AfrIFact: Cultural Information Retrieval, Evidence Extraction And Fact Checking for African Languages

Anonymous ACL submission

Abstract

Assessing the veracity of a claim made online is a complex and important task with real-world implications. When these claims are directed at communities with limited access to information and the content concerns issues such as healthcare and culture, the consequences intensify, especially in low-resource languages. In this work, we introduce AfrIFact, a dataset that covers the necessary steps for automatic fact-checking (i.e., information retrieval, evidence extraction, and fact checking), in ten African languages and English. Our evaluation results show that even the best embedding models lack cross-lingual retrieval capabilities, and that cultural and news documents are easier to retrieve than healthcare-domain documents, both in large corpora and in single documents. We show that LLMs lack robust multilingual fact-verification capabilities in African languages, while few-shot prompting improves performance by up to 43% in `AfriqueQwen-14B`, and task-specific fine-tuning further improves fact-checking accuracy by up to 26%. These findings, along with our release of the AfrIFact dataset, encourage work on low-resource information retrieval, evidence retrieval, and fact checking.

1 Introduction

Information poverty, defined as the lack of equitable access to information, presents a significant global challenge that disproportionately impacts marginalized communities and regions where low-resource languages are spoken (Gebremichael and Jackson, 2006). With the advent of Large Language Models (LLMs) capable of generating, rewriting, and summarizing information, ensuring the validity and correctness of their outputs remains a critical concern. This issue is especially acute in linguistically diverse contexts, as most current systems are primarily engineered for English-speaking populations (Liu et al., 2025). Therefore, effectively addressing access to reliable and factual information requires not only improving access but also developing accurate and equitable language technologies, including extracting valuable information from online sources, identifying valid evidence from given documents, and verifying claims with substantiating evidence (Bender et al., 2021).

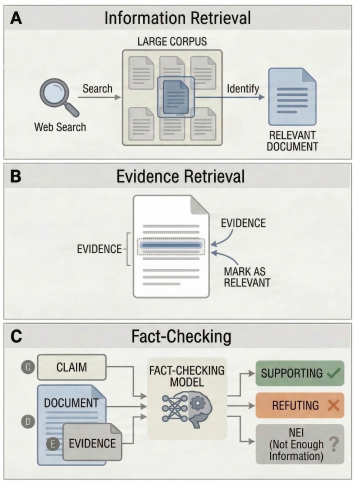


Figure 1: Three steps for verifying factual claims are covered by our dataset. (A) information retrieval to find documents related to a claim, (B) extraction of sentence-level evidence supporting or refuting the claim, and (C) fact-checking using the claim, retrieved document, and extracted evidence with LLMs.

Fact-checking and evidence retrieval (Thorne et al., 2018; Schlichtkrull et al., 2023; Thorne et al., 2018) aim to verify the truthfulness of claims using reliable sources. In practice, claims made by individuals or organizations cannot be classified as true or false without supporting evidence. Consequently, real-life fact-checking typically involves three steps: identifying relevant documents or web resources related to a claim, extracting specific sections that support or refute the claim, and classifying the claim as either supported or refuted by the evidence.

In real-world settings, claims disseminated

Dataset	# Lang.	# Native African Lang.	Domains	Tasks
X-FACT (Gupta and Srikumar, 2021)	25	0	News, Politics	Claim Verification
XFEVER (Chang et al., 2023)	6	0	Wikipedia	Multilingual Claim Verification
MultiClaim (Panchendrarajan et al., 2025)	78	1 (Sudanese)	Social Media, News	Claim Matching, Retrieval
Poly-FEVER (Zhang et al., 2025a)	11	1 (Amharic)	Wikipedia, Scientific Articles	Multilingual Fact Verification
AfrIFact(ours)	**11**	**10**	**Health, Culture(Wiki)-News**	**Fact Verification, Evidence Extraction, Information Retrieval**

Table 1: **Comparison of multilingual fact-checking datasets.** We report the number of supported languages, number of African languages, domain coverage, and task types. **AfrIFact** focuses on African languages and introduces multilingual information retrieval, evidence extraction and claim verification across culturally grounded domains.

through social media can significantly influence public health decisions, social behaviors, and cultural perspectives. As a result, the ability to reliably evaluate and verify the veracity of such claims is increasingly essential. Furthermore, current systems are rarely evaluated across diverse cultural and linguistic contexts, such as those in Africa, which may undermine reliable access to information by allowing unreliable information to proliferate.

In this work, inspired by the fact-checking process shown in Figure 1, we introduce **AfrIFact**, a multilingual benchmark containing more than 18,000 claims in 10 African languages and English to assess the capability of **embedding models** and **LLMs** to answer the following research questions: First, to investigate the capability in extracting relevant documents from large-scale resources, we ask ***(RQ1)*** *What is the effectiveness of embedding models in performing information retrieval for under-resourced African languages?* Secondly, to identify and select the specific sentences or passages that support or refute a claim, assessing the quality and reliability of their evidence extraction, we explore ***(RQ2)*** *How precisely can embedding models identify and extract relevant evidence from single document for claim verification?* Finally, we ask ***(RQ3)*** *To what extent can current LLMs correctly verify the factuality of claims?* by asking LLMs to classify claims as `SUPPORTED`, `REFUTED`, or, when the available evidence is insufficient to draw a reliable conclusion, `NOT_ENOUGH_INFORMATION`. To answer this question, this work makes the following contributions:

- We introduce **AfrIFact**, a multilingual information retrieval, evidence retrieval and fact checking benchmark covering healthcare, culturally grounded content, with the aim of creating a large-scale evaluation [1].
- We investigate both monolingual and multilingual **information retrieval** performance using multilingual embedding models.
- We assess the document-level **evidence extraction** abilities of embedding models in monolingual, culturally grounded single document settings.
- We evaluate the **fact-checking capabilities** of LLMs across 10 African languages through **claim classification**, both with and without labeled evidence spans, and examine methods for improvement, such as few-shot prompting and fine-tuning.

2 Related Work

Automated Fact Checking (AFC) has become an important research area with the goal of combating misinformation by verifying factual claims using computational methods (Vlachos and Riedel, 2014; Thorne et al., 2018; Schlichtkrull et al., 2023). Existing research typically model AFC as a multi-stage pipeline that includes claim detection, evidence retrieval, and claim verification (Guo et al., 2022). Prior work has explored identifying check-worthy claims in several discourse (Hassan et al., 2015), as well as retrieving evidence from large knowledge sources such as Wikipedia (Thorne et al., 2018; Augenstein et al., 2019) and reasoning over them to verify claims (Nakashole and Mitchell, 2014; Thorne et al., 2018; Potthast et al., 2018).

Table 1 summarizes widely used fact-checking benchmarks, their language coverage, their domains and tasks they can be used for. However, most existing datasets and approaches focus primarily on English, and relatively little work has explored AFC for African languages. MultiClaim (Panchendrarajan et al., 2025) is multilingual on a scale, but includes limited African language coverage with one language i.e., primarily Sudanese. In contrast, AfrIFact focuses explic-

[1] Datasets, seed documents, IR documents, and models will be made available upon acceptance under a CC-BY-4.0 license

itly on African languages and supports culturally grounded domains, enabling multilingual fact verification with evidence extraction and retrieval.

Information Retrieval (IR) Classical evidence retrieval (ER) in AFC courses on retrieving evidence that directly refutes a specific claim (Zheng et al., 2024a). In contrast, IR is a broader task that aims to retrieve relevant documents for downstream applications, not limited to claim verification. For African languages, several research efforts have focused on information retrieval, including the development of IR datasets such as CLIRMatrix (Sun and Duh, 2020), AfriCLIRMatrix (Ogundepo et al., 2022), MIRACL (Zhang et al., 2023), CIRAL (Adeyemi et al., 2023); however, these datasets are often limited in scale or domain coverage, and not designed for fact checking purpose. In this work, we repurpose AfrIFact for IR, enabling benchmarking of embedding models in monolingual and multilingual IR.

3 AfrIFact Dataset

AfrIFact covers two domains, health and culture-news, covering ten African languages. These domains were selected because of their high societal impact, their susceptibility to misinformation, and the availability of seed data for dataset curation.

Language Coverage: Included languages are Amharic, Hausa, Igbo, Oromo, Shona, Swahili, Twi, Wolof, Yorùbá, and isiZulu. The selection of these languages was influenced by the number of speakers, the availability of web resources, typological diversity, geographical diversity (West, East and Southern Africa), the coverage in previous datasets, and the availability of annotators. For each selected language, native-speaking annotators were recruited through an African AI community. Language coordinators, who are also native speakers, were recruited to oversee the annotators. Linguists were prioritized and, when unavailable, annotators with prior experience in similar projects were prioritized. More details about the languages and annotators are provided in Appendix A.

Figure 2 shows the data curation pipeline for both domains in AfrIFact. In the following sections, we describe the steps in detail.

3.1 *AfrIFact-Health*

To create *AfrIFact-Health*, which focuses on the factuality of healthcare information, particularly common diseases, recurrent outbreaks, and public

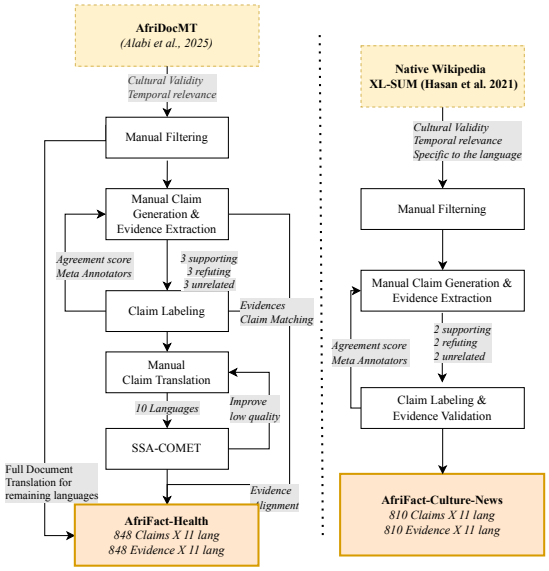


Figure 2: Illustration of the data construction process for *AfrIFact-Health* on shared health-data approach for lack of structured native health documents and *AfrIFact-Culture-News* culturally grounded, natively sourced data approach

health concerns, we build on the health split of the AfriDocMT (Alabi et al., 2025). This dataset is a document-level human-translated machine translation (MT) corpus containing 334 articles from the World Health Organization (WHO) website. The articles are translated from English into five African languages, all of which are included in this project. In *AfrIFact-Health*, claim generation, evidence extraction, and factuality labeling are conducted in English. These annotations are then projected to the corresponding African language data through a translation and alignment process.

Manual Filtering: We manually filtered 93 documents from AfriDocMT for use in this work. The selection process was guided by three criteria: cultural relevance and temporal validity. To support temporal validity, we excluded time-dependent materials, such as short-term reports, and retained only documents whose content is expected to remain relevant over extended periods. These 93 documents, totaling 2,600 sentences, are translated into five additional African languages not covered by AfriDocMT. The translations are done by human expert translators using the same guidelines as in the AfriDocMT dataset.

