FINDING THE PARALLELS: A PATH TO EVALUATING SUMMARIZATION FROM LOW-RESOURCE LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Across benchmarks, large language models have demonstrated impressive summarization capabilities for high-resource languages. In contrast, summarization benchmarks are largely unavailable for low-resource languages. This scarcity poses a challenge for both evaluating LLM summarization capabilities and measuring the effectiveness of LLM-as-a-Judge approaches in these languages. However, previous work in machine translation has contributed parallel datasets that pair sentences in low-resource languages with higher-resource translations. We introduce PARAWEAVE, a method for automatically transforming these existing sentence-level parallel datasets into coherent document-level benchmarks while preserving their original cross-lingual alignment. We used PARAWEAVE to construct a summarization benchmark that spans ten language varieties from three families (Afro-Asiatic, Indo-Iranian, and Nilo-Saharan). We use the resulting benchmark to evaluate summarization faithfulness and LLM-as-a-Judge performance for summaries generated from documents in these languages. Pairing low-resource language documents with high-resource translations enables studying the effect of source language on both summarization quality and LLM-as-a-Judge performance. These comparisons reveal gaps in model performance and demonstrate that even current frontier models struggle with faithful low-resource language summarization and evaluation in certain languages. For example, GPT-5.5's summaries of Nilo-Saharan documents are significantly less faithful than their English document counterparts. Native-speaker annotations validate the model judgments in our construction pipeline and judge faithfulness metrics. The proposed pipeline offers a general approach for repurposing existing sentence-level parallel resources for evaluation of document-level tasks.

1 INTRODUCTION

Due to the relative scarcity of text representing low-resource languages and the general cost of manual dataset construction, there are few long-form datasets available in such languages for evaluating long-form tasks such as summarization. The lack of evaluation resources beyond English and other high-resource languages has already been identified as a bottleneck in multilingual model development (Romanou et al., 2025). In the absence of these datasets, model capabilities and performance gaps in numerous languages remain unclear.

Typically, prior work relies on two primary families of methods for evaluating summaries, both of which pose challenges when applied to low-resource languages. First, *reference-based metrics* compare model-generated summaries against human-written summaries (e.g., ROUGE (Lin, 2004), BLEURT (Sellam et al., 2020)). While some low-resource languages have news paired with summaries (e.g., Moroccan Darija on Goud.ma (Issam & Mrini, 2022), Persian from BBC Persian (Hasan et al., 2021) and the languages of LR-Sum (Palen-Michel & Lignos, 2023)), coverage is uneven across languages and dialects, and collecting summaries can be prohibitively expensive at scale.

Second, *reference-free metrics* verify a summary against the source document(s) (e.g., using NLI or LLM-as-a-Judge; Laban et al. (2022); Honovich et al. (2022); Min et al. (2023)), alleviating the need for human-written references. These metrics, however, presuppose that models or LLM judges sufficiently understand the language under evaluation. In either case, translation may be a natural

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

approach to alleviating such resource scarcity, although similarly, *translate-then-evaluate* approaches (Qiu et al., 2023) then confound model performance with translation quality.

One method of addressing these concerns is through *parallel long-form documents*: articles in a low-resource language paired with human-translated articles in English or another high-resource language. Assuming a capable summarizer in the high-resource language, comparing summaries produced from a pair of low resource/English articles enables isolating the effect of input language on model performance. Similarly, this *also* enables assessing the impact of input language on evaluation metrics and LLM judges.

These document-level parallel benchmarks do exist but are limited. NTREX-128 (Federmann et al., 2022), for example, renders 123 news articles into 128 languages, including low-resource languages, but lacks representation of some dialects like Moroccan Arabic. In contrast, *sentence*-level parallel benchmarks, namely those intended for machine translation, are available for many languages and dialects that lack document-level evaluation data. FLORES-200 (NLLB Team et al., 2022) alone translates the same 2,009 dev and devtest sentences into over 200 languages, MADAR (Bouamor et al., 2018) covers 25 Arabic city dialects, and community efforts such as SALT (Akera et al., 2022) provide tens of thousands of sentence pairs for Ugandan languages with little other digital text.

Document-level Benchmarks from Parallel Sentences. In our paper, we leverage the availability of machine translation datasets to address gaps in summarization evaluation in low-resource languages. We introduce an automated benchmark construction framework, PARAWEAVE, which transforms parallel-sentence pairs into coherent documents for summarization benchmarks (Section 3). PARAWEAVE first clusters translation sentence pairs using embedding models on the high-resource language side, then uses LLMs to order and filter sentences into coherent documents (see Figure 1).

Using PARAWEAVE, we contribute a benchmark of 2,330 parallel articles across four Arabic dialects and 1,341 in Lugbara, Acholi, Ateso, Persian, Dari and Tajik.[1] We validate a subset of the dialectal Arabic documents with judgments from native speakers; the other languages are validated automatically. Notably, the resulting documents are LLM-constructed but maintain the authenticity of human-written text in the original translation benchmark (Section 3.2 and Appendix G). We utilize the resulting benchmark to investigate two research questions.

RQ1) How faithfully do multilingual LLMs summarize documents in low-resource languages? We **assess cross-lingual summarization faithfulness**, using the constructed documents to evaluate whether LLMs summarize low-resource documents into English as well as they do English documents (Section 4). These experiments reveal performance gaps between models. In many cases large frontier models outperform specialized open-source models. Exceptions include Ateso and Lugbara. While scaling reasoning compute improves frontier model performance, a *9 billion parameter model* fine-tuned on Lugbara slightly outperforms it.

RQ2) Does LLM-as-a-Judge Performance Generalize to Low-Resource Languages? To measure summarization faithfulness, we previously evaluated summaries generated from low-resource source articles against parallel English source articles. We also **assess cross-lingual, LLM-based faithfulness evaluation**. Here, we evaluate whether an LLM can accurately verify claims against source documents in low-resource languages (Section 5). In contrast to reliable performance on English source documents, we find that an accurate LLM judge can fail to correctly verify claims against low-resource source documents. These results highlight the importance of having access to parallel high-resource documents in low-resource settings, and future work on improving multilingual LLM-as-a-Judge performance in low-resource settings.

In summary, we propose PARAWEAVE, a method for automatically constructing parallel documents from existing sentence-level machine translation datasets. We first use PARAWEAVE to construct a cross-lingual summarization benchmark containing ten low-resource source languages (Section 3). Evaluation on the benchmark reveals large performance gaps in both LLM summarization (Section 4) and judging capabilities (Section 5) across several languages.

[1] Lugbara (Central Sudanic), Acholi (Western Nilotic) and Ateso (Eastern Nilotic) are Nilo-Saharan languages of northwestern, northern and eastern Uganda; Persian, Dari and Tajik are the Iranian varieties of Iran, Afghanistan and Tajikistan, Tajik written in Cyrillic; the Arabic dialects are the regional groups of AraBench (Sajjad et al., 2020) (Egyptian, Gulf, Levantine and Maghrebi).

2 Related Work

Multilingual and low-resource summarization. Multilingual summarization datasets such as XL-Sum (Hasan et al., 2021), CrossSum (Bhattacharjee et al., 2023), MassiveSumm (Varab & Schluter, 2021), and WikiLingua (Ladhak et al., 2020) pair articles with summaries in many languages, but they are harvested from web news and how-to guides, so they cover languages with substantial online text and rarely separate dialects. Evaluation built on FLORES-200 reaches far more languages, with short inputs and closed answers: Belebele asks multiple-choice questions about its passages in 122 language variants (Bandarkar et al., 2024) and SIB-200 classifies its sentences by topic (Adelani et al., 2024), and suites such as AfroBench collect existing tasks for African languages (Ojo et al., 2025). For Arabic, EASC (El-Haj et al., 2010) and KALIMAT (El-Haj & Koulali, 2013) provide human-written and system-generated summaries of Modern Standard Arabic, respectively, dialect and culture benchmarks such as AraDiCE (Mousi et al., 2025), PALM (Alwajih et al., 2025) and DialectalArabicMMLU (Altakrori et al., 2026) test knowledge through instruction–response pairs and multiple-choice questions, BALSAM (Al-Matham et al., 2025) includes summarization among its tasks, and MArSum pairs 19,806 Moroccan Darija news articles with their headlines but has no English references to validate against (Gaanoun et al., 2022). None of these works pair long-form low-resource input with a parallel high-resource version to evaluate summarization faithfulness.

Faithfulness evaluation. Datasets and benchmarks such as FRANK (Pagnoni et al., 2021), SummEval (Fabbri et al., 2021), TRUE (Honovich et al., 2022), AggreFact (Tang et al., 2023), and SummaC (Laban et al., 2022) assess whether generated summaries remain grounded in the source article. FActScore improves this approach for longer text generations by decomposing outputs into smaller atomic claims and verifying the support for each claim (Min et al., 2023). ReIFE (Liu et al., 2025) evaluates the LLM evaluators themselves across many base models and evaluation protocols, and finds that the base model matters more than the protocol. Our judge follows both: a FActScore-inspired decompose-and-verify protocol run by a frontier model. JudgeBench (Tan et al., 2025) likewise builds a test of LLM judges from existing datasets with verifiable answers; ours checks the judge against the parallel English version. Nevertheless, these resources are almost all English: SEAHORSE rates summaries in six languages, none of them among our ten (Clark et al., 2023), and AraHalluEval evaluates Arabic without separating its dialects (Alansari & Luqman, 2025).

Article-level data from sentence-level corpora. Other research has likewise attempted to generate article-level parallel data from sentence-level machine translation corpora. ParaDocs extracts article information from datasets like ParaCrawl, Europarl, and News Commentary to place the sentence pairs back into the article context from which they came (Wicks et al., 2024); other studies use webpage or article metadata in the same way (Al Ghussin et al., 2023; Pal et al., 2024), and DocHPLT (O'Brien et al., 2025) and DocBlocks (Ramos et al., 2025) maintain or create article-level parallel resources for long-context machine translation. These methods recover or preserve existing article structure. Earlier document-level MT work built "fake articles" from unrelated parallel sentences (Junczys-Dowmunt, 2019; Bérard et al., 2020), but as training data, with no aim of coherence. For many low-resource languages and dialects no article-structured resource with references exists: the available data are collections of independent sentence pairs from phrasebooks, conversations or MT test sets, with no article structure to recover, and our pipeline composes new coherent articles from them while preserving authenticity to the source sentences.

3 Automatic benchmark creation from parallel sentences

We first introduce our framework for automatically constructing summarization data from sentence-level translation benchmarks, enabling further evaluation of LLMs' multilingual summarization and evaluation capabilities. We then use the scalable framework to construct a benchmark for cross-lingual summarization evaluation.

