Whether the script gap seen in LLMs persists for larger models is unresolved, since the study used modest-scale models.

Frequently Asked Questions (FAQs)

How do you verify claims when the evidence is in another language?

Retrieve from a fixed multilingual pool, store the original passage and its translation as separate evidence objects, judge support on the original where a competent reader or model exists, and use translation only as a recorded fallback.

What is a cross-language evidence verification audit for claim and evidence language mismatch?

It compares three arms (native, translated, cross-language) at the same K over the same pool, then attributes each error to retrieval, translation, source coverage or verification. In CONCRETE, performance change after translating passages correlated negatively with claim-passage language distance (aclanthology.org).

Which multilingual claim verification datasets exist?

X-Fact (25 languages, native claims) (huggingface.co), XFEVER (translated FEVER) (github.com), MultiClaim for claim matching, and AVeriTeC for English claims with justifications. Only the first two support verdict scoring in several languages.

How common are translation errors in fact checking?

No source reviewed measures the rate on fact-checking text across many languages. The measured effects are task specific: small on a 600-pair XFEVER subset (see above), and large in XOR-TyDi QA, where entity errors can derail retrieval (arxiv.org).

How should multilingual sources be verified for data provenance?

Record original language, URL, publisher, translation method and observation date, then resolve translations and syndication to originating publishers (Table 3) before counting corroboration. Use PROV derivation for the chain (www.w3.org) and ClaimReview fields for the claim and reviewed item (schema.org).

What is the difference between cross-lingual misinformation detection and claim verification?

Detection and matching decide whether a post contains a claim already checked; verification judges the claim against evidence. Matching is scored by Success@10 (arxiv.org), verification by label metrics such as macro-F1 (scikit-learn.org).

Does a language with more web text perform better?

Not reliably. The RANLP study found no clear resource-level pattern (see above), while Common Crawl page shares differ by a factor of roughly 900 between English and Georgian (commoncrawl.github.io).

Conclusion

Cross-lingual fact checking is five measurements, not one. The anchor papers each measure a slice: X-Fact native claims with snippet evidence and a verdict score, XFEVER translated claims and evidence with accuracy, CONCRETE a retrieval intervention measured against a stated row of its own table, and the RANLP study LLM verdicts by seen and unseen language and by script. Their numbers use different splits, metrics and baselines and should be reported side by side, never pooled.

A defensible per-language result holds the candidate pool and retrieval budget fixed, keeps original and translated evidence apart, scores retrieval, support and verdict separately, resolves duplicates to originating publishers, reports abstention, and attaches intervals to every cell. Where the sources are silent, including controlled native-versus-translated comparisons at matched budgets, the correct statement is that no public benchmark exists, not that the effect is small.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://arxiv.org/html/2106.09248v1
2. https://aclanthology.org/2021.acl-short.86
3. https://aclanthology.org/2022.coling-1.86.pdf
4. https://aclanthology.org/2025.ranlp-1.131.pdf
5. https://github.com/Aniezka/cross-lingual-fact-verification
6. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.csv
7. https://arxiv.org/html/2010.11856v3
8. https://arxiv.org/pdf/2310.16278
9. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm

10. https://lunadong.com/publication/dependence_vldb.pdf
11. https://reporterslab.org/2026/06/12/2026-census-fact-checking-losses-continue-amid-funding-pressure-but-most-projects-persist
12. https://aclanthology.org/N18-1074.pdf
13. https://huggingface.co/datasets/utahnlp/x-fact
14. https://github.com/khuangaf/concrete
15. https://arxiv.org/pdf/2505.10740
16. https://aclanthology.org/2023.emnlp-main.1027.pdf
17. https://arxiv.org/abs/1911.03347
18. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html
19. https://github.com/utahnlp/x-fact
20. https://github.com/utahnlp/x-fact/raw/refs/heads/main/LICENSE
21. https://huggingface.co/datasets/utahnlp/x-fact/raw/main/README.md
22. https://github.com/nii-yamagishilab/xfever
23. https://disai.eu/about
24. https://disai.eu/replication-challenge
25. https://arxiv.org/html/2004.04721v4
26. https://www.w3.org/TR/prov-o
27. https://www.w3.org/International/questions/qa-choosing-language-tags
28. https://www.rfc-editor.org/rfc/rfc5646.html
29. https://www.unicode.org/reports/tr24
30. https://developers.google.com/fact-check/tools/api/reference/rest/v1alpha1/claims/search
31. https://schema.org/ClaimReview
32. https://developers.google.com/search/docs/appearance/structured-data/factcheck?hl=en
33. https://web.archive.org/web/2023/https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-unranked-retrieval-sets-1.html
34. https://web.archive.org/web/2023/https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html
35. https://arxiv.org/pdf/1705.08500
36. https://aclanthology.org/2020.emnlp-main.745
37. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html
38. https://meta.wikimedia.org/wiki/List_of_Wikipedias
39. https://arxiv.org/abs/1505.02463
40. https://lunadong.com/publication/lifespan_vldb.pdf
41. https://ifcncodeofprinciples.poynter.org/know-more/the-commitments-of-the-code-of-principles
42. https://arxiv.org/abs/2305.13117
43. https://reporterslab.org/2021/06/02/fact-checking-census-shows-slower-growth
44. https://reporterslab.org/latest-news

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.