Claim Generation and Evidence Extraction: Following the AVeriTeC framework (Schlichtkrull et al., 2023), which guides claim generation via

Split	# Lang.	Total Split			Total
		Train	Val	Test	
Health	11	1100	550	7678	
Culture & News	11	1228	550	7150	
		2328	1100	14828	18256

Table 2: **Dataset statistics** (train, validation, test) for *AfrIFact-Health* and *AfrIFact-Culture-News* across 11 languages. Detail data distribution in Appendix A.

question answering, and considering that *AfrIFact-Health* is a multiparallel dataset translated from English, as well as the strong performance of recent LLMs on English tasks, we first generated questions to steer the claim generation process using `GPT-4.1`. Given the generated questions, annotators, under the supervision of language coordinators, created three supporting, three refuting, and three unrelated (insufficient information) claims per document. The questions helped them focus on important topics and ensure that each claim was relevant and coherent. To support or refute claims, annotators also identified evidence spans within the source documents, where evidence is restricted to the sentence level and consists of one or more contiguous full sentences. In total, this process produced 848 claims.

Claim Labeling: Three annotators followed the three-class annotation schema (Thorne et al., 2018) for AFC to label the generated claims, and to verify the associated evidence spans.

Manual Claim Translation: At this stage, we have the English documents, generated claims and their labels, corresponding evidence spans. To construct the claims across African languages, we performed manual translation with native speakers for each of the ten target languages.

Translation Quality: To verify the quality of the human translations of documents and claims from English into the ten target languages, we employed SSA-COMET (Li et al., 2025). Following prior work (Azime et al., 2025; Alabi et al., 2025), we filtered out translations with scores below 0.6 and engaged language coordinators and translators to either revise them or flag them as low-quality. While SSA-COMET occasionally produced false positives as identified by language coordinators, we found that the metric served as a useful quality-assistance tool for identifying and improving sub-optimal translations.

Evidence Alignment: Given the sentence-level parallelism provided within AfriDocMT, we leveraged evidence spans annotated and verified in English to construct multilingual evidence for the *AfrIFact-Health* split. This heuristic approach enables a reliable transfer of evidence, as all evidence spans are defined at the sentence level. Subsequently, we involved language coordinators to verify the evidence extracted in each target language and to ensure alignment with the corresponding English evidence.

3.2 *AfrIFact-Culture-News*

For *AfrIFact-Culture-News*, which focuses on the factuality of information related to cultural knowledge and news, we manually curated Wikipedia articles covering Africa-centric topics inspired by Thorne et al. (2018); Chang et al. (2023). We collected Wikipedia articles in the ten target languages covering locally relevant topics, including political leaders, food, regions, languages, historical events, sports, religion, currencies, and tourism. To further expand regionally relevant content, we added the News domain dataset, which comprises news articles written in the study languages and reporting on regional events. We sourced these articles from XL-Sum (Hasan et al., 2021), a large-scale news summarization dataset, and retained documents after filtering for document length and safety considerations.

Manual Filtering: We prioritized culturally focused Wikipedia articles and supplemented them with news documents where cultural coverage was limited, resulting in 200 documents per language, each containing more than 400 sentences and culturally and linguistically authentic documents.

Claim Generation and Evidence Extraction: To accommodate the diversity of documents, annotators were tasked with generating two `SUPPORTS`, two `REFUTES`, and two `NOT_ENOUGH_INFORMATION` claims for each document based on selected evidence, each accompanied by annotated evidence sentences. Annotators were also responsible for verifying document quality and validity; if a document contained missing information or did not meet regional relevance criteria, they were instructed to skip it without generating claims.

Claim Labeling and Evidence Validation: Following the same claim labeling procedure used in the health data collection, annotators worked with native-language claims, evidence spans, and source documents to assign of three labels: `SUPPORTS`, `REFUTES`, or `NOT_ENOUGH_INFORMATION`.

Dataset	Metric	REFUTES	NEI	SUPPORTS
Health	Fleiss' K.	0.74	0.79	0.82
	Cohen's K.	0.69	0.74	0.79
	Krippendorff's	0.74	0.79	0.82
Culture-News	Fleiss' K.	0.82	0.83	0.83
	Cohen's K.	0.81	0.82	0.82
	Krippendorff's	0.85	0.85	0.85

Table 3: **Agreement scores** for English *AfrIFact-Health* and averaged per language for *AfrIFact-Culture-News* split. Detailed result found in the Appendix A.

Agreement Scores For both *AfrIFact-Health* and *AfrIFact-Culture-News* splits, each claim was independently labeled as SUPPORTS, REFUTES, or NOT_ENOUGH_INFORMATION using three annotators. To assess the reliability of the annotations, we computed Cohen's kappa (Cohen, 1960), Fleiss' kappa (Fleiss, 1971), Krippendorff's alpha (Krippendorff, 2011) alpha as inter-annotator agreement measures. As shown in Table 3, the dataset achieves substantial agreement across all metrics, indicating high annotation quality and supporting its reliability for evaluating the factuality of large language models.

4 Methodology

Given AfrIFact, we address our research questions and detail the experimental setup used in this study.

Information Retrieval Documents: Our main evaluation setting uses a universal multilingual multi-domain corpus for all types of queries, which covers health, news, and Wikipedia articles from all languages.[2] This design is to simplify cross-lingual cross-task results comparison on our benchmark.

As data on *AfrIFact-Health*are parallel across languages, we evaluate it on both monolingual and multilingual settings: where the monolingual setting evaluates retrieval results on only the relevant documents from query's language,[3] and the multilingual setting evaluates retrieval results using relevant documents in all languages.[4]

Information Retrieval & Evidence Extraction Evaluation: In both extraction tasks, we identify the subset of textual units within a document or collection of documents that provide supporting or refuting evidence for a given claim. In our setting, the system retrieves the top-k ranked segments as candidate evidence. These extracted evidence segments are subsequently used by a downstream fact verification model to determine whether the claim is SUPPORTS, REFUTES, or NOT_ENOUGH_INFORMATION.

For information retrieval following prior literature (Enevoldsen et al., 2025; Uemura et al., 2025), we use nDCG@10 as the primary metric and also report Recall@100 in the Appendix for reference. Similarly, for Evidence extraction, we reported nDCG@3 as the primary metric and also reported Recall@3 in the Appendix, given the reduced search space for this task.

Claim Labeling Evaluation: The claim-labeling task asks LLMs to label each document pair and its associated claims as SUPPORTS, REFUTES, or NOT_ENOUGH_INFORMATION. In this work, this evaluation is extended by presenting labeled evidence or balanced few-shot examples. For evidence-based evaluations, we have modified a prompt to guide the LLM to focus on the evidence text extracted from the document, and for few-shot examples, we present three distinct examples with different labels. Evaluation details and reproducibility are explained in Appendix G.

Language Models: In this work, we evaluate a diverse set of both embedding-based and lexical retrieval models. For information and evidence extraction, we use models that generate semantically meaningful sentence embeddings. These embeddings can be compared using similarity measures, such as cosine similarity, to rank candidate sentences with respect to a given claim (Reimers and Gurevych, 2019).

For embedding-based retrieval, we evaluate several multilingual and instruction-tuned models, including Qwen embeddings (Qwen3-Embedding-{0.6B, 4B, 8B}) (Zhang et al., 2025b), multilingual-e5-large (Wang et al., 2024), and AfriE5-Large-instruct (Uemura et al., 2025).

For the task of claim classification, we evaluate several LLMs from multiple model families, including Gemma (Team, 2025a) (Gemma-3-12B & 27B-it), Qwen (Team, 2025b) (Qwen-3-14B), Cohere(Command R v01, Tiny-Aya-Global), and AfriqueLLM (Yu et al., 2026) (AfriqueGemma-12B, AfriqueQwen-14B). Almost all models in the main

[2] In detail, we include documents from AfriDocMT, XL-Sum news, and manually collected, language-specific Wikipedia articles. In addition, we create language-specific retrieval collections using documents exclusively drawn from each corresponding language.

[3] relevant documents that are not in the same language as query are removed from both the retrieval results and labels during evaluation.

[4] Meanwhile, Appendix C provides more evaluation results when retrieving from a corpus containing documents from a single language or a single domain.

model	English	Amharic	Hausa	Igbo	Oromo	Shona	Swahili	Twi	Wolof	Yoruba	Zulu	A-Avg.	Δ
AfrIFact-Health (monolingual eval)													
multilingual-e5-large	0.465	0.383	0.231	0.207	0.387	0.189	0.236	0.241	0.216	0.156	0.254	0.250	-0.215
multilingual-e5-large-inst.	0.614	0.426	0.372	0.436	0.473	0.380	0.516	0.41	0.455	0.313	0.432	0.421	-0.193
AfriE5-Large-instruct	0.512	0.421	0.340	0.323	0.358	0.354	0.456	0.328	0.346	0.278	0.379	0.358	-0.154
Qwen3-Embedding-0.6B	0.506	0.284	0.161	0.250	0.324	0.195	0.308	0.259	0.314	0.128	0.194	0.242	-0.264
Qwen3-Embedding-4B	0.501	0.323	0.229	0.251	0.299	0.209	0.353	0.303	0.362	0.181	0.250	0.276	-0.225
Qwen3-Embedding-8B	0.534	0.368	0.192	0.272	0.341	0.246	0.387	0.333	0.343	0.216	0.278	0.298	-0.236
AfrIFact-Health (multilingual eval)													
multilingual-e5-large	0.149	0.123	0.113	0.119	0.138	0.141	0.189	0.107	0.105	0.133	0.171	0.134	-0.015
multilingual-e5-large-inst.	0.229	0.094	0.092	0.118	0.133	0.116	0.148	0.105	0.115	0.085	0.123	0.113	-0.116
AfriE5-Large-instruct	0.298	0.235	0.221	0.248	0.215	0.224	0.287	0.183	0.173	0.172	0.257	0.222	-0.077
Qwen3-Embedding-0.6B	0.171	0.065	0.070	0.106	0.111	0.073	0.112	0.098	0.111	0.081	0.082	0.091	-0.080
Qwen3-Embedding-4B	0.184	0.087	0.077	0.100	0.103	0.073	0.132	0.103	0.126	0.095	0.102	0.100	-0.084
Qwen3-Embedding-8B	0.207	0.123	0.077	0.109	0.115	0.097	0.158	0.115	0.126	0.104	0.117	0.114	-0.093
AfrIFact-Culture-News (monolingual eval)													
multilingual-e5-large	0.597	0.487	0.341	0.471	0.261	0.394	0.406	0.286	0.303	0.173	0.329	0.345	-0.252
multilingual-e5-large-inst.	0.735	0.622	0.567	0.633	0.462	0.576	0.533	0.446	0.554	0.483	0.445	0.532	-0.203
AfriE5-Large-instruct	0.752	0.631	0.544	0.623	0.380	0.513	0.530	0.408	0.543	0.359	0.368	0.490	-0.262
Qwen3-Embedding-0.6B	0.414	0.500	0.317	0.422	0.358	0.455	0.332	0.346	0.329	0.335	0.307	0.370	-0.044
Qwen3-Embedding-4B	0.428	0.544	0.382	0.520	0.361	0.492	0.386	0.359	0.340	0.426	0.314	0.412	-0.016
Qwen3-Embedding-8B	0.426	0.563	0.347	0.457	0.354	0.428	0.377	0.330	0.342	0.398	0.344	0.394	-0.032

Table 4: nDCG@10 of Information Retrieval Scores on Multilingual Corpus; **A-Avg.** column shows average score on African languages (i.e., excluding English); Δ column shows the score gap between **A-Avg.** English scores. Grey marks on best A-Avg. scores.

work, with the exception of `Tiny-Aya-Global`, have more than 128k context length to eliminate the effect of context size on performance. Additional models are included in the Appendix with lower context sizes. In addition, we include the closed-source model `GPT-5.2` for comparison. Appendix includes more model coverage, while the main paper focuses on a small subset.