3.1 ParaWeave

Problem Setting. We assume access to a sentence-level parallel dataset $D_{\text{sentence}} = \{(a_i, b_i)\}_i$, where a_i is a sentence in the low-resource language and b_i is its gold human translation into a high-resource language. Our goal is to build pairs of coherent multi-sentence articles from these

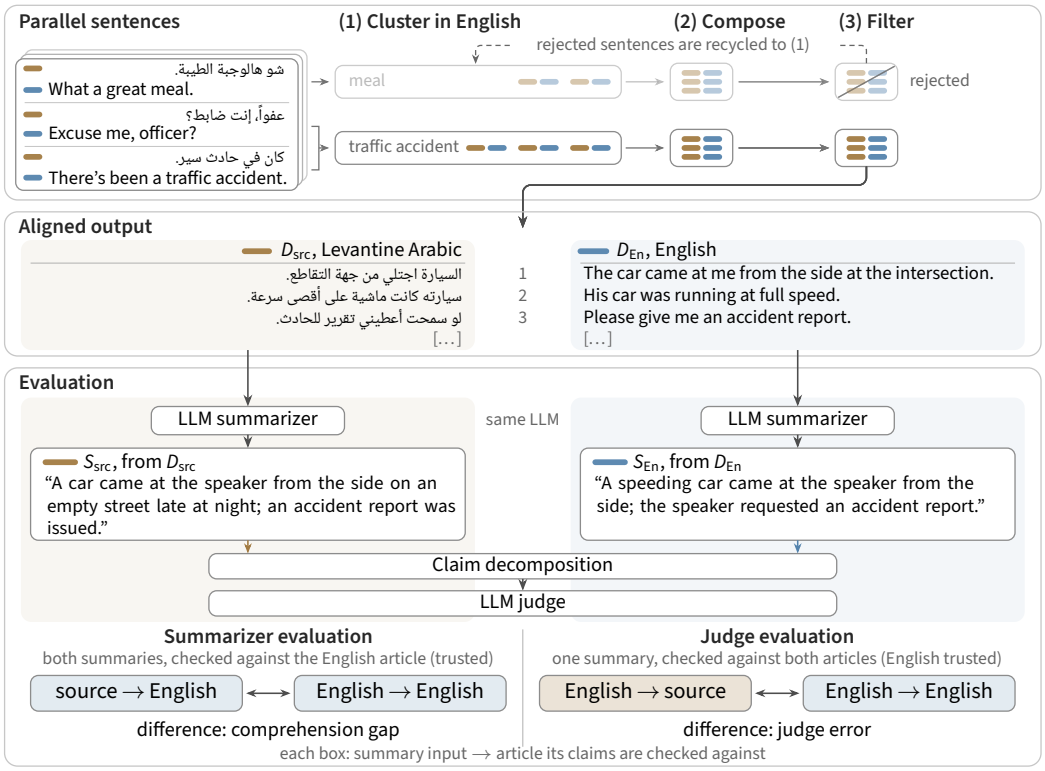

Figure 1: PARAWEAVE constructs articles from independent sentence pairs. Prompted English embeddings are clustered (1). An LLM selects and orders original pairs into articles (2). A separate LLM accepts coherent articles (3). Rejected candidates are recycled to step 1 and accepted documents form the aligned output. During evaluation, the same model summarizes each version into English. Following Min et al. (2023), each summary is decomposed into claims, and an LLM judge checks the faithfulness of each claim. We first evaluate the faithfulness of generated summaries against the English document. We then evaluate cross-lingual LLM-as-a-Judge capabilities by comparing judged faithfulness scores from the English document to those computed against the low-resource reference document.

sentences (Figure 1). Each output article is a pair $(D_{\text{src}}, D_{\text{En}})$ with $D_{\text{src}} = (a_{i_1}, \ldots, a_{i_n})$ and $D_{\text{En}} = (b_{i_1}, \ldots, b_{i_n})$ constructed from ordered sentence pairs. The pipeline has three steps, numbered in Figure 1. We first cluster related pairs in the high-resource language (*step 1*). We then compose candidate articles from each cluster (*step 2*), filtering with a separate LLM (*step 3*). Then we recycle the sentences of rejected candidates into further rounds (the loop back to step 1).

Step 1) Clustering. We first cluster unrelated pairs of sentences into groups of topically related sentences. We cluster on the high-resource side for better accuracy, producing clusters of 150 sentence pairs on average that could all refer to the same event. We embed texts with Qwen3-Embedding-8B (Zhang et al., 2025), prompted to capture whether or not the sentences could plausibly co-occur in the same document. Prompts and additional details are included in Appendices B and C.

Step 2) Composition. A chosen frontier LLM (GPT-5.5; OpenAI, 2026a) takes each individual large cluster and is instructed to select, order, and divide the sentences into candidate articles. The composer prompt is reproduced in Appendix B and its settings in Appendix C. Sentence pairs are selected verbatim; the one edit it may make is proper-noun unification, when two otherwise compatible sentences refer to different names, since independently written lines that fit one event can name different people. The principle is that the composer never adds content from its own knowledge.

Step 3) Filtering. Not every candidate produced by the generation model is coherent. We therefore apply a separate frontier LLM-as-a-Judge filter for coherence (Claude Sonnet 4.6; Anthropic, 2026). A different model is given each constructed article and instructed to rate its overall coherence on a scale from 1 to 5. We define coherence as whether the sentences flow naturally into one article. The prompt for filtering is reproduced in Appendix B. We keep articles that receive a coherence score of at least 4. Any article scoring below 4 is rejected, broken back down into its original sentence pairs, and returned to the candidate pool. On a sample of 32 accepted and rejected articles, the filter agrees with native speakers' average rating at Cohen's $\kappa = 0.62$ (Appendix A).

Finally, we take the rejected pool of sentences and repeat all steps from scratch. Rejected sentences can be reclustered into different groups, allowing them to form new coherent articles. Each complete iteration of this process is considered one round. The full process ends either after a chosen maximum number of rounds or when the passing rate falls below a selected threshold. The maximum number of rounds and the stopping threshold used for each build are listed in Appendix C. The later rounds produce most of the articles (Table 5 in Appendix E), and their articles pass through the same coherence filter as the first round's, so they meet the same quality bar.

3.2 A TEN-LANGUAGE SUMMARIZATION BENCHMARK

PARAWEAVE enables constructing pairs of synthetic documents for any low-resource language with a high-resource sentence-level machine translation dataset. Notably, these datasets are readily available for many languages and dialects. We utilize the method to construct a benchmark of ten languages that is later used for multilingual summarization evaluation.

We ran our pipeline across three language families with existing parallel corpora (Table 3):

- *Afro-Asiatic*: We use four Arabic dialect groups from 105,590 dialect-English pairs (Bouamor et al., 2018; Sajjad et al., 2020). This produced 2,330 articles (301 Egyptian, 914 Gulf, 435 Levantine and 680 Maghrebi).
- *Nilo-Saharan*: We use Lugbara, a Central Sudanic language of Uganda, from 24,946 Lugbara-English pairs of SALT (Akera et al., 2022), giving 240 articles whose multi-way translations give the same articles in Acholi and Ateso, two Nilotic languages of the same family.
- *Indo-Iranian branch of the Indo-European family*: We use Persian, Dari and Tajik, from the 2,009 dev and devtest sentences of FLORES-200 (NLLB Team et al., 2022), translated from English, giving 232, 230 and 159 articles.

In total, the resulting dataset has 3,671 articles, of which 2,330 are in the Arabic dialects and 1,341 are in the other two families. These languages span multiple resource classes. All three Nilo-Saharan languages are in the lowest resource class of Joshi et al. (2020), whereas Persian is in class 4 and Egyptian Arabic in class 3.

SALT's larger pool also allows longer articles (15 sentences on average); on the Arabic articles, length has little effect on faithfulness (Appendix L). The FLORES varieties also serve as a control because their sentences come from known pages in fragments of two to five sentences (Appendix F). We validate the resulting dataset via a native-speaker study. We include additional automatic metrics in Appendix D.

Human Benchmark Validation. We conducted a native-speaker study of synthetic documents in four Arabic dialects with assistance from crowd workers.[2] Seven native Arabic dialect speakers participated (two Egyptian, two Levantine, two Maghrebi, and one Gulf). Annotators evaluated the quality of the accepted synthetic articles against natural news, shuffled variants of the accepted articles, and rejected articles from the earlier coherence gate in step 3 (Figure 1; Section 3.1).

Every annotator completed three tasks on articles in their own dialect (full task descriptions are shown in Appendix A). Annotators (1) evaluated the PARAWEAVE composition step by identifying a random sentence from sentences in the same constructed document, (2) completed a quality checklist to evaluate overall article quality, and (3) provided coherence and human-likeness ratings for each type of document.

[2]The study was reviewed and approved by our institution's IRB. Annotators gave informed consent and were paid 17 USD per hour.

	Natural news	Accepted	Rejected	Shuffled
Coherence (1-4)	3.93	3.07	2.11	2.04
Human-written (1-4)	3.75	2.57	2.25	2.18

Table 1: Native-speaker ratings by hidden condition (7 annotators): natural news, articles the gate accepted, articles the gate rejected (not in the benchmark), and accepted articles with their sentence order shuffled.

Human Validation Results. Every annotator identified the intruder in all eight odd-one-out items from task 1. On the task 3 ratings (Table 1), natural news scores highest, accepted articles next, and the two other controls the lowest as expected. Annotators most often faulted rejected articles as "shared setting only" (independent lines from one domain rather than one continuous scene), exactly what the gate was built to reject. The annotators marked shuffled ones for broken order. Accepted and shuffled articles contain the same sentences, so the difference between their ratings is the order the composer imposed. Accepted articles still score below natural news, which we partially attribute to their source: MADAR sentences are conversational phrasebook lines, so an accepted article reads like a dialogue rather than a news article. By contrast, the coherence gate rates the Nilo-Saharan articles, built from written prose, as much more coherent, likely owing to their source register (Appendix N).

4 RQ1) How Well Do LLMs Summarize Low-Resource Languages?

In this section, we evaluate the quality of summaries generated by LLMs from low-resource source documents. We focus on evaluating summary *faithfulness* to the source document.

Summaries and claims. Each system generates an English summary for every article under two input conditions: the input article in the low-resource language, or the English article formed from the gold translations. We limit summary generation to 30 words to ensure comparable, one-sentence summaries across systems. Each summary is broken into atomic claims by the LLM judge, and each claim is checked against the English article as reference. We write $F_{r,g}$ for the faithfulness of the summary written from article r when its claims are checked against article g (Figure 1).

Judging Faithfulness Against English Source Documents. Each summary claim receives one of four labels: *supported*, *partially supported*, *not supported*, or *contradicted* by the LLM-as-a-Judge. Faithfulness is the fraction of supported claims with half credit for partial support: [3]

$$F_{r,g} = \frac{N^{r,g}_{\text{sup}} + \frac{1}{2} N^{r,g}_{\text{par}}}{N^{r}_{\text{claims}}}, \tag{1}$$

where $N^{r,g}_{\text{sup}}$ and $N^{r,g}_{\text{par}}$ count the claims of the summary generated from input r that reference g supports fully or partially, and N^{r}_{claims} counts all claims of that summary. Figure 7 in Appendix N shows the claims of one article's two summaries, each checked for consistency with each reference article. We freeze one decomposition per summary and judge identical claims under both languages. All judging is performed by Claude Sonnet 4.6. Here, we build on existing work that has demonstrated the effectiveness of LLM-as-a-Judge for evaluating English summaries of English documents (Song et al., 2024). We also validate the judge in this specific setting with human annotations.

As part of the native-speaker study, human annotators also classified whether claims extracted from generated summaries were faithful or unfaithful to the source document. Claude Sonnet 4.6 and annotators agreed on 70% of claims ($\kappa = 0.40$), close to the inter-annotator agreement (73%, $\kappa = 0.45$). This suggests that much of the remaining disagreement is noise rather than a systematic judge error. We additionally evaluate faithfulness with AlignScore (Zha et al., 2023) and MiniCheck (Tang et al., 2024), which correlate with the LLM-as-a-Judge on English documents (Appendices H and I). Together, these results provide evidence that LLM-as-a-Judge scores of faithfulness are reliable when comparing English summaries with English source documents. We therefore use this judge,

[3] Half credit is empirically validated; scoring the annotators' ratings of the claims the judge marks partially supported with the same weights gives them 0.63 credit on average (Appendix A).

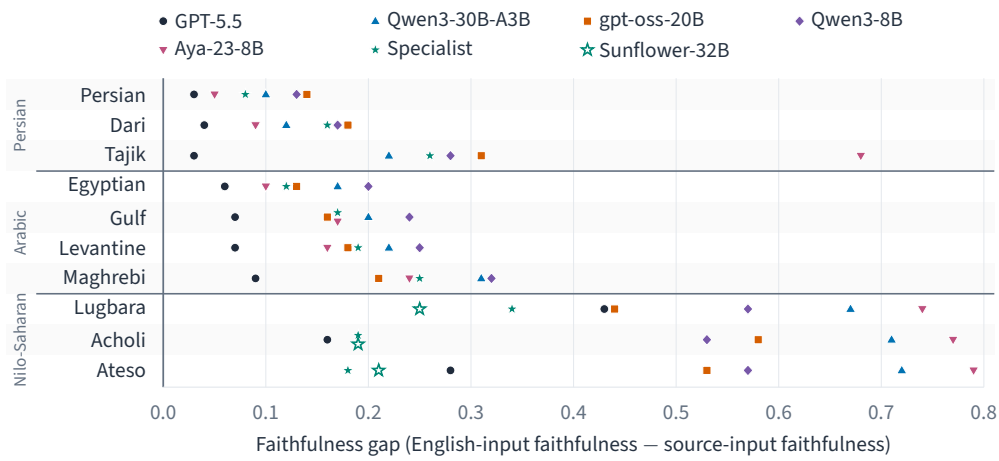

Figure 2: Each summarization system's faithfulness with English input minus its faithfulness with source-language input. Both are judged against the English article (0 means equally faithful), with rules separating the three language families. The specialist is Dorna2-8B for the Persian varieties, Fanar-1-9B for the Arabic dialects and Sunflower-Qwen3.5-9B for the Nilo-Saharan languages, which also show Sunflower-32B (hollow stars). Table 11 provides every value.

along with the English source document, to evaluate the faithfulness of summaries generated from low-resource language documents.

4.1 Experiments

We evaluate the faithfulness of summaries generated from both the low-resource language document and its corresponding English translation against the English article (Figure 1). We use *faithfulness gap* to refer to the difference in scores between summaries generated from the high-resource and low-resource documents, $F_{\mathrm{En,En}}$ and $F_{\mathrm{src,En}}$.

Models. We evaluate six systems on faithful summarization: GPT-5.5, gpt-oss-20B (OpenAI, 2025), Qwen3-30B-A3B and Qwen3-8B (Yang et al., 2025) as open general models, Fanar-1-9B (Fanar Team et al., 2025) as an Arabic-specialized model, and Aya-23-8B (Aryabumi et al., 2024) as a multilingual model. On the Nilo-Saharan languages, we use the specialists Sunflower-Qwen3.5-9B (Sunbird AI, 2026) and Sunflower-32B (Akera et al., 2026), and for the Persian varieties, the regional specialist Dorna2-8B (Part AI, 2024). On Lugbara, we also run GPT-5.6-Sol (OpenAI, 2026b) at each of its five reasoning efforts.

Results. Table 11 in Appendix K contains faithfulness scores for each summarization system. On all four Arabic dialects and all three Persian varieties evaluated, GPT-5.5 outperforms the corresponding specialized language models. The notable exception is the Nilo-Saharan languages, where the specialist language models are most performant in two of three languages. The Sunflower specialist models are the most faithful systems on Lugbara and Ateso, outperforming all other models. Notably, on Lugbara even *Sunflower-9B* slightly outperforms GPT-5.5 (0.50 to 0.47).

The strongest frontier model loses much of its faithfulness when the input article is in a Nilo-Saharan language: GPT-5.5 scores 0.90 with English input, but 0.47 with Lugbara, 0.74 with Acholi and 0.62 with Ateso input. Summarization from very low-resource languages therefore remains unsolved even for frontier models, and without a parallel English article this loss would be hard to measure, since the judge cannot reliably verify claims against the Lugbara or Ateso article.

Figure 2 plots the difference in faithfulness scores when summarizing from English versus the low-resource language. Every system is on average less faithful when summarizing from the low-resource document. The difference is minimal for GPT-5.5 on the Persian varieties and Arabic dialects (at most 0.09), but larger for the four general open-weight models (0.21 on average), and these losses track how well each system translates the source sentences (Appendix O). The difference for GPT-5.5 is

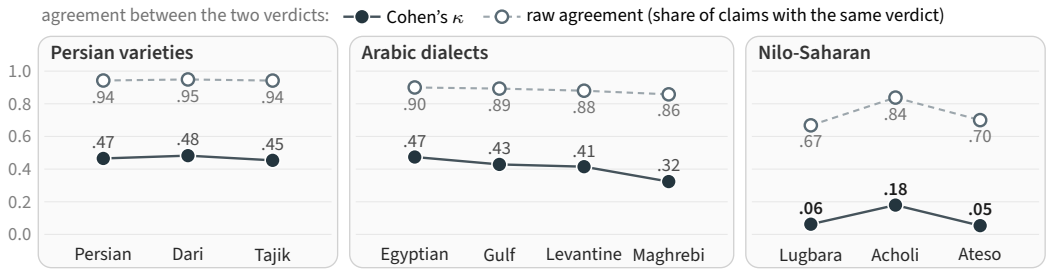

Figure 3: Comparison of LLM-as-a-Judge (Claude Sonnet 4.6) verdicts on summary claims checked against the source-language article versus the English article. κ is 0.32 to 0.48 for the Persian varieties and Arabic dialects and at most 0.18 for the Nilo-Saharan languages.

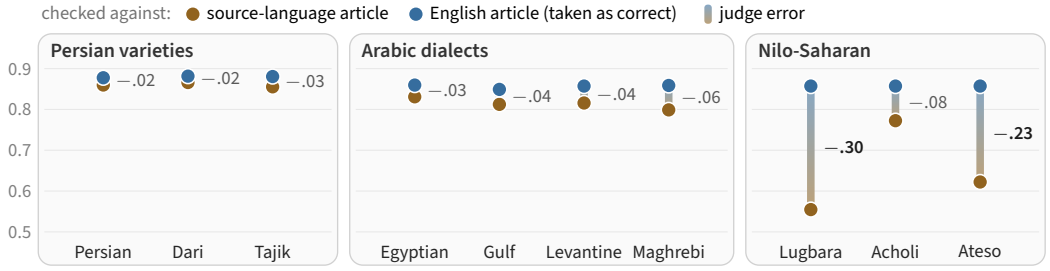

Figure 4: English-input summaries scored by Claude Sonnet 4.6 against the source-language article (gold) and the English article (blue), averaged over the systems of Table 11.

largest on the Nilo-Saharan languages, where faithfulness roughly halves when generating summaries from Lugbara documents (0.74 and 0.62 in Acholi and Ateso). In reasoning intensity experiments with GPT-5.6-Sol, faithfulness with Lugbara input rises with increasing reasoning effort, from 0.37 without reasoning to 0.49 with reasoning set to xhigh (Figure 6 in Appendix E).

Table 10 in Appendix J shows the claim error types annotated by the LLM-as-a-Judge. On the Arabic dialects, summaries including incorrect names and wrong numbers become several times as frequent when the input article is in the dialect. In Lugbara, models are more likely to include unsupported claims, twelve times the English-input rate. Additionally, the summaries themselves change style, and overgeneralize. Half of gpt-oss-20B's Lugbara-input summaries and a quarter of Qwen3-8B's yield no claim. Overall, the faithfulness of summaries generated from documents in low-resource languages is lower than those generated from English document counterparts, with error types dependent on source language. Next, we evaluate the effect of source language on LLM-as-a-Judge performance.

5 RQ2) Does LLM-as-a-Judge Performance Generalize to Low-Resource Languages?

In this section, we study whether LLM-as-a-Judge capabilities are maintained when checking claims against reference documents in low-resource languages.

In the previous section, we found that Claude Sonnet 4.6 was a reliable judge of faithfulness, aligning with previous work on LLM judges for high-resource summarization tasks (Song et al., 2024). To evaluate whether this capability generalizes to other languages, we compare performance when utilizing the low-resource text as the reference document.

5.1 Experiments

We use the summaries generated from English source documents in the summarization evaluation experiment. As before, we extract claims from each generated English summary. However, in contrast to the summary faithfulness evaluation in the previous section, we now use the Claude Sonnet 4.6 LLM-as-a-Judge to check claims against *both* the low-resource and English versions of each source

document (Figure 1, bottom). We examine whether claim judgments change based on reference document language, treating the verdict against the English document as correct. We consider *claim agreement* and the difference in average faithfulness scores.

Details. We evaluate Claude Sonnet 4.6 as the Judge language model, with mean faithfulness scores calculated over the same systems that we evaluated faithfulness on: GPT-5.5, gpt-oss-20B, Qwen3-30B-A3B, Qwen3-8B and Aya-23-8B, plus the specialists Dorna2-8B (Persian varieties), Fanar-1-9B (Arabic dialects) and both Sunflower models (Nilo-Saharan languages).

Results. In Figure 3, we plot claim agreement between the LLM judge when evaluating against the English and low-resource source document. Claim agreement is highest for Persian varieties, followed by Arabic dialects. Nilo-Saharan languages have low agreement, with a Cohen's κ between 0.05 and 0.18. In Figure 4, we additionally compare the differences in average faithfulness scores across source document language. Notably, when evaluating against the low-resource source document, the LLM-as-a-Judge approach rates summaries of Nilo-Saharan documents as significantly less faithful. The difference is most pronounced for Lugbara (-0.30). Overall, these results show that LLM-as-a-Judge approaches that are validated in high-resource settings can fail to generalize to low-resource language reference documents.

6 Discussion

There is scarcity of benchmarks for evaluating the multilingual summarization capabilities of language models from low-resource languages and dialects. Using ParaWeave, we automatically crafted a cross-lingual summarization benchmark of paired low-resource and high-resource documents that covered ten source languages. We utilize the benchmark to evaluate language model summarization faithfulness and LLM-as-Judge performance.

RQ1: How well do LLMs summarize low-resource languages? Summarization quality is inconsistent across languages, and *large gaps can exist between low-resource and high-resource settings*. The frontier model we evaluated, GPT-5.5, outperformed both open-source and specialist models in many settings. However, lower-resource Nilo-Saharan languages were an exception where specialist Sunflower-32B and Sunflower-9B models were able to outperform GPT-5.5 in faithfulness. The strongest frontier model loses much of its faithfulness when the input article is in a Nilo-Saharan language: GPT-5.5 scores 0.90 with English input, but 0.47 with Lugbara, 0.74 with Acholi and 0.62 with Ateso input. Summarization from very low-resource languages therefore remains unsolved even for frontier models. Without a parallel English article this loss would be hard to measure, since the judge cannot reliably verify claims against the Lugbara or Ateso article.

RQ2: Does LLM-as-a-Judge Performance Generalize to Low-Resource Languages? In our experiments, *a strong LLM-as-a-Judge can fail* for certain languages. Sonnet 4.6 had high agreement with native-speaker annotators when comparing summaries to English documents. However, its behavior degraded substantially when verifying claims against Nilo-Saharan reference documents.

Overall, our work identifies gaps in model performance across languages in current language models, as well as highlights ways that existing datasets can be repurposed to elucidate these limitations.