Few-shot and Fine-tuning Settings: For zero-shot settings, we evaluated open LLMs using 3 different prompts, as shown in the appendix, and reported the average scores. For `GPT-5.2`, we used a single prompt to reduce cost.

For a few-shot experiment, we selected 3 random shots from the validation set, each with a balanced class label. We limited the number of shots to 3, given the evaluation size and the context length.

For the Finetuning experiment, we created an alpaca-style (Taori et al., 2023) version of our training dataset by converting it to an instruction input-output format, and we have both evidence- and no-evidence versions to double our dataset size. We worked on LoRA (Hu et al., 2022), QLoRA (Dettmers et al., 2023), and full fine-tuning settings on `AfriqueQwen-14B` for reasons discussed in the results section. Details of reproducibility are explored under Appendix G.

5 Results and Analysis

5.1 Information Retrieval Capability

Table 4 reports our main information retrieval results on a unified multilingual multi-domain corpus. *AfrIFact-Health* is evaluated on both monolingual and multilingual settings, as parallel relevant documents are available across all languages.[5] *AfrIFact-Culture-News* is evaluated only on the monolingual setting.

Cross-domain comparison on monolingual retrieval. Given that the target corpus is identical, we can compare monolingual results across domains. **we found that embedding models have different behavior across-domains.** where e5-based embeddings (i.e., top-3 rows) are better on *AfrIFact-Culture-News*, and Qwen3-based embeddings (bottom-3 rows) are better on *AfrIFact-Health*. Yet on both domains, a clear performance gap remains between African languages and English; for example, all models show gaps of more than 0.15 nDCG@10 on *AfrIFact-Health*.

Language bias dominates multilingual retrieval. Models' limitations on African languages become even more pronounced under multilingual evaluation — where even the best embedding models achieve only 0.298 nDCG@10 when retrieving from English, and 0.222 when retrieving from African languages on average, which is only half the best scores on their monolingual counterparts. This suggests that current models still struggle to rank relevant documents in other languages above the non-relevant documents in the query language, i.e., the strong cross-lingual alignment (Roy et al.,

[5] For monolingual settings, relevant documents that are not in the same language as the query are removed from both retrieval results and ground truths.

model	English	Amharic	Hausa	Igbo	Oromo	Shona	Swahili	Twi	Wolof	Yoruba	Zulu	A-Avg.	Δ
AfrIFact-Health													
multilingual-e5-large	0.770	0.363	0.415	0.349	0.386	0.450	0.541	0.333	0.418	0.288	0.438	0.398	-0.372
multilingual-e5-large-inst.	0.724	0.346	0.403	0.346	0.379	0.447	0.517	0.341	0.412	0.269	0.420	0.388	-0.336
AfriE5-Large-instruct	0.742	0.371	0.419	0.361	0.395	0.455	0.527	0.363	0.443	0.290	0.442	0.407	-0.335
Qwen3-Embedding-0.6B	0.782	0.305	0.370	0.275	0.340	0.368	0.474	0.298	0.397	0.244	0.362	0.343	-0.439
Qwen3-Embedding-4B	0.788	0.325	0.371	0.295	0.339	0.373	0.510	0.305	0.405	0.249	0.381	0.355	-0.433
Qwen3-Embedding-8B	0.785	0.349	0.373	0.320	0.357	0.404	0.530	0.284	0.405	0.270	0.404	0.370	-0.416
AfrIFact-Culture-News													
multilingual-e5-large	0.725	0.610	0.518	0.545	0.505	0.751	0.558	0.447	0.404	0.353	0.352	0.504	-0.221
multilingual-e5-large-inst.	0.685	0.619	0.510	0.567	0.539	0.760	0.554	0.426	0.438	0.366	0.350	0.513	-0.172
AfriE5-Large-instruct	0.701	0.618	0.528	0.577	0.536	0.765	0.583	0.441	0.436	0.377	0.352	0.521	-0.179
Qwen3-Embedding-0.6B	0.747	0.581	0.471	0.539	0.505	0.759	0.529	0.434	0.435	0.340	0.336	0.493	-0.254
Qwen3-Embedding-4B	0.752	0.586	0.488	0.559	0.498	0.768	0.565	0.425	0.436	0.352	0.332	0.501	-0.252
Qwen3-Embedding-8B	0.759	0.608	0.511	0.569	0.506	0.758	0.567	0.411	0.443	0.360	0.345	0.508	-0.251

Table 5: nDCG@3 evidence extraction performances across languages and domains in AfrIFact. **A-Avg.** The column shows the average score on African languages (i.e., excluding English); Δ column shows the score gap between **A-Avg.** English scores. Grey shows the best A-Avg. scores.

2020).

5.2 Evidence Extraction Capability

In Table 5, we present evidence extraction scores and nDCG scores using the top 3 pieces of evidence extracted from a single document presented. We observe a large gap between English and African languages, except in the case of Shona, for the health domain, with around 0.76 nDCG@3 scores.

English African Language evidence retrieval gap is larger on health compared to culture-news. The Evidence extraction gap between English and African languages is particularly pronounced in the Health domain compared to the Culture and News domains. In *AfrIFact-Health*, the top-performing English score reaches 0.79, whereas the best-performing African language, Swahili, achieves only 0.52, indicating a substantial disparity. In contrast, the gap narrows in *AfrIFact-Culture-News*, where English attains 0.76, and Shona reaches 0.80, even slightly surpassing English. This can be attributed not only to the availability of Wikipedia and cultural domains on the web, but also to a lack of cultural data (i.e., the motivation behind our various dataset-creation strategies). Evidence extraction tasks achieves better performances in *AfrIFact-Culture-News* (0.52-0.56 nDCG@3) compared to results in *AfrIFact-Health* (0.39-0.44 nDCG@3).

Increasing embedding model size is not as important as language coverage. Despite the substantial parameter differences between models such as `Qwen3-Embedding-0.6B` and `Qwen3-Embedding-8B`, the evidence extraction scores remain very similar. In contrast, `multilingual-e5-large` and `AfriE5-Large-instruct` achieved much better improvement due to their better language coverage, and the latter, fully adapted to African languages. Performance ranges only from 0.39 to 0.41 in the health domain and from 0.52 to 0.53 in the culture domain. This suggests that scaling model size alone yields limited gains, and that performance differences are more likely driven by dataset characteristics than by model capacity.

5.3 Fact Checking Capability

In Figure 3, we report African language average scores for models covered. We present $document+claim$, or $document + claim + evidence$ evaluations. Additionally, we show variants of both evaluations by pre-setting 3 balanced few-shot examples aligned with the evaluations. Finally, we include the closed-source model `GPT-5.2` as a closed-source comparison. Details in language scores are shown in Appendix E.

Zero-shot fact verification remains extremely challenging across languages. Mid-size open models show near-random fact checking capability. While larger models like `Gemma-3-27B-it` score close to 60% accuracy, most models achieve around 33% average accuracy on African languages in zero-shot evaluations. Large models such as `Gemma-3-27B-it`and `Llama-3-70B-instruct`achieve strong average scores among open-source models. However, improvements from increased model size are often modest compared to gains from few-shot prompting.

Few-shot dominates evidence effects. Across both health and culture-news domains, few-shot examples help improve models'fact-checking ca-

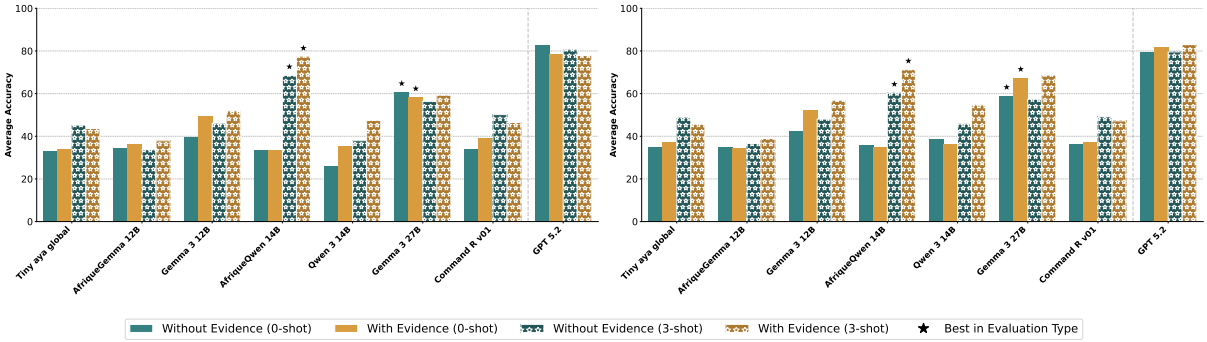


Figure 3: Average accuracy scores on African languages of different language models on the AfrIFact fact-checking task under zero-shot and three-shot settings, with and without evidence.

pability. models like `AfriqueQwen-14B`, `Command R v01`, and `Tiny-Aya-Global` show improved results in few-shot compared to zero-shot. We see significant improvement with few-shot examples, with the biggest being by 43% from 34% to 77% accuracy for `AfriqueQwen-14B`.

Providing evidence does not consistently improve fact-checking accuracy. Across many models, the difference between with-evidence and without-evidence settings is small or inconsistent, suggesting that models often rely more on their internal knowledge than on explicitly provided evidence.

Approach	*AfrIFact-Health* W0	*AfrIFact-Health* W	*AfrIFact-Culture-News* W0	*AfrIFact-Culture-News* W
Baseline	33.43	33.24	35.30	33.50
LoRA	45.95	56.53	42.55	43.54
QLoRA 8-bit	47.15	59.99	49.24	49.71
QLoRA 4-bit	36.30	45.25	36.04	38.67
Full finetuning	33.31	33.25	35.12	35.14

Table 6: Average Accuracy for African languages finetuning experiments on AfrIFact training on Baseline(`AfriqueQwen-14B`). **WO** is Without evidence provided and **W** is with evidence. Grey marks on the best Average scores.

A small scale dataset, coupled with parameter efficient finetuning, can help improve the performance of models. Table 6 shows usage of our multilingual training dataset on a promising model(`AfriqueQwen-14B`), given its few-shot performance, and we wanted to improve its performance using our small scale dataset. All training configurations except full finetuning seem to boost fact-checking classification. Specifically, QLoRA 8-bit fine-tuning shows the biggest improvement for the *AfrIFact-Health* domain, with a 26% average accuracy improvement across African languages.