7 Limitations

Our approach has several limitations. First, the quality of the benchmark depends on the register of its source data. The Arabic articles are built from travel phrasebooks and conversational sentences, so they test comprehension of conversational dialect and read less like documents than real news does. The Nilo-Saharan articles, built from written prose, read much closer to real documents (Appendix N), which suggests that the gap comes from the source data rather than the pipeline. Second, the English side is only as good as its human translations. Third, the dialect labels are the regional groups of AraBench, so Gulf also covers Iraqi and Yemeni cities and Egyptian covers Sudanese ones; every article is released with its country code for finer analysis. Finally, our pipeline's performance is based on many factors from composer choice to judge choice, every model has to be chosen carefully, and so does the cluster size relative to the size of the dataset.

AI USE STATEMENT

AI was used for assistance with literature review, polishing the manuscript, and for agentic coding. Additionally language models were employed for synthetic data generation.

ETHICS STATEMENT

The native-speaker study was reviewed and approved by our institution's IRB. Annotators were recruited through Prolific and Upwork with the same screening questionnaire, gave informed consent, and were paid 17 USD per hour for a session of about one hour. Responses are reported only in aggregate. About half a percent of articles were manually identified as containing charged material inherited from the source corpus, which were excluded from the survey. The Arabic sentence pool is the public text release of AraBench (MADAR and QAraC credited), and the ten articles that contain sentences of the research-only MDC corpus are withheld; the Persian pools derive from FLORES and the Lugbara pool from SALT, each released under its own terms with attribution (Appendix P).

REPRODUCIBILITY STATEMENT

The anonymized supplementary material contains the pipeline code, the exact prompts (Appendix B), the released benchmark with all results and scripts. Model versions and settings are in Appendix C. Construction depends on API models and is not bit-reproducible, so the full construction log is released for re-auditing.

REFERENCES

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 226–245, 2024.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. Machine translation for african languages: Community creation of datasets and models in uganda. In *3rd Workshop on African Natural Language Processing (AfricaNLP) at ICLR 2022*, 2022. URL https://openreview.net/forum?id=BK-z5qzEU-9.

Benjamin Akera, Evelyn Nafula Ouma, Gilbert Yiga, Patrick Walukagga, Phionah Natukunda, Trevor Saaka, Solomon Nsumba, Lilian Teddy Nabukeera, Joel Muhanguzi, Imran Sekalala, Nimpamya Janat Namara, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. Sunflower: A new approach to expanding coverage of African languages in large language models. In *Proceedings of the AI for African Languages Conference*, volume 314 of *Proceedings of Machine Learning Research*, pp. 1–20, 2026. URL https://proceedings.mlr.press/v314/akera26a.html.

Yusser Al Ghussin, Jingyi Zhang, and Josef van Genabith. Exploring Paracrawl for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 1304–1310, Dubrovnik, Croatia, 2023.

Rawan Al-Matham, Kareem Darwish, Raghad Al-Rasheed, Waad Alshammari, Muneera Alhoshan, Amal Almazrua, Asma Al Wazrah, Mais Alheraki, Firoj Alam, Preslav Nakov, et al. BALSAM: A platform for benchmarking Arabic large language models. *arXiv preprint arXiv:2507.22603*, 2025.

Aisha Alansari and Hamzah Luqman. AraHalluEval: A fine-grained hallucination evaluation framework for Arabic LLMs. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pp. 148–161, Suzhou, China, 2025.

Malik H. Altakrori, Nizar Habash, Teresa Lynn, Younes Samih, Abed Alhakim Freihat, Kirill Chirkunov, Muhammed AbuOdeh, Radu Florian, Preslav Nakov, and Alham Fikri Aji. DialectalArabicMMLU: Benchmarking dialectal capabilities in Arabic and multilingual language models.

In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC)*, pp. 3199–3219, Palma de Mallorca, Spain, 2026.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim Elmadany, et al. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, 2025.

Anthropic. System card: Claude Sonnet 4.6. https://www.anthropic.com/claude-sonnet-4-6-system-card, 2026. February 2026.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, et al. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*, 2024.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, 2024.

Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

Alexandre Bérard, Vassilina Nikoulina, Ioan Calapodescu, and Jerin Philip. Naver Labs Europe's participation in the robustness, chat, and biomedical tasks at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*, pp. 462–472, 2020.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. CrossSum: Beyond English-centric cross-lingual summarization for 1,500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2541–2564, Toronto, Canada, 2023.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. A multidialectal parallel corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1240–1245, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA). URL https://aclanthology.org/L14-1435/.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan, 2018.

Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves: How humans interpret Topic Models. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pp. 288–296, 2009.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur P. Parikh. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9397–9413, 2023.

Hoa Trang Dang. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference (DUC 2005)*, 2005.

Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pp. 225–232, Banff, Canada, 2004.

Mahmoud El-Haj and Rim Koulali. KALIMAT: a multipurpose Arabic corpus. In *Proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL-2)*, pp. 22–25, 2013.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. Using Mechanical Turk to create a corpus of Arabic summaries. In *Proceedings of the LREC 2010 Workshop on Language Resources and Human Language Technologies for Semitic Languages*, pp. 36–39, Valletta, Malta, 2010.

Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. Development of a TV broadcasts speech recognition system for Qatari Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, et al. Fanar: An Arabic-centric multimodal generative AI platform. *arXiv preprint arXiv:2501.13944*, 2025.

Christian Federmann, Tom Kocmi, and Ying Xin. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation (SUMEval)*, pp. 21–24. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.sumeval-1.4. URL https://aclanthology.org/2022.sumeval-1.4/.

Kamel Gaanoun, Abdou Mohamed Naira, Anass Allak, and Imade Benelallam. Automatic text summarization for Moroccan Arabic dialect using an artificial intelligence approach. In *Business Intelligence: 7th International Conference, CBI 2022*, Lecture Notes in Business Information Processing, pp. 158–177. Springer, 2022. doi: 10.1007/978-3-031-06458-6_13.

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*, 2023.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 3905–3920, Seattle, United States, 2022.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Abderrahmane Issam and Khalil Mrini. Goud.ma: a news article dataset for summarization in Moroccan Darija. In *3rd Workshop on African Natural Language Processing (AfricaNLP), ICLR*, 2022.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202, 2016.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.560.

Marcin Junczys-Dowmunt. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 225–233, Florence, Italy, 2019.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A. Hearst. Can Transformer models measure coherence in text? Re-Thinking the Shuffle Test. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1058–1064, 2021. URL https://aclanthology.org/2021.acl-short.134/.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4034–4048, 2020.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Political Analysis*, 32(1):84–100, 2024. doi: 10.1017/pan.2023.20. URL https://doi.org/10.1017/pan.2023.20.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in LLM-as-a-judge. *arXiv preprint arXiv:2502.01534*, 2025a.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. Evaluating scoring bias in LLM-as-a-judge. *arXiv preprint arXiv:2506.22316*, 2025b.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Yixin Liu, Kejian Shi, Alexander Fabbri, Yilun Zhao, Peifeng Wang, Chien-Sheng Wu, Shafiq Joty, and Arman Cohan. ReIFE: Re-evaluating instruction-following evaluation. 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)*, pp. 12247–12287, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.naacl-long.610/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12076–12100, Singapore, 2023.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pp. 4186–4218, Abu Dhabi, UAE, 2025.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias. *arXiv preprint arXiv:2606.19544*, 2026.

Dayyán O'Brien, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. DocHPLT: A massively multilingual document-level translation dataset. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, pp. 286–300, Suzhou, China, 2025.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19048–19095, Vienna, Austria, 2025.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

OpenAI. GPT-5.5 system card. https://openai.com/index/gpt-5-5-system-card/, 2026a. April 2026.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026b. July 2026.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 4812–4829, 2021.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13185–13197, Bangkok, Thailand, 2024.

Chester Palen-Michel and Constantine Lignos. LR-Sum: Summarization for less-resourced languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6829–6844, Toronto, Canada, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.427/.

Part AI. Dorna2-Llama3.1-8B-Instruct. https://huggingface.co/PartAI/Dorna2-Llama3.1-8B-Instruct, 2024. Hugging Face model card.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, 2015. Association for Computational Linguistics. URL https://aclanthology.org/W15-3049/.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. Detecting and mitigating hallucinations in multilingual summarisation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8914–8932, 2023.

Miguel Moura Ramos, Patrick Fernandes, Sweta Agrawal, and André F. T. Martins. Multilingual contextualization of large language models for document-level machine translation. *arXiv preprint arXiv:2504.12140*, 2025.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, et al. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations (ICLR)*, 2025.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 5094–5107, 2020.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.acl-main.704/.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, et al. Atlas-chat: Adapting large language models for low-resource Moroccan Arabic dialect. *arXiv preprint arXiv:2409.17912*, 2024.

Guokan Shang, Hadi Abdine, Ahmad Chamma, Amr Mohamed, Mohamed Anwar, Abdelaziz Bounhar, Omar El Herraoui, et al. Nile-chat: Egyptian language models for Arabic and Latin scripts. *arXiv preprint arXiv:2507.04569*, 2025.

Aili Shen, Meladel Mistica, Bahar Salehi, Hang Li, Timothy Baldwin, and Jianzhong Qi. Evaluating document coherence modeling. *Transactions of the Association for Computational Linguistics*, 9: 621–640, 2021. URL https://aclanthology.org/2021.tacl-1.38/.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 906–922, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.51/.

Sunbird AI. Sunflower-Qwen3.5-9B. https://huggingface.co/Sunbird/Sunflower-Qwen3.5-9B, 2026. Preview release.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations (ICLR)*, 2025.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11626–11644, Toronto, Canada, 2023.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.499/.

Daniel Varab and Natalie Schluter. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10150–10161, 2021.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

Rachel Wicks, Matt Post, and Philipp Koehn. Recovering document annotations for sentence-level bitext. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9876–9890, Bangkok, Thailand, 2024.

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

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, Toronto, Canada, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.634/.

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. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A Human Study Detail

Each dialect has its own survey of 58 items: 8 odd-one-out items, 17 checklist articles, 16 articles for the two-pass ratings (4 per hidden condition), and 17 faithfulness claims. Each annotator is a native speaker of one dialect, screened with an eligibility questionnaire, and works on that dialect only for about one hour. Annotators were picked by the region they grew up in and their self-reported native dialect; for Gulf no such annotator was available, so we used annotators from other regions who reported Gulf proficiency and confirmed they understood the articles.

The four tasks. Every annotator completed four tasks on articles in their own dialect: the first three validate the benchmark (Section 3.2) and the fourth the LLM judge (Section 4); the instructions, question wording and answer options follow under Instrument.

1. *Odd one out.* Five sentences, four from one article and one from another, and the annotator picks the one that does not belong, an intrusion test inspired by Chang et al. (2009) and Shen et al. (2021). This checks that articles are distinct enough for a native speaker to tell apart. This is to validate article grouping.
2. *Quality checklist.* For an accepted article, the annotator ticks every quality issue that applies from a fixed list (repetition, unclear reference, grammaticality, focus, lacking sensible meaning, cut off, other) or none; the first four are the DUC quality questions (Dang, 2005), as used by SummEval (Fabbri et al., 2021). This is to evaluate overall article quality.
3. *Coherence and human-likeness ratings.* Sixteen articles from four conditions that the annotator is not told about: natural news, accepted articles, articles the gate rejected, and accepted articles with their sentence order destroyed by k-block shuffling (Barzilay & Lapata, 2008; Laban et al., 2021). Each article is rated twice on a four-point scale, once for "How coherent is this article?" and once for "How human-written does this article sound?" (inspired by Ippolito et al., 2020); a rating of 1 or 2 asks for the reasons from a checklist. This is to evaluate the gating mechanism.
4. *Claim faithfulness.* The annotator sees the Arabic source article and one English claim taken from a system summary and rates how well the source supports it, from fully supported to fully contradicted (Min et al., 2023), naming the error type for anything below fully supported (Pagnoni et al., 2021; Alansari & Luqman, 2025). This is the task that validates the LLM judge of Section 4.