Low-rank finetuned models learn to use evidence, shown by all LoRA and QLoRA with evidence models surpassing those without. In our benchmarking experiments, we show that average African language scores with and without evidence are either very similar or that evidence reduces performance, but all LoRA and QLoRA finetuning show that models learn to use the provided evidence span and improve scores with evidence present. QLoRA 8-bit quantization shows the best improvement, with evidence of a 12.84% improvement in the model.

6 Conclusion and Future Works

In this work, we introduce **AfrIFact**, a dataset for information retrieval, evidence retrieval, and fact-checking, comprising more than 18,000 human-generated claims, along with annotated source documents and evidence spans across 10 African languages. Our results show a large gap between information retrieval and evidence extraction scores, even when relevant evidence is found in other languages (i.e., the *AfrIFact-Health* multilingual experiment), and work done to create Afri-centric models shows promising results. Additionally, we show that open models lack 0-shot fact-verification capabilities, which can be improved with a few-shot examples or dedicated fine-tuning.

We believe this benchmark reveals systematic gaps and challenges in culture-specific information retrieval, evidence retrieval, and fact verification, and encourages researchers to address each of the three problems by expanding dataset coverage and creating robust systems that aim to eliminate misinformation.

Limitations

Language inclusion in this dataset was guided by several factors, including the number of speakers, availability of web resources, typological and geographical diversity, coverage in prior datasets, and the availability of annotators.

AfrIFact focuses on the health and culture-news domains because of their societal importance and exposure to misinformation. However, the need for fact-checking extends far, beyond these areas, expanding AfrIFact to cover other additional domains would offer a more comprehensive assessment of multilingual fact-checking capabilities.

Although the paper discusses a full pipeline for verifying claims by searching documents, extracting evidence spans, and classifying veracity, we treat each step as a standalone section and evaluate it on our dataset. This design allows us to isolate the challenges of each task, which are often studied as separate research problems. Future work can explore end-to-end pipelines that integrate all stages and evaluate their joint performance.

The focus of this work is on creating a substantially larger test set to support robust, high-quality, human-generated, and cross-lingual evaluation, and to accommodate automated fact-checking training datasets. This design choice forced us to create a smaller training and validation set. We intentionally made this choice due to resource constraints and to leave a research gap of fine-tuning better fact verification and extraction models for other works.

Fine-tuning experiments in this work overlook crucial steps such as hyperparameter optimization, data augmentation, and in-depth analysis of trained models, as the scope of this paper is limited to the introduction of dataset creation and benchmarking.

7 Ethical Consideration

All annotators, language coordinators, and meta-annotators were compensated fairly in accordance with the agreement(Appendix B). Annotators signed contracts specifying the purpose of the data collection and their compensation. We ensured that working conditions were fair and that annotators could skip documents they found unsuitable without penalty.

The source data for AfrIFact is collected from publicly available resources: AfriDocMT (health documents), Wikipedia (cultural content), and XL-Sum (news articles). No private or personally identifiable information has been collected. The claims were generated by annotators from public data.

References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, and 1 others. 2025. Irokobench: A new benchmark for african languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757.

Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, and Jimmy Lin. 2023. Ciral at fire 2023: Cross-lingual information retrieval for african languages. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 4–6.

Jesujoba Oluwadara Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025. AFRIDOC-MT: Document-level MT corpus for African languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27770–27806, Suzhou, China. Association for Computational Linguistics.

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.

Israel Abebe Azime, Tadesse Destaw Belay, Dietrich Klakow, Philipp Slusallek, and Anshuman Chhabra. 2025. Bridging the Culture Gap: A Framework for LLM-Driven Socio-Cultural Localization of Math Word Problems in Low-Resource Languages. *arXiv preprint arXiv:2508.14913*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Yi-Chen Chang, Canasai Kruengkrai, and Junichi Yamagishi. 2023. XFEVER: Exploring fact verification

across languages. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 1–11, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Meseret D. Gebremichael and Jason W. Jackson. 2006. Bridging the gap in Sub-Saharan Africa: A holistic look at information poverty and the region's digital divide. *Government Information Quarterly*, 23(2):267–280.

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. 2021. 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.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced African languages? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12979–12998, Suzhou, China. Association for Computational Linguistics.

Shuliang Liu, Hongyi Liu, Aiwei Liu, Duan Bingchen, Zheng Qi, Yibo Yan, He Geng, Peijie Jiang, Jia Liu, and Xuming Hu. 2025. A survey on proactive defense strategies against misinformation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18144–18155, Vienna, Austria. Association for Computational Linguistics.

Ndapandula Nakashole and Tom M. Mitchell. 2014. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019, Baltimore, Maryland. Association for Computational Linguistics.

Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh, and Jimmy Lin. 2022. AfriCLIRMatrix: Enabling cross-lingual information retrieval for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8721–8728, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rrubaa Panchendrarajan, Rubén Míguez Pérez, and Arkaitz Zubiaga. 2025. MultiClaimNet: A massively multilingual dataset of fact-checked claim clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11203–11215, Suzhou, China. Association for Computational Linguistics.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gemma Team. 2025a. Gemma 3.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Kosei Uemura, Miaoran Zhang, and David Ifeoluwa Adelani. 2025. Afrimteb and afrie5: Benchmarking and adapting text embedding models for african languages. *arXiv preprint arXiv:2510.23896*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

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*.

Hao Yu, Tianyi Xu, Michael A. Hedderich, Wassim Hamidouche, Syed Waqas Zamir, and David Ifeoluwa Adelani. 2026. Afriquellm: How data mixing and model architecture impact continued pre-training for african languages. *Preprint*, arXiv:2601.06395.

Hanzhi Zhang, Sumera Anjum, Heng Fan, Weijian Zheng, Yan Huang, and Yunhe Feng. 2025a. Polyfever: a multilingual fact verification benchmark for hallucination detection in large language models. *arXiv preprint arXiv:2503.16541*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

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, and Zheyan Luo. 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)*, pages 400–410.

A Data Statistics

AfrIFact maintains balanced training and validation sizes per language, while the test split is substantially larger to support robust, high-quality, human-generated, and cross-lingual evaluation, and to accommodate automated fact-checking training datasets. The languages included are Amharic (Ethiopia), Twi (Ghana), Hausa (Nigeria, Niger, Chad, Cameroon, and Benin), Igbo (Nigeria), Oromo (Ethiopia and Kenya), Shona (Zimbabwe and Mozambique), Swahili (East Africa and the Democratic Republic of Congo), Wolof (Senegal, Gambia, and Mauritania), Yoruba (Nigeria, Benin, and Togo), and isiZulu (South Africa, Lesotho, and Eswatini)

B Annotators, Structure, Recruitment And Payment

The annotation team consists of 30 annotators, 10 language coordinators, 3 meta-annotators, and 2 administrative assistants. For each language, three native speakers serve as annotators under the supervision of a dedicated language coordinator. The language coordinators have prior experience in data annotation and research and are responsible for guiding the annotation process and ensuring alignment and consistency of the outputs. The recruitment process involved appointing language coordinators based on prior experience in similar projects. Annotators signed contracts specifying the purpose of the data collection and the compensation provided. Everyone involved in the work received training in fact-checking and verification, along with video and textual guidelines for using the tools at each stage of the data annotation process. Before full participation, annotators completed trial tasks that were reviewed and verified by the language coordinators to ensure quality. Appropriate compensation was provided to all contributors involved in the project in accordance with international wage standards.

C Information Retrieval Results

Table 11 presents additional evaluation results on the *AfrIFact-Health* dataset as complement to Table 4. Specifically:

1. *Monolingual Corpus on Health*: Queries from each language only search for relevant documents from documents in the same language under the health domain, e.g., Amharic queries are only searched against Amharic documents.
2. *Multilingual Corpus on Health*: Queries in each language are searched against documents in all 11 languages. As documents are parallel across languages, now each query has 11 copies of relevant documents, which all contribute to the evaluation of retrieval results. All documents are still from the health domain.
3. *Multilingual Corpus on Health and Culture-News*: **[in main table]** Queries in each language are additionally searched against documents in all languages from culture-news domains. Evaluations stays the same as *Multilingual Corpus on Health*.
4. *Multilingual Corpus on Health and Culture-News (eval on monolingual ground-truth* : **[in main table]** Identical corpus as above, but exclude relevant documents from non-query-language in retrieval results and ground-truth during evaluation.

In the above settings, the retrieval corpus gradually moved from monolingual mono-domain to multilingual multi-domain, where settings 3 and 4 are used as the main evaluation reported in Table 4 so as to achieve a unified corpus with culture-news domains.

Comparing results from the above 4 settings, we have two major observations: First, **embedding models struggle with placing relevant documents from non-query languages on top of non-relevant documents from query languages**: Moving from setting 1 to 2, i.e., monolingual-health to multilingual-health, scores of models are at least cut by half if not smaller, and it is not better when using English queries.

Second, **embedding models are overall robust to additional non-relevant documents from other languages or domains**. Moving from 1 to 4 (corpus includes more non-relevant documents from other languages and domains), while some embeddings are heavily impacted (e.g., `multilingual-e5-large`), the scale is smaller than when introducing relevant documents from other languages as mentioned above. Moreover, some embedding models are barely affected by the "noise" document — `multilingual-e5-large-inst.` maintains the same effectiveness overall on both English and African languages (i.e., 0.629 v.s. 0.614 on English,

language	Amharic	English	Hausa	Igbo	Twi	Oromo	Shona	Swahili	Wolof	Yoruba	Zulu	Total
AfrIFact-Health												
SUPPORT	284	284	284	284	284	284	284	284	284	284	284	3124
REFUTES	283	283	283	283	283	283	283	283	283	283	283	3113
NEI	281	281	281	281	281	281	281	281	281	281	281	3091
Total	848	848	848	848	848	848	848	848	848	848	848	9328
AfrIFact-Culture-News												
SUPPORT	285	263	282	278	278	271	291	307	246	275	286	3062
REFUTES	253	267	213	272	272	271	212	211	256	269	225	2721
NEI	275	283	336	263	263	271	310	295	311	269	302	3178
TOTAL	810	810	828	810	810	810	810	810	810	810	810	8928
TOTAL	1658	1658	1676	1658	1658	1658	1658	1658	1658	1658	1658	18256

Table 7: Detailed distribution of annotated examples in the AfrIFact dataset across languages, domains, and verification labels. The table reports the number of instances labeled as `SUPPORTS`, `REFUTES`, and `NOT_ENOUGH_INFORMATION` (Not Enough Information) for each language in the *AfrIFact-Health* and *AfrIFact-Culture-News* domains.