Instrument. The survey was run in Qualtrics; the wording below is reproduced from the survey file. The welcome page told annotators that the articles are "a mix of two kinds: some were written entirely by a human, and some were assembled by an AI system out of real human-written sentences (the AI selects and orders existing sentences; it does not invent new ones). You will not be told which is which."

- **Odd one out**: "Which ONE of these five sentences does not belong with the others?"
- **Quality checklist**: "Read the article below, then select EVERY quality issue that applies to it from the checklist underneath. You may select as many issues as apply, or none at all if the article has no such issues." Options: unnecessary repetition; referential clarity issue; grammaticality problem (bad enough to make the meaning unclear); focus issue; lacks sensible meaning; incomplete or feels cut off; other; none of the above.
- **Coherence**: "How coherent is this article?" with 1 not at all coherent, 2 somewhat coherent, 3 mostly coherent, 4 completely coherent. For a rating of 1 or 2, the follow-up asks for every type of incoherence observed: topic scatter; shared setting only / phrasebook (sentences share a domain or setting but are independent, one-off lines, not a continuous scene); broken order or structure; contradictory or inconsistent content.
- **Human-written**: "How human-written does this article sound?" with 1 does not sound human-written at all, 2 somewhat, 3 mostly, 4 completely. For a rating of 1 or 2, the follow-up lists: repetitive or redundant phrasing; mismatched voice, register or formality; generic or formulaic phrasing; uneven level of detail or pacing.
- **Claim faithfulness**: the Arabic source article is shown above an English claim, then "How well is this claim supported by the source? Rate from Fully supported to Fully contradicted. Judge the claim against the Arabic source above." Options: (2) fully supported, (1) partially supported, (0) irrelevant / not supported, (-1) partially contradicted, (-2) fully contradicted. Anything below fully supported asks for one error type: entity, number, relation, unsupported, overgeneralization, other. For comparison with the judge's four labels, the two contradicted levels are merged.

Gate against the annotators. On the 32 accepted and rejected articles of task 3, the gate agrees with the annotators' mean coherence rating (coherent at 3 or above) on 26 articles ($\kappa = 0.62$), while

two annotators of the same dialect agree with each other at $\kappa = 0.13$. Table 2 gives the reasons for low ratings.

Condition	Rated 1–2	Broken order	Shared setting only	Topic scatter	Contradictory
Natural news	0/28	0	0	0	0
Accepted	9/28	9	3	1	3
Rejected	21/28	18	13	4	4
Shuffled	22/28	20	9	3	7

Table 2: Reasons given for coherence ratings of 1 or 2 (several allowed; 28 ratings per condition). Rejected articles are faulted for "shared setting only" far more often than accepted ones.

Judge against the annotators. On whether a claim is fully supported, the judge and the annotators agree on 70% of claims ($\kappa = 0.40$), and two annotators agree on 73% ($\kappa = 0.45$) of the 51 claims both rated. The claims were drawn from those on which the judge's verdicts against the Arabic and English articles agree, so this task cannot adjudicate the disagreements of Section 5.

B PROMPTS

The prompts below are reproduced from the released code without change; placeholders in double underscores or braces are filled per call, and the source-language name is a parameter ("dialectal Arabic" for the Arabic build). System prompt first, then the user message.

Composer.

```
You are given a topical CLUSTER of dialectal Arabic sentences with
    English translations (numbered SOURCE ||| EN pairs). Select and
    order sentences into ARTICLES that each score >=4 on this rubric.
STRICT COHERENCE RUBRIC (the ONLY thing that counts):
  5 = ONE topic/event/scene: the SAME participants and a single
    continuous through-line (one conversation, one incident, one story).
  4 = one scene with only minor drift.
  3 = loosely related: a shared SETTING/DOMAIN but DIFFERENT sub-things
    -- e.g. several separate accident-related interactions, or a list of
     payment/shopping/tourism phrases. A shared setting/domain/topic is
    NOT coherence. REJECT (do NOT output).
  1-2 = grab-bag / phrasebook / independent one-off utterances (tourism
    Qs, food orders, symptom lists, single-word commands).
Each article = 5+ pairs forming a GENUINE single connected scene/story (
    rubric >=4). ALLOWED edits ONLY: (1) reorder pairs for flow; (2)
    unify PROPER-NOUN names (same change to BOTH the dialectal Arabic
    and the English). FORBIDDEN: rewording, adding connectives, changing
     common nouns/verbs, adding/removing content. Be STRICT: only output
     articles you are confident are >=4 (a single connected scene). A
    cluster that is only a shared setting yields NOTHING -- return [].
    If no single connected scene exists, return []. Return ONLY JSON: {"
    articles":[{"pairs":[{"ar":"..","en":".."}, ...]}, ...]}.

<the cluster as numbered lines: i. SOURCE ||| EN>
```

Composer, rebuild after a rejected candidate (user message; same system prompt).

```
This article was judged loosely coherent. These sentence indices do not
    belong: __DROP__. Rebuild ONE tighter article as a connected subset
    (drop those + anything else off-scene; reorder allowed; NO rewording
    /new content). Keep >=5 pairs or return empty. Numbered SOURCE |||
    EN pairs:
__BODY__

Return ONLY JSON {"pairs":[{"ar":"..","en":".."}, ...]}.
```

Gate, coherence check.

```
You rate the COHERENCE of a candidate article on this exact scale. Judge
     STRICTLY; a shared setting/domain is a 3, NOT a 4.
STRICT COHERENCE RUBRIC (the ONLY thing that counts):
  5 = ONE topic/event/scene: the SAME participants and a single
    continuous through-line (one conversation, one incident, one story).
  4 = one scene with only minor drift.
  3 = loosely related: a shared SETTING/DOMAIN but DIFFERENT sub-things
    -- e.g. several separate accident-related interactions, or a list of
     payment/shopping/tourism phrases. A shared setting/domain/topic is
    NOT coherence. REJECT (do NOT output).
  1-2 = grab-bag / phrasebook / independent one-off utterances (tourism
    Qs, food orders, symptom lists, single-word commands).

Numbered English sentences:
__BODY__

Return ONLY JSON {"coherence": <1-5 per the rubric>, "drop":[<indices
    NOT part of the single connected scene>], "reason":"<=15 words"}.
```

Summarizer (identical for every evaluated system; maxw = 30).

```
You are a precise cross-lingual summarizer. You read Arabic (including
    dialect) and write English.

Summarize the following {dialect}-Arabic text in ONE English sentence of
     at most {maxw} words. Output ONLY the summary as plain declarative
    English -- no preamble, no quotes, no commentary. Do not add any
    fact that is not stated in the text.

TEXT:
{src}
```

Claim decomposition.

```
You break a summary into atomic facts. An atomic fact states ONE thing,
    is fully self-contained (resolve every pronoun to an explicit entity
    ), and copies no opinion or hedge. Output ONLY facts that the
    summary asserts. Many summaries are unattributed first-person
    dialogue with no named speaker -- in that case resolve 'I'/'me' to '
    the speaker' and 'you' to 'the listener' (or 'the addressee'); do
    NOT return zero claims just because no proper name is given.

Decompose this summary into atomic factual claims. Return ONLY JSON: {"
    claims": ["<claim 1>", "<claim 2>", ...]}. Each claim is one short
    standalone English sentence with all references made explicit. If
    the summary states nothing factual, return {"claims": []}.

SUMMARY:
<summary text>
```

Claim judging.

```
You are a strict cross-lingual faithfulness judge. You read Arabic (
    including dialect) and English. Given a SOURCE text and CLAIMS taken
     from an English summary of it, decide for EACH claim whether the
    SOURCE supports it. Judge ONLY against the source; use no outside
    knowledge.

SOURCE:
__SRC__

CLAIMS:
__CLAIMS__
```

```
For each claim return a verdict. Labels: SUPPORTED (fully stated/
    entailed by the source), PARTIAL (partly supported, some detail
    missing or vague), NOT_SUPPORTED (the source does not state it --
    fabricated/extrinsic), CONTRADICTED (the source states the opposite).
     error in { none | entity: wrong/invented named thing (person/place/
    org) | number: wrong quantity/date/amount/count | relation: right
    entities but wrong who-did-what/attribution/causality | unsupported:
     content absent from the source (fabrication) | contradiction:
    asserts the opposite of the source | overgeneralization: broader/
    stronger than the source supports }. Return ONLY JSON: {"verdicts
    ":[{"i":1,"label":"...","error":"...","evidence":"<short source
    quote or empty>","reason":"<=15 words"}]} with one entry per claim
    index.
```

Translation probe (Appendix O; greedy decoding).

```
You are a precise translator. Translate the given __LANG__ sentence into
     English. Output ONLY the English translation as plain text -- no
    preamble, no quotes, no commentary, no transliteration.

__LANG__:
__SENTENCE__

English:
```

C Models and Settings

Variety	Source pairs	Pool	Articles	Sentences	Sent./art.	En words/art.
Egyptian	MADAR/BTEC	17,236	301	2,397	8.0	69
Gulf	MADAR/BTEC	36,045	914	7,656	8.4	82
Levantine	MADAR/BTEC	21,507	435	3,481	8.0	70
Maghrebi	MADAR/BTEC	30,802	680	5,406	8.0	69
Arabic dialects		**105,590**	**2,330**	**18,940**	**8.1**	**74**
Persian	FLORES-200	2,009	232	997	4.3	91
Dari	FLORES-200	2,009	230	1,001	4.4	91
Tajik	FLORES-200	2,009	159	701	4.4	93
Lugbara	SALT	24,946	240	3,571	14.9	142
Acholi	SALT	–	240	3,571	14.9	142
Ateso	SALT	–	240	3,571	14.9	142

Table 3: Composition of every article set and the size of the sentence pool it was built from; the Acholi and Ateso articles are the Lugbara articles in those languages.

- **Embedding.** Qwen3-Embedding-8B (`Qwen/Qwen3-Embedding-8B`, sentence-transformers, float16, batch 16) embeds the English side under the instruction "Instruct: Group sentences that belong to the same conversation, scene, or connected event.\nQuery: " and the 4096-d vectors are L2-normalised. Each round reduces the surviving pool with scikit-learn `PCA` to 50 dimensions and clusters it per sub-dialect with `MiniBatchKMeans` ($k = \lfloor n/150 \rfloor$, batch 512, `n_init=auto`), both with `random_state=0`.
- **Composition.** GPT-5.5, one cluster per call. The model is `gpt-5.5-2026-04-23` through the OpenAI API in JSON mode with `reasoning_effort=low` and `max_completion_tokens=4096`; GPT-5 models fix temperature at the API default. At most 250 sentence pairs of a cluster are shown, every article must have at least five pairs, and a candidate the gate rejects may be rebuilt at most twice from the gate's list of off-scene sentences. The released Arabic articles were composed in July 2026 (final build 2026-07-08).
- **Gate.** A separate frontier model (Claude Sonnet 4.6). It runs as `claude-sonnet-4-6` through the Messages API at temperature 0 in two calls per candidate: a coherence

check on the English side that returns a 1–5 score and the indices of off-scene sentences (`max_tokens=300`), then the four-axis rubric on both sides (`max_tokens=150`); a candidate is kept when the rubric coherence is at least 4 and it has at least 7 sentences. A build runs for at most 12 rounds (the Arabic dialects) or 6 (the FLORES varieties) and stops earlier when a round keeps fewer than 5 articles (2 for FLORES); the Lugbara build ran 3 rounds and was then resumed for 10 more, 13 in all (Table 5).

- **Evaluated systems.** GPT-5.5 (`gpt-5.5-2026-04-23`), gpt-oss-20B, Qwen3-30B-A3B, Qwen3-8B, Fanar-1-9B-Instruct, Aya-23-8B. The open-weight systems are `openai/gpt-oss-20b`, `Qwen/Qwen3-30B-A3B`, `Qwen/Qwen3-8B`, `QCRI/Fanar-1-9B-Instruct` and `CohereForAI/aya-23-8B`, run in-process with transformers (`torch_dtype=auto`) on one GPU; GPT-5.5 runs through the OpenAI API. Every system receives the same prompt (Appendix B) asking for one English sentence of at most 30 words, decoded with temperature 0.3, top-p 0.8 and at most 90 new tokens; GPT-5.5 uses `reasoning_effort=low` at its default temperature, Qwen3 runs with thinking off, and gpt-oss runs at low reasoning effort with only its final channel kept. Summaries were generated on 2026-07-15 and 2026-07-16.
- **Judge.** Claude Sonnet 4.6 (`claude-sonnet-4-6`) through the Batches API, temperature 0. Each call judges all claims of one summary against one reference with `max_tokens=900` and the source truncated to 8,000 characters; the full run is about 105,700 calls made between 2026-07-15 and 2026-07-17. Claims are extracted once per summary by the same model with the decomposition prompt (temperature 0, `max_tokens=400`) and frozen, so both references are checked on identical claims; a live re-run of the decomposer recovered 5,537 of the 5,995 summaries for which the first pass returned no claims. Judged summaries exist for 2,170 of the 2,330 released articles, and the result tables are computed on those.
- **Transfer runs.** The Acholi and Ateso versions of the Lugbara articles were summarized on 2026-09-25 with every system's Lugbara settings and judged by the same judge; their multi-way SALT text is read from the text columns of `mekaneeky/Synthetic_Acholi_MMS`, an ungated release that carries SALT's text table. Lugbara and the Persian varieties (Section 3.2, Appendix E) use the same composer, gate, prompts and judge with the language named in place of the dialect; the Lugbara build ran with a minimum of 12 sentences per article and the FLORES builds with a minimum of 4 (their sentences average 21 English words against 8 for MADAR), both in September 2026. The added systems are `Sunbird/Sunflower-Qwen3.5-9B` (Qwen3.5-9B with continued pretraining, instruction tuning and GRPO on Ugandan and other African-language data, Lugbara in its middle coverage tier; preview weights, revision `3537737`, downloaded 2026-09-21, served in bf16 by vLLM with thinking off and greedy decoding; its card does not list training corpora but evaluates on the SALT test split), `Sunbird/Sunflower-32B` (Qwen3-32B fine-tuned on 31 Ugandan languages, whose training mix includes SALT) and `PartAI/Dorna2-Llama3.1-8B-Instruct`. Sunflower-32B is loaded in 4-bit NF4 with bitsandbytes, and for these sets Qwen3-30B-A3B is served by vLLM with fp8 weight-only quantisation, because neither fits one of our 48 GB GPUs in bf16 and sharded bf16 inference on the shared machine produced degenerate output; both decode greedily, every other system as in the Arabic run. Summaries whose decomposition yields no claim are excluded from the means; with Lugbara input there are 120 such summaries for gpt-oss-20B, 57 for Qwen3-8B, 30 for Qwen3-30B-A3B and at most 7 for any other system. The reasoning-effort sweep of Figure 6 runs `openai.gpt-5.6-sol` through Bedrock Mantle at each effort it accepts (none, low, medium, high and xhigh; minimal is rejected) with `max_completion_tokens=32000` and the summarizer prompt unchanged, on 2026-09-23 and 2026-09-24.
- **Dialect specialists, excluded.** Nile-Chat-12B (Egyptian; Shang et al., 2025) and Atlas-Chat-9B (Moroccan; Shang et al., 2024) were run on an earlier article cohort and excluded because they return Arabic rather than an English summary.
- **Seeds.** PCA and KMeans use `random_state=0`; bootstrap confidence intervals and the paired-bootstrap tiers use 1,000 to 2,000 resamples with seed 0; the survey items were drawn with seed 20260909, and the shuffled condition moves blocks of $k = \max(2, \lfloor n/4 \rfloor)$ sentences, redrawn until at least 60% of the sentences change position.

- **Anchors and metrics.** AlignScore-large checkpoint (Zha et al., 2023) in `nli_sp` mode;`MoritzLaurer/mDeBERTa-v3-base-mnli-xnli`(He et al., 2023; Laurer et al., 2024); MiniCheck `flan-t5-large` (Tang et al., 2024); `intfloat/multilingual-e5-large-instruct` (Wang et al., 2024); chrF (Popović, 2015) from sacrebleu. The translation probe runs every open-weight system through vLLM 0.21 on one GPU in bfloat16 with greedy decoding (Qwen3-30B-A3B in fp8, as for its summaries; Sunflower-32B in 4-bit NF4 through transformers); the Arabic translations of gpt-oss-20B, Qwen3-8B, Fanar-1-9B and Aya-23-8B come from an earlier run of the same prompt through transformers.

We reduce the 4096-dimensional representation to $d = 50$ with PCA (Jolliffe & Cadima, 2016) before clustering, motivated by the connection between the principal components and a continuous relaxation of the K-means objective (Ding & He, 2004).

The size of the clusters can be adjusted according to the amount of parallel data available and the context window of the composer model to be used.

D Construction Ablation Detail

High-resource clustering. Our pipeline clusters the corresponding English side rather than the low-resource sentence itself. We evaluate this choice only on real articles, whose boundaries nobody in this project chose, where positive pairs come from the same article and negatives come from different articles, and measure whether compatible pairs rank above negatives:

$$\mathrm{AUC} = \Pr\left(s(p^+) > s(p^-)\right). \tag{2}$$

Here s is the cosine similarity between the embeddings of the two sentences of a pair $p = (i, j)$,

$$e_i = f(t_i; \iota), \qquad s(p) = \cos(e_i, e_j), \tag{3}$$

so a configuration fixes the side t_i that is embedded (the low-resource language or English) and the instruction ι (none or the scene instruction of Appendix C), and character n-gram TF-IDF on the same pairs gives a lexical baseline (Salton & Buckley, 1988). The articles are FLORES-200 pages (Appendix F), which every language shares, so each side is scored on identical pairs with each fragment's ten most similar fragments in English as hard negatives, and BBC articles from CrossSum (Bhattacharjee et al., 2023) in English and MSA, with hard negatives mined within each set.

On FLORES the low-resource embeddings reach 0.84 to 0.86 AUC on the Persian varieties and 0.86 to 0.90 on the eight Arabic dialects, above character TF-IDF at 0.73 to 0.78, but the English side reaches 0.92 on the same pairs (Table 4). We therefore cluster in English and transfer the resulting indices back to the corresponding low-resource sentences. This makes sense because smaller embedding models barely encounter these low-resource languages compared to the more high-resource languages.

Instruction conditioning. Prompting the English embedding model on a scenic instruction raises AUC from 0.924 to 0.943 on FLORES and from 0.722 to 0.734 on CrossSum, and ties the uninstructed embedding on CrossSum MSA (0.716 against 0.719), within the run-to-run variation of about 0.01.

Articles	Side	TF-IDF	Embedding	+ scene instruction
FLORES pages	English	0.728	0.924	0.943
	Persian	0.751	0.862	0.883
	Dari	0.777	0.857	0.876
	Tajik	0.735	0.841	0.870
	MSA Arabic	0.749	0.900	0.907
	Arabic dialects (mean of 8)	0.749	0.888	0.897
CrossSum (BBC)	English	0.584	0.722	0.734
	MSA Arabic	0.656	0.719	0.716

Table 4: AUC of same-article sentence pairs on real articles (Qwen3-Embedding-8B). On FLORES every language is scored on the same pairs, and English beats every low-resource side.

E TRANSFER RESULTS: OTHER LANGUAGE FAMILIES

The transfer runs use the Arabic composer, gate, prompts and judge with the language name swapped in (settings and quantisation in Appendix C); the Lugbara build kept 240 articles of 12 to 29 sentences (mean 14.9) from 13 rounds over 24,946 pairs at a cost of $79.95, and the FLORES sets are the recovered fragments of Appendix F.

Variety	Round 1	All rounds	Rounds	All / round 1
Levantine	82	435	12	5.3×
Lugbara	57	240	13	4.2×
Persian	85	232	7	2.7×
Dari	100	230	6	2.3×
Tajik	53	159	6	3.0×

Table 5: Accepted articles after the first round and after the last. The first rounds of the Egyptian, Gulf and Maghrebi builds were not logged.

Summaries from English score about 0.86 against English but about 0.56 against Lugbara for every system (Table 11), so only the English-reference score measures the summariser, and the Sunflower models' margins over GPT-5.5 with Lugbara input (0.03 for the 9B, 95% interval -0.02 to 0.08; 0.14 for the 32B, 0.09 to 0.18) come with a caveat, since SALT is in the 32B's training mix and the 9B card evaluates on it. Against the Lugbara article the judge passes 76% and 82% of the false Lugbara-input claims of GPT-5.5 and gpt-oss-20B but 38 to 48% of Aya-23-8B's and the Sunflower models', and rejects 29 to 31% of the Sunflower models' true claims against 10 to 12% of the GPT family's.

Acholi and Ateso. SALT is multi-way parallel, so replacing every sentence of the 240 Lugbara articles by its Acholi or Ateso translation gives articles with the identical English side (all 3,571 sentences map), which every system summarized with its Lugbara settings. The Sunflower models stay near 0.65 on Acholi and Ateso, level with GPT-5.5 on Ateso (9B $+0.04$, 95% interval 0.00 to 0.09) and below it on Acholi, perhaps because Acholi is a Luo language close to Dholuo, which FLORES-200 covers.