Language	Class	Fleiss' Kappa	Cohen's Kappa (avg)	Krippendorff's Alpha
Amharic	REFUTES	0.80	0.80	0.80
	NOT_ENOUGH_INFORMATION	0.88	0.88	0.88
	SUPPORTS	0.88	0.88	0.88
Yoruba	REFUTES	0.92	0.92	0.92
	NOT_ENOUGH_INFORMATION	0.94	0.94	0.94
	SUPPORTS	0.96	0.96	0.96
igbo	REFUTES	0.67	0.67	0.67
	NOT_ENOUGH_INFORMATION	0.76	0.76	0.76
	SUPPORTS	0.83	0.83	0.83
Swahili	REFUTES	0.82	0.82	0.82
	NOT_ENOUGH_INFORMATION	0.82	0.83	0.83
	SUPPORTS	0.85	0.85	0.85
oromo	REFUTES	0.87	0.87	0.87
	NOT_ENOUGH_INFORMATION	0.94	0.94	0.94
	SUPPORTS	0.87	0.87	0.87
Hausa	REFUTES	0.74	0.75	0.75
	NOT_ENOUGH_INFORMATION	0.80	0.81	0.81
	SUPPORTS	0.86	0.87	0.87
Shona	REFUTES	0.99	0.99	0.99
	NOT_ENOUGH_INFORMATION	0.96	0.96	0.96
	SUPPORTS	0.97	0.97	0.97
Wolof	REFUTES	0.79	0.79	0.79
	NOT_ENOUGH_INFORMATION	0.78	0.78	0.78
	SUPPORTS	0.82	0.82	0.82
English	REFUTES	0.84	0.84	0.84
	NOT_ENOUGH_INFORMATION	0.72	0.72	0.72
	SUPPORTS	0.78	0.78	0.78
Twi	REFUTES	0.74	0.74	0.74
	NOT_ENOUGH_INFORMATION	0.79	0.79	0.79
	SUPPORTS	0.84	0.84	0.84
Zulu	REFUTES	0.84	0.84	0.84
	NOT_ENOUGH_INFORMATION	0.72	0.72	0.72
	SUPPORTS	0.77	0.78	0.77

Table 8: Inter-Annotator Agreement Metrics by Language and Category for *AfrIFact-Culture-News*, showing substantial to great agreement score.

Language	Country	Gender Distribution (LC/Annotator)	Qualifications	Resides in Country
Amharic	Ethiopia	M / 2M1F	Linguistics, CS, Translation	✓
Hausa	Nigeria	M / 2M1F	Linguistics and translation	✓
Igbo	Nigeria	F / 1M2F	Language studies	✓
Twi	Ghana	F / 1M2F	Linguistics, education	✓
Oromo	Ethiopia	F/ 2F1M	Linguistics and journalism	✓
Shona	Zimbabwe	F / 2M1F	Language studies and translation	✓
Swahili	Kenya	M / 2M1F	Linguistics, education	✓
Wolof	Senegal	M / 2M1F	Linguistics and translation	✓
Yoruba	Nigeria	F / 1M2F	Linguistics, NLP annotation	✓
Zulu	South Africa	F / 1M2F	Linguistics and education	✓

Table 10: Summary of annotator teams involved in the creation of the AfriFact dataset. For each language, we report the country context, gender of the language coordinator (LC), gender distribution of the three annotators, a summary of annotator qualifications, and whether annotators reside in the corresponding language region.

0.424 v.s. 0.421 on average on African languages). The same observation is made moving from 2 to 3 (a multilingual corpus includes more non-relevant documents from other domains), where the overall

scores are not impacted.

D Evidence Extraction Detailed Metrics

Table 12 shows recall@3 alternative metric scores for evidence retrieval from a single document.

E Fact Classification Detailed Scores

In Table 14, we observe that providing labeled evidence does not consistently improve fact-checking accuracy. Many models collapse to near-random performance in the zero-shot setting, indicating difficulty in understanding the task without demonstrations. Few-shot examples often yield dramatic improvements of up to 43 points for models such as `AfriqueQwen-14B`. Increasing model size generally improves performance, but gains plateau beyond mid-scale models. English consistently achieves higher scores than African languages, highlighting a persistent multilingual performance gap. Languages such as Wolof, Twi, and Yoruba show lower scores across most models, while `GPT-5.2` demonstrates substantially higher performance across all evaluation settings.

Finetuning experiments in Table 15 and Table 16, most models learn to use evidence and improve on the evidence span based evaluation. Providing evidence substantially improves fact verification in the Health domain but has mixed effects in culture–news, while QLoRA-8 consistently outperforms other adaptation strategies, reaching 58.77% in the culture-news domain and 65.57% in the health domain in some languages. Given the consistent performance trends across all languages in model training with a fixed 3-epoch setup and similar hyperparameters, this demonstrates the effectiveness of our multilingual data.

F Translation Verification Tasks

In *AfrIFact-Health*, all claim generation, evidence extraction, and labeling are performed in English. This design choice offers three main advantages. First, it ensures that new annotators can more easily understand and perform the task. Second, given the availability of sentence-level translations in AfriDocMT, evidence spans annotated in English can be directly translated into other languages, significantly reducing the human effort required to transfer evidence across languages and helping us cover the healthcare domain, which lacks sufficient native resources. Figure 6 shows translation qualities and modifications and corrections made by translators.

G Setup and Reproducibility

G.1 Evaluation

We conducted all evaluations using EleutherAI's open-source Language Model Evaluation Harness (`lm-eval`) (Gao et al., 2024). The framework supports multiple evaluation strategies, including log-likelihood, with experiments configured and managed via configuration files. For open-source models, we employed log-likelihood and generation-based evaluation for multiple-choice tasks as shown in (Adelani et al., 2025). In multiple-choice, each candidate option is appended to the corresponding question prompt, after which the log-likelihood is computed. Model accuracy is then reported based on the option with the highest log-likelihood score. For the `GPT-5.2`model, we created the same prompts and used batched API generation to reduce costs by selecting only the best prompt. Evaluation costs are close to 400 USD for GPT models across all variants. All experiments use temperature=0, top_p=1.

G.2 Finetuning

To fine-tune LLLMs, we leveraged **LLaMA-Factory** (Zheng et al., 2024b), an open-source framework that supports efficient fine-tuning through parameter-efficient methods and scalable training pipelines. We performed LoRA-based supervised fine-tuning on our dataset using one H100 GPU. Our Training split was converted to the Alpaca instruction format and will be shared as part of our released resources. Training was conducted for three epochs using a LoRA rank of 8 and an effective batch size of 8 via gradient accumulation. For the QLoRA trainings we did, we kept the same hyper parameters with LORA and modified it to have 8 and 4-bit bits and bytes quantization. For full finetuning, we leveraged LLaMA-Factory, configurations with the same 3 epochs across 4 H100 GPUs. Detailed training configuration will be released in our GitHub training code examples. A systematic exploration and optimization of hyperparameters, including LoRA configuration and training dynamics, is left for future work.

Figure 4: **Why is evidence not helping improve accuracy?** Evidence introduces a conservative shift in model predictions: in Health, it reduces hallucinated `SUPPORTS` but increases `NOT_ENOUGH_INFORMATION` predictions, while in Culture, it significantly improves NEI detection by reducing false SUPPORTS classifications. In the Health domain `NOT_ENOUGH_INFORMATION` and `SUPPORTS` improve while in the culture domain, the model tends to be good at identifying `NOT_ENOUGH_INFORMATION` when evidence is provided. In both domains, the model confusion between `SUPPORTS` and `REFUTES` improves with evidence provided.

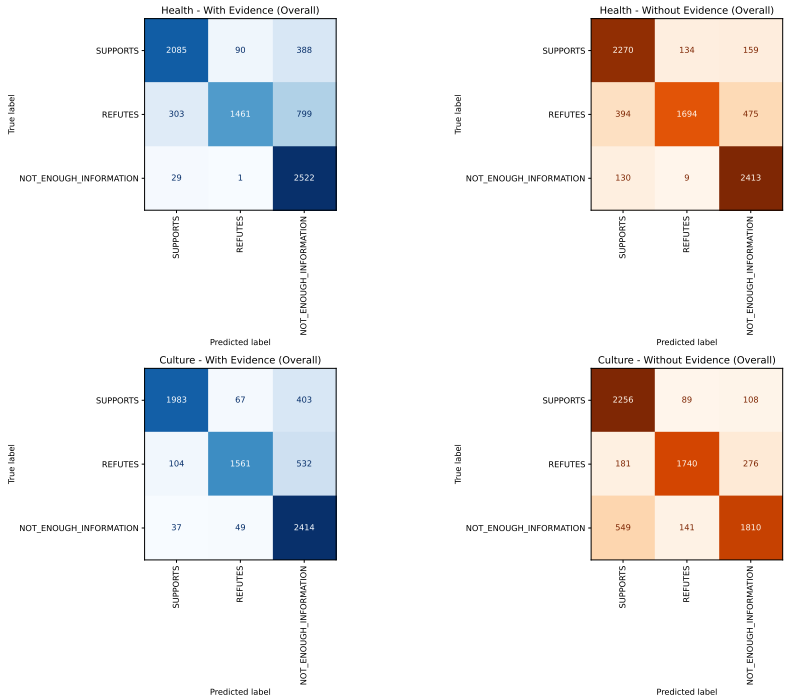
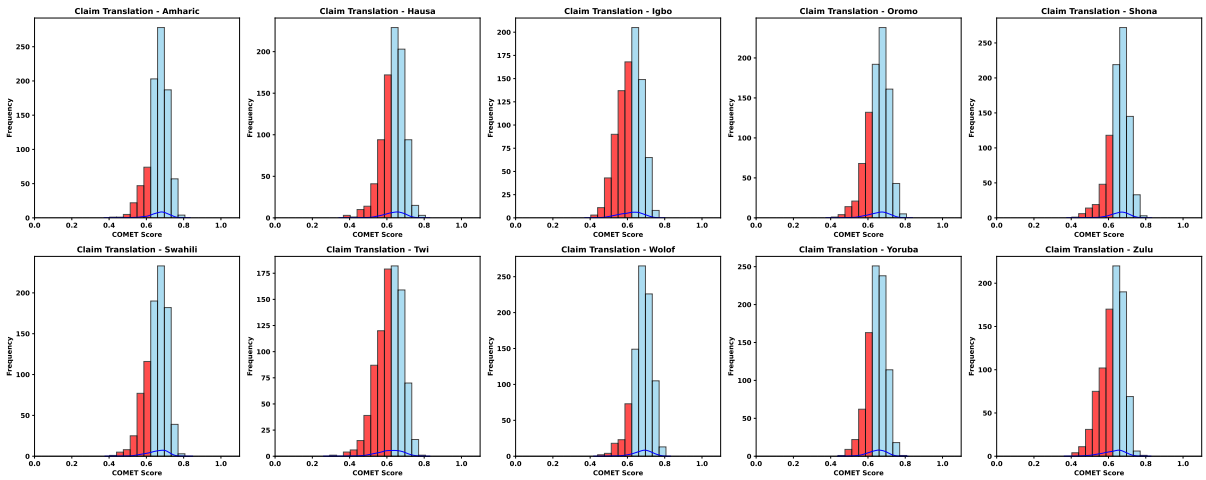


Figure 5: Distribution of COMET scores for claim and document translations across African languages. The upper grid shows histograms of COMET scores for claim translations across ten languages, while the lower grid presents score distributions for document translations in five languages (Igbo, Oromo, Shona, Twi, and Wolof). This fixed or false alarm percentages items that has below 0.6 COMET scores colored red and addressed by annotators.



Model	English	Amharic	Hausa	Igbo	Oromo	Shona	Swahili	Twi	Wolof	Yoruba	Zulu	A-Avg.	Δ
Monolingual Corpus on Health													
multilingual-e5-large	0.524	0.422	0.366	0.385	0.444	0.354	0.371	0.323	0.348	0.259	0.387	0.366	-0.158
multilingual-e5-large-instruct	0.629	0.426	0.372	0.438	0.476	0.384	0.517	0.420	0.464	0.313	0.433	0.424	-0.205
AfriE5-Large-instruct	0.593	0.431	0.394	0.431	0.449	0.412	0.525	0.391	0.408	0.301	0.424	0.417	-0.176
Qwen3-Embedding-0.6B	0.558	0.286	0.169	0.267	0.360	0.206	0.326	0.282	0.336	0.150	0.204	0.259	-0.299
Qwen3-Embedding-4B	0.546	0.326	0.233	0.274	0.341	0.212	0.363	0.325	0.394	0.208	0.262	0.294	-0.252
Qwen3-Embedding-8B	0.582	0.373	0.208	0.293	0.385	0.258	0.403	0.354	0.386	0.234	0.285	0.318	-0.264
Multilingual Corpus on Health													
multilingual-e5-large	0.200	0.149	0.145	0.176	0.141	0.147	0.192	0.114	0.114	0.136	0.174	0.149	-0.051
multilingual-e5-large-instruct	0.234	0.094	0.092	0.118	0.134	0.117	0.149	0.109	0.116	0.085	0.123	0.114	-0.120
AfriE5-Large-instruct	0.299	0.237	0.222	0.250	0.215	0.227	0.287	0.186	0.174	0.172	0.258	0.223	-0.076
Qwen3-Embedding-0.6B	0.175	0.065	0.071	0.107	0.118	0.075	0.115	0.102	0.115	0.081	0.083	0.093	-0.082
Qwen3-Embedding-4B	0.186	0.088	0.077	0.101	0.111	0.075	0.133	0.109	0.132	0.096	0.105	0.103	-0.083
Qwen3-Embedding-8B	0.212	0.124	0.081	0.111	0.121	0.099	0.159	0.120	0.132	0.104	0.118	0.117	-0.095
Multilingual Corpus on Health and Culture-News													
multilingual-e5-large	0.149	0.123	0.113	0.119	0.138	0.141	0.189	0.107	0.105	0.133	0.171	0.134	-0.015
multilingual-e5-large-inst.	0.229	0.094	0.092	0.118	0.133	0.116	0.148	0.105	0.115	0.085	0.123	0.113	-0.116
AfriE5-Large-instruct	0.298	0.235	0.221	0.248	0.215	0.224	0.287	0.183	0.173	0.172	0.257	0.222	-0.077
Qwen3-Embedding-0.6B	0.171	0.065	0.070	0.106	0.111	0.073	0.112	0.098	0.111	0.081	0.082	0.091	-0.080
Qwen3-Embedding-4B	0.184	0.087	0.077	0.100	0.103	0.073	0.132	0.103	0.126	0.095	0.102	0.100	-0.084
Qwen3-Embedding-8B	0.207	0.123	0.077	0.109	0.115	0.097	0.158	0.115	0.126	0.104	0.117	0.114	-0.093
Multilingual Corpus on Health and Culture-News (eval on monolingual ground-truth)													
multilingual-e5-large	0.465	0.383	0.231	0.207	0.387	0.189	0.236	0.241	0.216	0.156	0.254	0.250	-0.215
multilingual-e5-large-inst.	0.614	0.426	0.372	0.436	0.473	0.380	0.516	0.41	0.455	0.313	0.432	0.421	-0.193
AfriE5-Large-instruct	0.512	0.421	0.340	0.323	0.358	0.354	0.456	0.328	0.346	0.278	0.379	0.358	-0.154
Qwen3-Embedding-0.6B	0.506	0.284	0.161	0.250	0.324	0.195	0.308	0.259	0.314	0.128	0.194	0.242	-0.264
Qwen3-Embedding-4B	0.501	0.323	0.229	0.251	0.299	0.209	0.353	0.303	0.362	0.181	0.250	0.276	-0.225
Qwen3-Embedding-8B	0.534	0.368	0.192	0.272	0.341	0.246	0.387	0.333	0.343	0.216	0.278	0.298	-0.236

Table 11: ***AfrIFact-Health*** nDCG@10 scores when retrieving queries from different corpus and evaluated on corresponding qrels.

Table 12: AfrIFact Recall@3 Evidence Extraction Scores.

model	English	Amharic	Hausa	Igbo	Oromo	Shona	Swahili	Twi	Wolof	Yoruba	Zulu	A-Avg.	Δ
AfrIFact-Health													
multilingual-e5-large	0.820	0.408	0.460	0.403	0.439	0.509	0.586	0.396	0.484	0.325	0.495	0.450	-0.369
multilingual-e5-large-inst.	0.822	0.380	0.455	0.401	0.422	0.507	0.572	0.398	0.473	0.317	0.478	0.440	-0.382
AfriE5-Large-instruct	0.822	0.412	0.464	0.412	0.437	0.513	0.574	0.409	0.506	0.337	0.495	0.456	-0.366
Qwen3-Embedding-0.6B	0.823	0.355	0.444	0.344	0.402	0.444	0.548	0.350	0.470	0.300	0.417	0.407	-0.416
Qwen3-Embedding-4B	0.825	0.373	0.443	0.349	0.395	0.458	0.567	0.347	0.480	0.294	0.440	0.415	-0.410
Qwen3-Embedding-8B	0.827	0.377	0.424	0.381	0.408	0.478	0.587	0.344	0.473	0.317	0.462	0.425	-0.402
AfrIFact-Culture-News													
multilingual-e5-large	0.807	0.637	0.589	0.606	0.560	0.808	0.616	0.512	0.474	0.391	0.390	0.558	-0.248
multilingual-e5-large-inst.	0.790	0.644	0.572	0.604	0.595	0.805	0.616	0.491	0.489	0.419	0.394	0.563	-0.227
AfriE5-Large-instruct	0.797	0.639	0.585	0.632	0.595	0.808	0.638	0.498	0.481	0.423	0.395	0.569	-0.228
Qwen3-Embedding-0.6B	0.825	0.613	0.552	0.591	0.570	0.801	0.603	0.495	0.484	0.391	0.379	0.548	-0.277
Qwen3-Embedding-4B	0.848	0.623	0.572	0.618	0.573	0.810	0.638	0.483	0.479	0.396	0.366	0.556	-0.292
Qwen3-Embedding-8B	0.836	0.634	0.581	0.627	0.571	0.803	0.638	0.473	0.484	0.410	0.382	0.560	-0.276

Figure 6: Distribution of COMET scores for document translations across African languages. This fixed or false alarm percentages items that has below 0.6 COMET scores colored red and addressed by annotators.

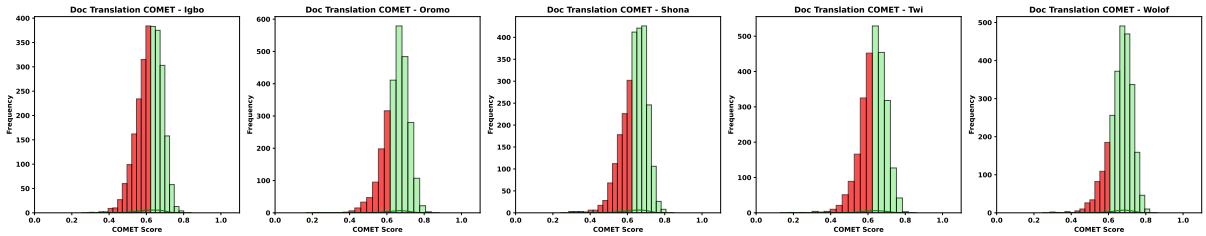


Table 13: ***AfrIFact-Health*** 0-shot/3-shot Accuracy Scores. The best open source models are Bold, and the best overall models are grayed.