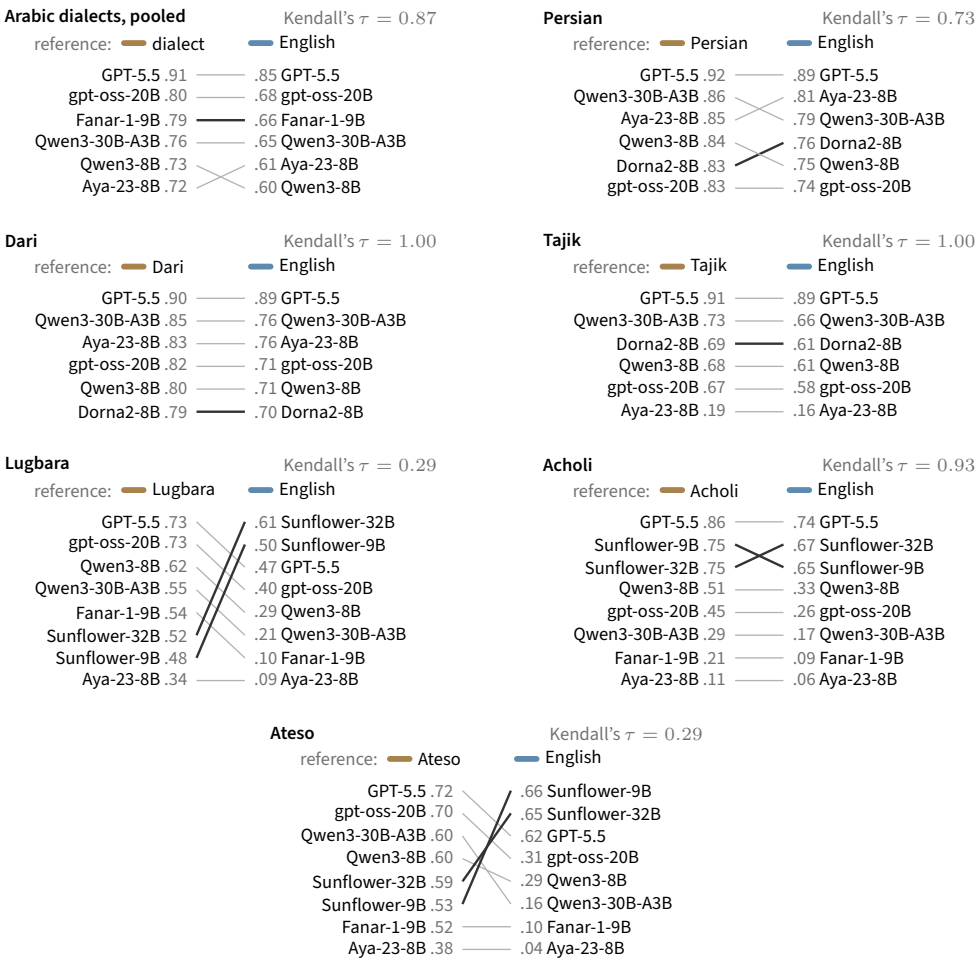

Figure 5: Source-input summaries ranked against the source-language article (left) and its English version (right) in every build; on Lugbara the Sunflower models move from sixth and seventh to first and second.

Figure 6 scores gpt-5.6-sol's Lugbara-input summaries at every reasoning effort it accepts against the English reference, since the judge cannot verify claims against a Lugbara article (Table 7). More effort helps up to medium, from 0.37 without reasoning to 0.48, and little beyond it (0.49 at xhigh).

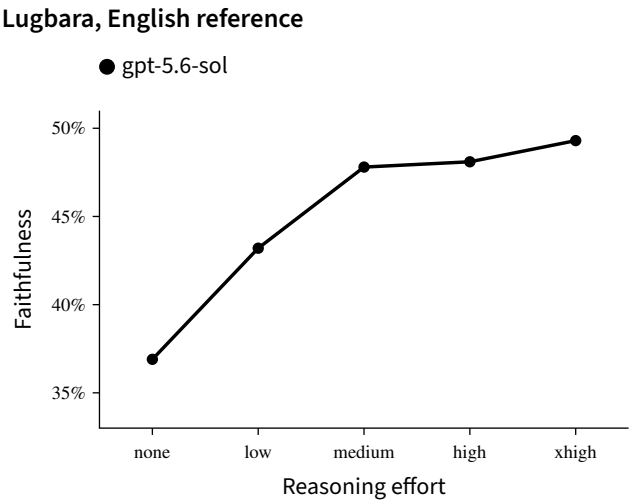

Figure 6: gpt-5.6-sol's faithfulness with Lugbara input at each reasoning effort, against the English reference (95% intervals about ± 3 points).

F Recovery Control on FLORES

FLORES-200 sentences come in fragments of 3.6 consecutive sentences on average from 562 known pages, so running the pipeline on the flat pool with the grouping withheld (GPT-5.5 composer, Sonnet 4.6 gate, minimum four sentences, six rounds, seven for Persian) tests what it recovers: Table 6 shows that 92 to 94% of the returned articles are exactly one original fragment, no article mixes unrelated pages, and when asked for articles that span sources the pipeline returned nothing in any of the three languages.

	Persian	**Dari**	**Tajik**
Sentences in the pool	2,009	2,009	2,009
Source fragments (3.6 sentences each)	562	562	562
Fragments with at least 4 sentences	275	275	275
Articles returned	232	230	159
Articles made of a single source fragment	214 (92%)	214 (93%)	150 (94%)
Of which the whole fragment	156	160	122
Of which in the original sentence order	198	202	145
Mean share of sentences from the main source	0.973	0.975	0.982
Distinct qualifying fragments recovered	214 (78%)	214 (78%)	150 (55%)
Cost (composer and gate)	\$9.4	\$7.9	\$10.4
Articles required to span sources	0	0	0

Table 6: With the FLORES grouping withheld, the pipeline reassembles the original fragments and composes nothing across sources.

G Verbatim Copies of the Source Sentences

After normalising whitespace and punctuation, 18,894 of the 18,940 Arabic sentences (99.76%) are exact copies of their pool entries, 43 differ only by a unified proper noun and 3 by meaning-preserving trims. On the transfer builds 98.9 to 99.9% of the source sentences and 99.9 to 100% of the English sentences are exact copies.

H JUDGE VALIDATION DETAIL

Judge choice and cross-judging. The gate and the judge are always a different model family from the composer, since LLM judges can favor outputs of related models (Li et al., 2025a), and they are frontier models rather than small ones, since prior work reports model-dependent scoring biases in LLM judges (Li et al., 2025b; Norman et al., 2026) and ReIFE finds that the ranking of base models is largely stable across protocols while a protocol's effect depends on the base model (Liu et al., 2025). We therefore spend the choice on the base model and keep the protocol standard: a FActScore-inspired decompose-and-verify judge (Min et al., 2023) (Section 4). All LLM judging was validated against human annotations (Sections 3.2 and 4).

Judge evaluation on the released articles. Table 7 scores each claim's faithful-or-not verdict against the source article against its verdict against the English article, over all systems of each build. The two verdicts agree moderately on the Arabic dialects and Persian varieties, where the source reference still passes about half of the claims the English reference rejects except on Tajik and Acholi, and barely above chance on Lugbara and Ateso, where the two rankings also come apart.

	Cohen's κ		**Source-input claims**		
Build	source input	English input	false passed	true rejected	**Ranking τ**
Arabic dialects	0.54	0.40	47%	4%	0.87
Persian	0.51	0.47	53%	3%	0.73
Dari	0.57	0.48	49%	2%	1.00
Tajik	0.79	0.45	21%	3%	1.00
Lugbara	0.21	0.06	58%	19%	0.29
Acholi	0.63	0.17	27%	9%	0.93
Ateso	0.18	0.05	58%	24%	0.29

Table 7: Agreement of the judge's verdicts against the source and the English article on the same claims, with the English verdict as truth. False passed and true rejected are for source-input claims; τ compares the systems' rankings under the two references.

I AUTOMATIC ANCHORS

All anchors are open-weight models run on every judged article and summary: AlignScore-large and MiniCheck with the English article as premise, mDeBERTa-v3-XNLI with the best sentence of either article, and multilingual-e5 retrieval of the closest sentence of either article. Scored claim by claim against the judge's English-reference verdict (Table 8), every anchor that uses the source article is weaker than its English counterpart, and MiniCheck on the English article reaches AUC 0.81 to 0.93.

Checker	**Arabic**	**Persian**	**Dari**	**Tajik**	**Lugbara**
Uses the source article					
Judge, source reference	0.737	0.739	0.751	0.869	0.569
mDeBERTa, source sentences	0.585	0.598	0.588	0.668	0.469
e5 evidence, source sentences	0.549	0.553	0.567	0.619	0.462
Uses the English article					
MiniCheck	0.813	0.826	0.858	0.916	0.929
mDeBERTa, English sentences	0.673	0.681	0.706	0.762	0.785
e5 evidence, English sentences	0.648	0.625	0.652	0.777	0.871
Claims	132,932	16,013	16,140	10,639	20,647

Table 8: AUC of each anchor against the judge's English-reference verdict, per claim. Every anchor that uses the source article is weaker than its English counterpart.

AlignScore, MiniCheck and English-side e5 rank the systems' source-input summaries in the order of the English-reference judge in every build (Table 9), including the Lugbara reversal of Figure 5,

which therefore does not rest on the Claude judge; no anchor that uses the source article exceeds $\tau = 0.20$.

Ranks the systems by	Arabic	Persian	Dari	Tajik	Lugbara
Judge, source reference	0.87	0.73	1.00	1.00	0.29
AlignScore-large, English article	0.60	0.73	1.00	0.73	0.71
MiniCheck, English article	0.73	1.00	0.60	0.87	0.93
e5 evidence, English sentences	0.73	0.47	0.60	0.87	0.79
mDeBERTa, English sentences	−0.20	0.47	−0.07	0.73	0.86
mDeBERTa, source sentences	−0.33	0.07	−0.20	−0.20	−0.29
e5 evidence, source sentences	−0.33	−0.60	0.07	0.20	−0.50

Table 9: Kendall τ between each anchor's system ranking and the judge's English-reference ranking. Only anchors that read the English article agree with it.

J Error Types

The error type is the judge's own tag for every claim below supported (Appendix M), and Table 10 gives each tag's share of all claims, judged against the English article. Of the 4,293 entity-tagged dialect-input claims, 16% give a name spelled like one in the article (e.g. Fahd for Fahad), which bounds transliteration at about one in six, and 26% give a name absent from the article.

Error tag	Arabic	Persian	Dari	Tajik	Lugbara
entity	0.062 / 0.003	0.064 / 0.003	0.089 / 0.002	0.091 / 0.003	0.071 / 0.008
number	0.022 / 0.003	0.009 / 0.004	0.008 / 0.004	0.013 / 0.005	0.007 / 0.003
unsupported	0.132 / 0.054	0.053 / 0.035	0.061 / 0.029	0.198 / 0.035	0.536 / 0.044
contradiction	0.028 / 0.011	0.011 / 0.005	0.013 / 0.006	0.024 / 0.005	0.012 / 0.006
relation	0.073 / 0.059	0.051 / 0.038	0.058 / 0.037	0.071 / 0.033	0.032 / 0.074
overgeneralization	0.031 / 0.023	0.047 / 0.050	0.048 / 0.049	0.049 / 0.055	0.065 / 0.034

Table 10: Per-claim error rates with source-language / English input, judged against the English article. Entity errors rise in every language, and half of the Lugbara-input claims are unsupported.

K Full Scores

Table 11 gives the faithfulness of both summaries against both articles for every system and variety; Figure 2 plots the gap between its English-reference scores.