Model	Amharic	English	Hausa	Igbo	Oromo	Shona	Swahili	Twi	Wolof	Yoruba	Zulu	A-Avg.
AfrIFact-Health With Evidence												
AfriqueGemma-12B	38.11/38.68	36.01/47.66	36.82/38.44	38.59/40.40	38.68/39.40	35.20/41.17	34.48/38.40	34.81/34.48	34.19/33.38	35.96/35.86	37.73/38.78	36.46/37.90
AfriqueQwen-14B	33.24/71.11	33.24/92.26	34.53/84.19	33.29/83.38	33.24/86.44	33.24/86.87	34.96/86.34	33.24/58.17	33.24/54.97	33.91/79.27	34.53/81.76	33.74/**77.25**
Command R v01	39.06/35.24	63.75/89.83	40.54/45.08	38.49/44.56	36.91/55.06	39.54/47.66	45.32/46.94	36.68/48.95	35.77/52.96	37.73/38.97	42.88/49.28	39.29/46.47
Gemma-3-12B-it	38.59/39.02	58.07/71.30	55.21/58.31	58.60/60.74	52.58/53.39	56.26/63.18	58.36/63.80	45.46/47.23	38.40/43.36	35.77/35.91	54.58/56.30	49.38/52.13
Gemma-3-1B-it	34.77/35.91	34.34/52.39	34.24/35.51	33.76/36.51	33.67/33.72	33.76/37.18	35.82/39.21	33.67/35.10	33.43/34.15	34.62/34.03	34.00/35.24	34.17/35.65
Gemma-3-27B-it	40.16/40.15	84.38/88.97	64.33/69.70	68.10/68.02	64.47/65.83	73.93/75.82	72.87/78.40	47.18/45.95	47.85/43.70	38.40/38.11	65.09/67.19	**58.24**/59.29
Gemma-3-4B-it	35.43/35.96	58.40/74.93	44.94/47.99	46.08/46.18	44.70/41.45	55.92/54.35	54.63/62.89	40.83/37.92	41.31/37.01	34.72/34.72	48.90/50.00	44.75/44.85
Llama-3-70B-instruct	49.90/42.17	79.13/89.68	60.70/73.97	63.51/73.69	61.27/69.58	56.06/65.52	67.72/86.44	57.55/59.22	53.77/60.46	50.48/43.74	57.50/67.14	57.85/64.19
Qwen-3-14B	36.29/63.56	34.67/73.59	33.52/45.56	33.81/43.70	35.82/40.69	34.57/44.56	39.97/55.44	33.24/45.89	35.48/41.98	34.48/42.31	35.82/51.00	35.30/47.47
Tiny-Aya-Global	33.24/44.70	33.38/68.15	34.05/42.36	33.24/47.80	33.24/36.96	33.29/51.43	37.63/48.90	33.24/34.91	33.24/45.80	34.19/36.87	34.72/46.56	34.01/43.63
GPT-5.2	80.95/78.94	83.24/84.10	78.22/77.65	79.51/79.94	78.65/78.65	80.80/79.80	82.52/80.23	75.64/76.36	73.93/74.50	78.51/77.08	77.36/75.79	78.61/77.89
AfrIFact-Health Without Evidence												
AfriqueGemma-12B	34.86/33.62	37.01/53.49	34.34/35.77	34.15/34.48	34.29/32.38	34.81/39.88	34.48/38.59	34.10/29.51	33.86/31.81	33.24/28.75	34.57/32.38	34.27/33.72
AfriqueQwen-14B	33.43/73.73	33.43/79.42	33.29/72.54	33.81/71.30	33.29/72.87	33.38/72.45	33.29/75.64	33.29/54.63	33.33/49.43	33.52/70.92	33.33/71.87	33.40/**68.54**
Command R v01	34.86/39.45	51.05/80.04	33.67/47.66	33.81/47.33	33.48/56.69	33.67/51.00	36.58/53.34	33.57/50.00	33.43/52.01	33.43/46.13	34.19/57.93	34.07/50.15
Gemma-3-12B-it	50.91/55.01	58.55/70.15	39.92/53.34	39.40/43.31	38.44/43.98	40.74/52.87	47.99/59.84	33.14/35.48	32.81/35.10	31.76/33.24	42.55/48.95	39.77/46.11
Gemma-3-1B-it	34.81/36.53	37.01/51.58	33.43/35.74	33.57/34.89	34.15/33.98	35.48/37.42	35.05/41.79	33.29/34.36	35.00/33.93	33.24/33.21	34.10/34.17	34.21/35.60
Gemma-3-27B-it	71.74/67.87	79.51/76.32	67.80/64.53	63.68/59.60	59.74/55.65	70.24/66.15	74.36/71.61	44.23/38.64	44.59/37.30	45.24/43.67	64.90/58.68	**60.65**/56.37
Gemma-3-4B-it	33.14/53.10	41.26/67.81	30.85/49.81	30.95/39.40	28.46/39.59	29.70/47.85	33.91/58.36	27.84/35.58	25.26/40.35	30.52/32.62	33.38/46.56	30.40/44.32
Llama-3-70B-instruct	51.09/50.86	70.95/79.37	56.50/68.39	55.53/64.66	47.68/60.08	48.51/57.69	61.72/75.17	47.74/51.58	42.75/52.63	45.07/47.18	54.50/59.89	51.11/58.81
Qwen-3-14B	29.99/34.19	28.61/60.98	25.74/39.68	24.21/37.15	28.13/37.01	25.07/38.63	27.79/40.74	24.83/36.25	27.08/39.06	25.93/35.77	22.78/40.83	26.16/37.93
Tiny-Aya-Global	33.29/50.96	33.24/62.23	33.24/47.23	33.19/46.90	33.29/37.82	33.29/49.71	33.29/53.15	33.29/34.96	33.29/41.55	33.05/43.51	33.19/46.47	33.24/45.22
GPT-5.2	86.96/83.38	88.25/86.82	82.95/80.37	83.67/81.38	82.52/80.95	85.10/81.66	84.81/82.23	79.23/78.94	77.36/78.08	82.09/79.37	80.66/79.08	82.54/80.54

Table 14: ***AfrIFact-Culture-News*** 0-shot/3-shot Accuracy Scores. The best open source models are Bold, and the best overall models are grayed.

Model	Amharic	English	Hausa	Igbo	Oromo	Shona	Swahili	Twi	Wolof	Yoruba	Zulu	A-Avg.
AfrIFact-Health With Evidence												
AfriqueGemma-12B	0.36/0.39	0.46/0.58	0.36/0.45	0.34/0.37	0.36/0.39	0.37/0.48	0.32/0.34	0.31/0.32	0.37/0.37	0.34/0.36	0.34/0.43	0.35/0.39
AfriqueQwen-14B	0.34/0.77	0.35/1.00	0.38/0.77	0.32/0.78	0.33/0.81	0.38/0.77	0.37/0.72	0.31/0.61	0.39/0.45	0.34/0.75	0.37/0.70	0.35/0.71
Command R v01	0.43/0.39	0.54/0.93	0.38/0.47	0.33/0.46	0.33/0.58	0.39/0.44	0.39/0.58	0.31/0.49	0.39/0.43	0.38/0.42	0.44/0.47	0.38/0.47
Gemma-3-12B-it	0.62/0.65	0.83/0.86	0.60/0.63	0.58/0.61	0.48/0.54	0.51/0.64	0.56/0.58	0.46/0.51	0.39/0.45	0.50/0.53	0.54/0.57	0.52/0.57
Gemma-3-1B-it	0.35/0.34	0.35/0.55	0.38/0.37	0.32/0.39	0.33/0.38	0.38/0.34	0.36/0.31	0.31/0.34	0.39/0.37	0.35/0.37	0.38/0.33	0.35/0.35
Gemma-3-27B-it	0.81/0.82	0.99/0.99	0.71/0.74	0.74/0.75	0.69/0.68	0.74/0.77	0.65/0.66	0.60/0.61	0.48/0.49	0.64/0.65	0.67/0.70	0.67/0.69
Gemma-3-4B-it	0.59/0.61	0.63/0.75	0.52/0.56	0.50/0.51	0.47/0.49	0.51/0.55	0.53/0.55	0.46/0.45	0.40/0.41	0.44/0.47	0.47/0.52	0.49/0.51
Qwen-3-14B	0.39/0.67	0.47/0.72	0.36/0.55	0.37/0.53	0.37/0.50	0.37/0.65	0.38/0.51	0.37/0.54	0.32/0.49	0.36/0.51	0.37/0.56	0.36/0.55
Tiny-Aya-Global	0.34/0.57	0.41/0.64	0.44/0.41	0.32/0.51	0.33/0.42	0.38/0.52	0.41/0.47	0.31/0.44	0.39/0.41	0.39/0.39	0.42/0.44	0.37/0.46
GPT-5.2	0.90/0.89	0.98/0.99	0.86/0.86	0.84/0.87	0.80/0.82	0.85/0.83	0.71/0.73	0.78/0.81	0.79/0.80	0.83/0.86	0.82/0.84	0.83/0.85
AfrIFact-Health Without Evidence												
AfriqueGemma-12B	0.34/0.35	0.34/0.34	0.40/0.41	0.31/0.33	0.33/0.34	0.37/0.45	0.36/0.38	0.31/0.33	0.38/0.37	0.32/0.34	0.38/0.38	0.35/0.37
AfriqueQwen-14B	0.35/0.66	0.33/0.87	0.40/0.67	0.33/0.67	0.34/0.65	0.38/0.68	0.37/0.58	0.31/0.53	0.39/0.38	0.32/0.60	0.38/0.63	0.36/0.60
Command R v01	0.38/0.42	0.44/0.81	0.40/0.48	0.33/0.53	0.33/0.58	0.39/0.50	0.39/0.55	0.31/0.51	0.39/0.40	0.32/0.48	0.39/0.50	0.36/0.49
Gemma-3-12B-it	0.41/0.45	0.45/0.51	0.50/0.56	0.39/0.43	0.39/0.46	0.48/0.58	0.47/0.52	0.37/0.42	0.39/0.41	0.36/0.42	0.48/0.55	0.42/0.48
Gemma-3-1B-it	0.34/0.36	0.34/0.44	0.40/0.35	0.32/0.37	0.33/0.38	0.38/0.38	0.37/0.30	0.31/0.34	0.39/0.35	0.33/0.38	0.38/0.34	0.35/0.35
Gemma-3-27B-it	0.60/0.61	0.67/0.66	0.68/0.66	0.59/0.57	0.60/0.57	0.66/0.65	0.65/0.63	0.51/0.51	0.44/0.40	0.54/0.53	0.63/0.63	0.59/0.57
Gemma-3-4B-it	0.39/0.43	0.43/0.49	0.47/0.53	0.37/0.41	0.35/0.40	0.41/0.50	0.44/0.52	0.35/0.38	0.39/0.40	0.35/0.42	0.42/0.50	0.39/0.45
Qwen-3-14B	0.38/0.42	0.48/0.54	0.42/0.49	0.39/0.46	0.38/0.47	0.39/0.53	0.44/0.43	0.37/0.48	0.34/0.40	0.38/0.44	0.38/0.48	0.39/0.46
Tiny-Aya-Global	0.34/0.55	0.34/0.51	0.40/0.47	0.32/0.52	0.33/0.46	0.38/0.54	0.37/0.53	0.31/0.45	0.39/0.41	0.32/0.48	0.38/0.49	0.35/0.49
GPT-5.2	0.84/0.83	0.98/0.98	0.82/0.83	0.85/0.86	0.87/0.88	0.74/0.74	0.71/0.72	0.92/0.90	0.71/0.70	0.81/0.81	0.69/0.69	0.81/0.81

Table 15: ***AfrIFact-Health*** without-evidence/with-evidence finetuning experiments

Model	Amharic	English	Hausa	Igbo	Oromo	Shona	Swahili	Twi	Wolof	Yoruba	Zulu
Lora	49.14 / 63.94	49.00 / 59.41	47.85 / 61.41	46.78 / 56.92	48.21 / 58.74	48.50 / 58.93	48.42 / 62.85	37.25 / 40.59	38.18 / 42.26	46.92 / 59.22	48.28 / 60.46
qLora8	49.43 / 63.66	49.86 / 65.23	48.93 / 64.23	48.17 / 61.46	49.21 / 62.89	49.21 / 63.80	49.36 / 65.57	39.83 / 45.22	39.68 / 46.90	48.50 / 62.37	49.14 / 63.75
qLora4	39.30 / 50.96	34.38 / 35.96	36.29 / 49.04	35.34 / 44.84	36.06 / 44.17	36.44 / 44.41	37.06 / 49.43	33.38 / 34.96	33.33 / 34.53	38.78 / 49.76	37.06 / 50.38
full	33.31 / 33.24	33.74 / 33.48	33.31 / 33.24	33.31 / 33.24	33.31 / 33.24	33.31 / 33.29	33.31 / 33.29	33.31 / 33.24	33.31 / 33.24	33.31 / 33.24	33.31 / 33.24

Table 16: ***AfrIFact-Culture-News*** without-evidence/with-evidence finetuning experiments

Model	Amharic	English	Hausa	Igbo	Oromo	Shona	Swahili	Twi	Wolof	Yoruba	Zulu
Lora	50.31 / 53.13	51.28 / 40.00	49.33 / 50.51	41.33 / 41.33	40.31 / 39.79	42.87 / 44.05	44.92 / 44.46	33.32 / 32.77	38.72 / 39.49	42.05 / 45.90	42.31 / 44.00
qLora8	55.64 / 57.13	58.77 / 54.00	57.79 / 57.90	49.79 / 50.00	50.41 / 48.05	49.49 / 52.82	52.26 / 49.49	39.33 / 38.72	39.90 / 42.10	50.15 / 53.95	47.59 / 46.97
qLora4	35.18 / 35.62	36.05 / 34.97	41.49 / 48.82	32.87 / 33.13	33.44 / 33.49	37.95 / 38.36	38.46 / 43.08	30.82 / 31.08	38.56 / 38.62	33.59 / 42.05	38.00 / 42.46
full	33.54 / 33.54	33.23 / 33.23	40.31 / 40.31	31.69 / 31.69	32.77 / 32.77	37.54 / 37.54	36.77 / 36.77	30.62 / 30.62	38.62 / 38.62	32.00 / 32.00	37.54 / 37.54

H Without Evidence and With Evidence Prompt Examples

We worked on prompts that focus on $docment + claim$ for without evidence and $document + evidence + claim$ for with evidence span. Given the log-likelihood eval we used for open models, we make the claim and evidence closer together.

Prompt 1

SYSTEM:
You are a helpful assistant for automated fact-checking. Your task is to analyze claims based solely on the provided document.

USER:
You are an intelligent decision support system designed for automated fact-checking. Determine whether the document **Supports**, **Refutes**, or provides **Not_Enough_Information** for the given claim.

Veracity: Support, Refute, Not_Enough_Information,
Justification: Detailed reasoning addressing clarity,relevance, consistency, and sufficiency of the document

Definitions for Veracity Labels:

- **Support:** The claim is accurate and there is nothing significant missing.
- **Refute:** The claim is inaccurate, contradicted by the document, or makes an incorrect assertion.
- **Not_Enough_Information:** The document is insufficient, unrelated, or does not provide enough relevant information to determine whether the claim is true or false.

Determine whether the claim is **Supported**, **Refuted**, or **Not_Enough_Information** based **ONLY** on the provided document.

Document:
{DOCUMENT}

Claim:
{CLAIM}

Question:
Based only on the document above, does it Support, Refute, or provide Not_Enough_Information for the claim?

Answer: (choose `SUPPORTS`, `REFUTES`, or `NOT_ENOUGH_INFORMATION`)

Prompt 2

Role:
You are a professional fact checker responsible for verifying factual claims.

Objective:
Assess the relationship between a claim and the provided document by determining whether the document **Supports**, **Refutes**, or provides **Not_Enough_Information** for the claim.

Constraints:

- Use only the document provided below.
- Do not rely on prior knowledge, assumptions, or external information.
- If the document does not clearly support or contradict the claim, select **Not_Enough_Information**.

Document:
{DOCUMENT}

Claim:
{CLAIM}

Question:
Based only on the document above, does it **Support**, **Refute**, or provide **Not_Enough_Information** for the claim?

Answer: (choose `SUPPORTS`, `REFUTES`, or `NOT_ENOUGH_INFORMATION`)

Prompt 3

SYSTEM:
You are a helpful assistant for Multilingual Evidence-Centered Fact Verification. Your task is to analyze claims based solely on the provided documents.

USER:
You are an intelligent decision support system designed for automated

fact-checking. Determine whether the document **Supports**, **Refutes**, or provides **Not_Enough_Information** for the given claim.

Role:
You are an independent fact checker tasked with evaluating factual claims in linguistically diverse settings.

Objective:
Determine the factual status of a claim by interpreting the provided document, which may appear in different languages or linguistic varieties, and decide whether the document confirms the claim, contradicts it, or fails to address it.

Constraints:

- Base your judgment exclusively on the document provided, regardless of language or linguistic variation.
- Do not use background knowledge, assumptions, or external sources.
- If the document does not directly confirm or contradict the claim, choose **Not_Enough_Information**.

Document:
{DOCUMENT}

Claim:
{CLAIM}

Question:
Based only on the document above, does it **Support**, **Refute**, or provide **Not_Enough_Information** for the claim?

Answer: (choose `SUPPORTS`, `REFUTES`, or `NOT_ENOUGH_INFORMATION`)

Prompt 1

SYSTEM:
You are a helpful assistant for automated fact-checking. Your task is to analyze claims based on the provided evidence.

USER:
You are an intelligent decision support system designed for automated fact-checking. Based **only** on the evidence provided below, determine whether the evidence **Supports**, **Refutes**, or provides **Not_Enough_Information** for the given claim.

Veracity: Support, Refute, Not_Enough_Information
Justification: Detailed reasoning addressing clarity, relevance, consistency, and sufficiency of the evidence

Definitions for Veracity Labels:

- **Support:** The claim is accurate and there is nothing significant missing.
- **Refute:** The claim is inaccurate, contradicted by the evidence, or makes an incorrect assertion.
- **Not_Enough_Information:** The evidence is insufficient, unrelated, or does not provide enough relevant information to determine whether the claim is true or false.

Determine whether the claim is **Supported**, **Refuted**, or **Not_Enough_Information** based **ONLY** on the provided evidence.

Document:
{DOCUMENT}

Evidence:
{EVIDENCE}

Claim:
{CLAIM}

Question:
Based only on the evidence above, does the evidence Support, Refute, or provide Not_Enough_Information for the claim?

Answer: (choose `SUPPORTS`, `REFUTES`, or `NOT_ENOUGH_INFORMATION`)

Prompt 2

Role:
You are a professional fact checker responsible for verifying factual claims.

Objective:
Assess the relationship between a claim and the provided evidence by determining whether the evidence **Supports**, **Refutes**, or provides **Not_Enough_Information** for the claim.

Constraints:

- Use only the evidence provided below.
- Do not rely on prior knowledge, assumptions, or external information.
- If the evidence does not clearly support or contradict the claim, select **Not_Enough_Information**.

Document:
{DOCUMENT}

Evidence:
{EVIDENCE}

Claim:
{CLAIM}

Question:
Based only on the evidence above, does the evidence **Support**, **Refute**, or provide **Not_Enough_Information** for the claim?

Answer: (choose `SUPPORTS`, `REFUTES`, or `NOT_ENOUGH_INFORMATION`)

Prompt 3

SYSTEM:
You are a helpful assistant for Multilingual Evidence-Centered Fact Verification. Your task is to analyze claims based on the provided evidence.

USER:
You are an intelligent decision support system designed for automated fact-checking. Based only on the evidence provided below, determine whether the evidence **Supports**, **Refutes**, or provides **Not_Enough_Information** for the given claim.

Role:
You are an independent fact checker tasked with evaluating factual claims in linguistically diverse settings.

Objective:
Determine the factual status of a claim by interpreting the provided evidence, which may appear in different languages or linguistic varieties, and decide whether the evidence confirms the claim, contradicts it, or fails to address it.

Constraints:

- Base your judgment exclusively on the evidence provided, regardless of language or linguistic variation.
- Do not use background knowledge, assumptions, or external sources.
- If the evidence does not directly confirm or contradict the claim, choose **Not_Enough_Information**.

Document:
{DOCUMENT}

Evidence:
{EVIDENCE}

Claim:
{CLAIM}

Question:
Based only on the evidence above, does the evidence **Support**, **Refute**, or provide **Not_Enough_Information** for the claim?

Answer: (choose `SUPPORTS`, `REFUTES`, or `NOT_ENOUGH_INFORMATION`)

I Annotation Tool Examples

Figures 7 and 8 show screenshots of tools used for claim generation and claim labeling tasks.

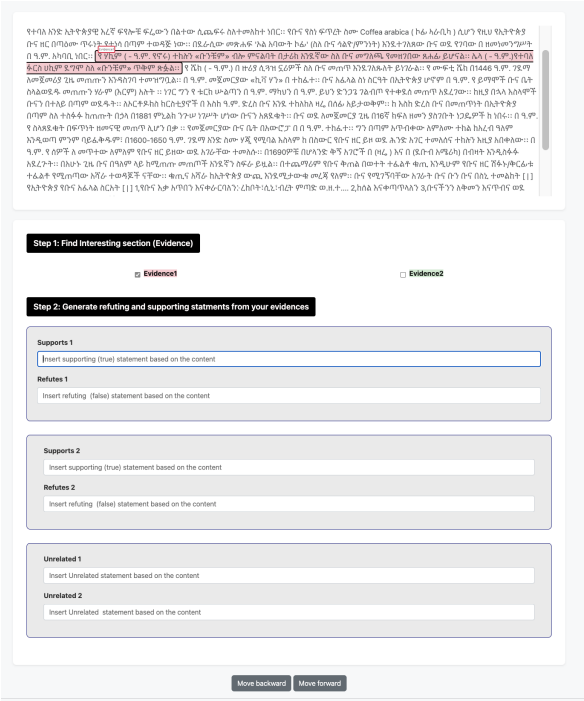


Figure 7: Example interface of the customized tool used for claim generation, showing an Amharic example.

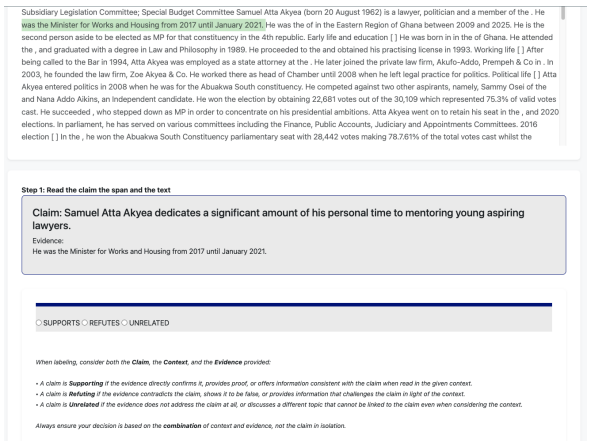


Figure 8: Example of the customized annotation interface used for labeling claims in the claim labeling, showing an English example.

J Annotation Guide

Due to multiple tasks and multiple platforms used in our task, we leveraged several guides, including video, textual, and similar practices on how to perform the tasks, followed by tests and back-and-forth with language coordinators to verify the validity of the tasks.

AfriFact Annotation Guidelines

Task 1: Claim Generation
Claim Definition: A claim is a single sentence expressing information about one target entity.
Requirements

- Must reference the main entity directly (no pronouns).
- Avoid vague expressions (e.g., “may be”, “it is reported”).
- Claims must be derivable from the source sentence and dictionary.
- Include temporal context when facts may change.

Evidence Labels

- **SUPPORTED**: Evidence agrees with the claim.
- **REFUTED**: Evidence contradicts the claim.
- **NOT ENOUGH INFO**: Evidence insufficient to determine truth.

Task 2: Claim Labeling
Objective: Label claims and select the minimal evidence needed.
Procedure

1. Read the claim and candidate evidence sentences.
2. Determine if evidence supports or refutes the claim.
3. Combine multiple sentences if needed.
4. If no sufficient evidence exists, label *NOT ENOUGH INFO*.

Rule of Thumb
If only the selected evidence is given, can the claim be verified as true or false? If not, label NOT ENOUGH INFO.
Edge Case Rules

- Avoid ambiguous claims (e.g., many, several, popular).
- Distinguish between actors and fictional characters.
- Filmographies and lists are not exhaustive evidence.
- Be time-aware when verifying roles or titles.

Skipping Rules

- Claim cannot be verified from Wikipedia.
- Claim is vague or personal.
- Claim contains major grammatical errors.