		Persian varieties						Arabic dialects								Nilo-Saharan					
		Persian		Dari		Tajik		Egyptian		Gulf		Levantine		Maghrebi		Lugbara		Acholi		Ateso	
System	$r \backslash g$	src	En	src	En	src	En	src	En	src	En	src	En	src	En	src	En	src	En	src	En
GPT-5.5	src	.92	.89	.90	.89	.91	.89	.92	.87	.91	.85	.90	.86	.89	.83	.73	.47	.86	.74	.72	.62
	En	.89	.92	.89	.93	.89	.92	.89	.93	.86	.92	.86	.93	.84	.92	.59	.90	.80	.90	.66	.90
Qwen3-30B-A3B	src	.86	.79	.85	.76	.73	.66	.80	.72	.77	.67	.76	.66	.73	.57	.55	.21	.29	.17	.60	.16
	En	.86	.89	.87	.88	.86	.88	.87	.89	.83	.87	.82	.88	.80	.88	.54	.88	.79	.88	.62	.88
gpt-oss-20B	src	.83	.74	.82	.71	.67	.58	.83	.73	.78	.68	.80	.67	.80	.65	.73	.40	.45	.26	.70	.31
	En	.85	.88	.87	.89	.86	.89	.82	.86	.80	.84	.83	.85	.81	.86	.55	.84	.77	.84	.64	.84
Qwen3-8B	src	.84	.75	.80	.71	.68	.61	.75	.66	.74	.61	.72	.61	.70	.54	.62	.29	.51	.33	.60	.29
	En	.87	.88	.87	.88	.85	.89	.85	.86	.81	.85	.84	.86	.81	.86	.58	.86	.77	.86	.61	.86
Aya-23-8B	src	.85	.81	.83	.76	.19	.16	.78	.69	.73	.62	.71	.62	.69	.55	.34	.09	.11	.06	.38	.04
	En	.84	.86	.84	.85	.82	.84	.76	.79	.76	.79	.77	.78	.75	.79	.55	.83	.76	.83	.62	.83
Specialist	src	.83	.76	.79	.70	.69	.61	.83	.72	.79	.67	.78	.66	.77	.60	.48	.50	.75	.65	.53	.66
	En	.84	.84	.85	.86	.84	.87	.80	.84	.80	.84	.80	.85	.78	.85	.53	.84	.74	.84	.60	.84
Sunflower-32B	src															.52	.61	.75	.67	.59	.65
	En															.54	.86	.77	.86	.60	.86

Table 11: Faithfulness of summaries from the source (r=src) or English (r=En) input, checked against the source (g=src) or English (g=En) article. Specialists: Dorna2-8B (Persian), Fanar-1-9B (Arabic) and Sunflower-Qwen3.5-9B (Nilo-Saharan, whose English-input entries repeat), plus Sunflower-32B on the Nilo-Saharan languages.

L Article Length

The released articles have 7 to 28 sentences (mean 8.1) and 25 to 329 English words (mean 74); the minimum sentences per article and the cluster size are set when the pipeline is run, so longer articles cost yield rather than a change of method, and Table 12 splits the 2,170 judged articles into quartiles of English length.

Quartile	**Words**	**Sent.**	GPT-5.5	gpt-oss	Qwen3-30B	Fanar	Qwen3-8B	Aya	**Gap**
Q1 (shortest)	25–58	7.3	0.922	0.840	0.829	0.834	0.793	0.769	0.070
Q2	59–68	7.7	0.912	0.822	0.828	0.816	0.789	0.753	0.076
Q3	69–83	8.2	0.916	0.825	0.816	0.804	0.797	0.757	0.074
Q4 (longest)	84–329	9.3	0.902	0.795	0.800	0.801	0.790	0.738	0.067

Table 12: Faithfulness by quartile of English article length; the faithfulness gap stays near seven points.

M Where the Error Taxonomy Comes From

The six error tags adapt the hallucination taxonomy of AraHalluEval (Alansari & Luqman, 2025) to a claim-versus-single-source setting (Table 13).

Tag	**Definition**	**AraHalluEval analogue**
entity	wrong or invented person, place, organisation or named thing	Named Entity (direct)
number	wrong or invented quantity, date, amount or count	Value (direct)
unsupported	content absent from the source	Fabrication, partly Context (direct)
overgeneralization	broader or stronger than the source supports	Inference (close)
contradiction	asserts the opposite of the source	Context Inconsistency (close)
relation	right entities, wrong relationship	none (extension)

Table 13: Each error tag maps onto a category of the AraHalluEval taxonomy.

N QUALITATIVE EXAMPLES

An example cluster (Lugbara). Clustering gathers sentences on one topic, from which the composer picks a single scene. One cluster of the first round (recomputed with the settings of Appendix C) holds 129 of the 24,946 Lugbara pairs, all about health in Uganda; its English side includes "Uganda is safe and has no Ebola cases", "Entebbe International Airport was closed as a measure to curb the spread of coronavirus", "The Uganda blood bank encourages people to donate blood" and "Cancer remains a silent killer in Uganda; therefore, we should all be screened". From it the composer selected the fourteen sentences about one scene, the second wave of coronavirus in Uganda, which became benchmark article lgg-0203, beginning "Uganda has been hit by the second wave of coronavirus. There were more severe cases in the second wave of coronavirus in Uganda. How many coronavirus cases are currently in Uganda?"

A Lugbara article that reads like news. SALT's sentences were written to cover "a range of topics of local relevance, such as agriculture, health and society" (Akera et al., 2022), so the coherence gate rates the Lugbara articles close to the FLORES fragments of real documents and well above the Arabic phrasebook articles (Table 14). Benchmark article lgg-0024, rated 5 out of 5, is one of them; its English side in full:

> Museveni and Kagame were once comrades. Wars in Uganda, Rwanda, and the Democratic Republic of Congo were commanded by Museveni and his lieutenant Kagame. Since becoming president, Kagame has sought to break away from being Museveni's lieutenant. The Rwanda leadership now views Museveni as an enemy. It is difficult for leadership in Kigali to trust Museveni's commitment to any agreement. The Daily Monitor newspaper in Kampala in December twenty eighteen quoted Kagame. Rwandan President Paul Kagame spoke strongly about people trying to destabilize Rwanda. Kagame promised to raise the cost on anybody who wants to destabilize our security. Kagame's remarks were a broad sweep. Others said he aimed at either Burundi or Uganda. According to analysts, Kagame's tone was a harbinger of doom for the upcoming summit. The meeting was to follow up on an earlier meeting in Kigali. The meeting was aimed at resolving the stand-off with Uganda. He added that Uganda was determined to keep Rwanda on the negotiating table. Oryem says Uganda is willing to give talks another chance irrespective of changes in dates. Henry Okello Oryem is Uganda's minister of state for international affairs. Rwanda's minister of state wrote a letter to the bloc's secretary-general. State Minister for East African Community told about the false accusations made by Burundi. Uganda's Ministry of Foreign Affairs delivered a diplomatic protest in November. Uganda's Ministry of Foreign Affairs delivered a diplomatic protest. President Museveni and Paul Kagame have blocked open public discussion on the issues.

These ratings come from the gate, not from native speakers; the human study covers the Arabic dialects only.

Build	Coherence	Rated 5	Reads like a document	Summarizability	Sent./art.
Arabic dialects	4.07	7%	2.70	2.93	8.1
Lugbara	4.71	71%	3.58	4.35	14.9
Persian (FLORES fragments)	4.89	89%	3.95	4.41	4.3
Dari (FLORES fragments)	4.91	91%	3.92	4.43	4.4
Tajik (FLORES fragments)	4.95	95%	4.01	4.52	4.4

Table 14: The coherence gate's 1–5 ratings of every accepted article. Lugbara articles rate close to the FLORES fragments of real documents and well above the Arabic ones.

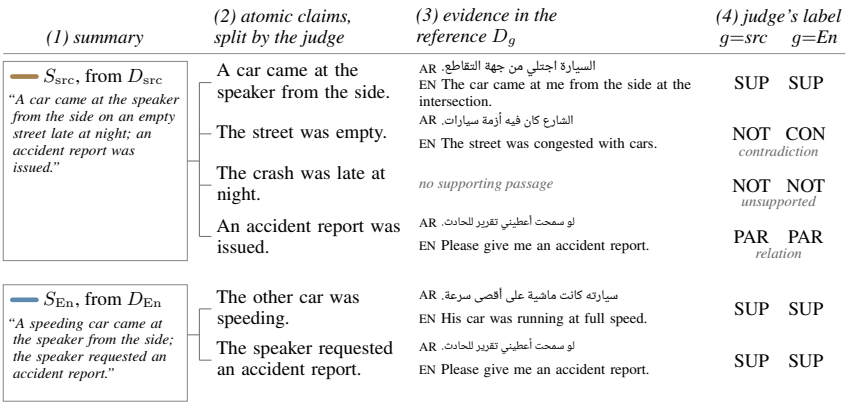

Figure 7: Example of faithfulness evaluation, with Arabic as the low-resource language. Each of the two summaries, S_{src} from the low-resource input and S_{En} from the English input, is decomposed once into its atomic claims. Every claim is then checked against the low-resource reference D_{src} and against the English reference D_{En}, receiving a label, the evidence span the label rests on, and an error tag (gray) for anything below supported: entity, number, unsupported, overgeneralization, contradiction or relation (Appendix M). Against the English reference the claims of S_{src} receive one of each label; against the Arabic reference the judge misses that the street was congested, the kind of failure the judge evaluation isolates.

O Translation Quality and Faithfulness

Each open-weight system also translated every source sentence of the judged articles into English (prompt in Appendix B, greedy decoding), scored with chrF against the human English and averaged per article (Table 15). Over all 27 system-build pairs mean chrF and $F_{\mathrm{src,En}}$ correlate at $r = 0.93$, and on Lugbara only the two Sunflower models translate at all (chrF 47 and 54 against 17 to 21).

System	Arabic	Persian	Dari	Tajik	Lugbara
gpt-oss-20B	49.7 / 0.68	61.7 / 0.74	60.4 / 0.71	50.1 / 0.58	18.7 / 0.40
Qwen3-30B-A3B	43.3 / 0.65	62.0 / 0.79	60.1 / 0.76	48.6 / 0.66	17.1 / 0.21
Qwen3-8B	41.0 / 0.60	60.4 / 0.75	59.2 / 0.71	48.6 / 0.61	18.2 / 0.29
Aya-23-8B	43.4 / 0.61	70.6 / 0.81	67.4 / 0.76	26.9 / 0.16	18.0 / 0.09
Fanar-1-9B	50.4 / 0.66	–	–	–	20.6 / 0.10
Dorna2-8B	–	60.1 / 0.76	58.4 / 0.70	48.7 / 0.61	–
Sunflower-9B	–	–	–	–	47.1 / 0.50
Sunflower-32B	–	–	–	–	54.0 / 0.61
r across systems	0.86	0.78	0.65	0.98	0.83

Table 15: Mean chrF of each system's sentence translations / its faithfulness with source input (– where not run). Last row: Pearson r across a build's systems.

P Licensing

Table 16 lists every text source the release draws on. Model outputs, extracted claims, judge labels, metric scores, and the survey instrument are our own and ship with the benchmark.

Source	License	Redistributed	Modifications
AraBench public text release (Sajjad et al., 2020): the MADAR-derived sentence pairs of 25 cities (Bouamor et al., 2018) and the QAraC Qatari transcripts (Elmahdy et al., 2014)	none stated (public text release); MADAR and QAraC credited	the sentences and English references of 2,320 of the 2,330 articles, with a city code and scene identifier per article	selection and ordering only; proper-noun unification on 110 sentences, applied to both sides
Multidialectal Arabic Parallel Corpus (MDC) (Bouamor et al., 2014)	research only, no redistribution	nothing: the 10 articles containing its 36 sentences are withheld	–
CrossSum, BBC Arabic articles (Bhattacharjee et al., 2023)	CC BY-NC-SA 4.0	the 16 natural-news survey items (4 per dialect) and the article pairs of the clustering diagnostic	none; verbatim excerpts, non-commercial terms apply to these items
FLORES-200 (NLLB Team et al., 2022), Persian, Dari and Tajik pools	CC BY-SA 4.0	articles from the transfer runs	selection and ordering only
SALT (Akera et al., 2022), Lugbara, Acholi and Ateso pairs with English	CC BY-SA 4.0	the 240 Lugbara articles and their Acholi and Ateso versions	selection and ordering only

Table 16: Every text source except MDC is redistributed under its own terms, and no source sentence was rewritten beyond proper-noun unification.