The Illusion of Fluency: Evaluating Figurative and Cultural Understanding in Multilingual LLMs with AFRI-PROVERB

Anonymous ACL submission

Abstract

The rapid global deployment of Multilingual Large Language Models (MLLMs) underscores the urgent need to evaluate their ability to understand culturally grounded figurative language, particularly for African proverbS, which are largely underrepresented in existing multilingual evaluation benchmarks. To address this gap, we introduce **AFRI-PROVERB**, a first large-scale multilingual evaluation dataset designed to assess both literal translation accuracy and figurative understanding of proverbs across East and Central African languages. **AFRI-PROVERB** comprises 8,415 curated proverbs collected from six African countries, covering 44 African languages. We evaluate modern LLMs using prompts on four translation tasks (literal and figurative) across two target languages: Swahili and English. We also present a comprehensive human evaluation framework and LLM-as-a-Judge to assess MLLMs performance. Our results reveal a substantial performance gap: current models struggle to interpret implicit cultural cues, despite different architectures they struggle to preserve the profound figurative and cultural significance. Critically, human analysis demonstrates that current MLLMs generate fluent translations, but frequently fail to preserve the intended meaning accuracy, adequacy and cultural grounding of African proverbs, models often rely on memorized patterns from high-resource languages, producing generic moral statements or proverb-like templates instead of recovering the intended meaning.

1 Introduction

Language is central to human communication, shaping expression, identity, and the transmission of cultural knowledge. In many African societies, languages carry communal values, social norms, and historical memory, with meaning often implicit, figurative, and shaped by oral traditions such as proverbs. African proverbs are culturally grounded,

"A man who calls his kinsmen to a feast does not do so to save them from starving. They all have food in their own homes. When we gather together in the moonlit village ground it is not because of the moon. Every man can see it in his own compound. We come together because it is good for kinsmen to do so."

— Chinua Achebe, Things Fall Apart

figurative expressions that convey meaning beyond their literal form and play a central role in communication, education, and storytelling (Martin, 2024). For instance, the Swahili proverb *"Kikulacho ki nguoni mwako"* ("What eats you is in your clothes") does not describe a physical threat, but conveys that harm often comes from those closest to us. Interpreting such expressions requires cultural knowledge and pragmatic inference beyond surface-level linguistic competence. Consequently, understanding proverbs requires models to go beyond surface-level pattern matching to culturally grounded interpretation (Thapa et al., 2025; Li et al., 2024; Hu et al., 2020), making proverbs a rigorous testbed for evaluating whether multilingual LLMs capture culturally grounded meaning or rely on superficial fluency (Thapa et al., 2025).

Despite tremendous progress on LLMs development on various NLP tasks (Adelani et al., 2022; Wang et al., 2024; Etori and Gini, 2024; Romero et al., 2024; Nimo et al., 2025), their ability to process culturally grounded and figurative language in African languages remains critically underexplored. These "culture gaps" (Liu et al., 2023) are particularly evident in proverb translation, where preserving figurative meaning exposes limitations in multilingual LLMs (Constant et al., 2017). Prior work (Ghosh and Srivastava, 2022; Liu et al., 2024; Thapa et al., 2025; Korre and Pavlopoulos, 2025; Magdy et al., 2025; Azime et al., 2025) has shown that proverbs are effective for evaluating figurative understanding in LLMs, their use for assessing culturally grounded reasoning in African languages remains limited. Table 1 compares our work

Dataset	Coverage	# Languages.	# African.	# E.Africa	# Pairs
ePiC (Ghosh and Srivastava, 2022)	English-only	1	0	–	2,750
PRONE (Thapa et al., 2025)	Nepali-only	1	0	–	2,830
ProverbEval (Azime et al., 2025)	Ethiopian + English	5	5	✓	6,854†
Proverbs Run in Pairs (Wang et al., 2025)	5 high-resource languages	5	0	–	424
P.Oracles? (Korre and Pavlopoulos, 2025)	Greek-only	1	0	–	–
Jawaher (Magdy et al., 2025)	Arabic-only	1	0	–	10,037
MAPS (Liu et al., 2024)	5 high-resource languages	5	0	–	4,446
Afri-Proverb (Ours)	**East African languages**	**45**	**45**	✓	**8,415**

Table 1: Total data statistics for AFRI-PROVERB compared with existing proverb benchmarks. † Not parallel pairs (monolingual).

with existing proverb benchmarks. To address the gaps, we introduce **AFRI-PROVERB**, first large-scale multilingual evaluation benchmark covering **8,415 proverbs in 44 East and Central African languages** across two target languages. **AFRI-PROVERB** expands proverb-based evaluation with literal translation and figurative interpretation. Automatic evaluation show current LLMs perform substantially better at *literal translation* than *figurative interpretation*. Human evaluation further reveals an ***illusion of fluency***: models consistently achieve higher fluency (2.28 ± 0.73) than semantic accuracy scores (1.73 ± 0.69), indicating that grammatically fluent and proverb-like outputs often fail to preserve the intended figurative meaning of the source proverb. Our contributions are as follows:

- We introduce **AFRI-PROVERB**, a first large-scale multilingual benchmark comprising **8,415 proverbs across 44 East and Central African languages**, for evaluation of culturally grounded figurative language understanding.
- We curate and computationally align existing proverb collections into a structured multilingual benchmark that distinguishes *literal translation* from *figurative interpretation* across two target languages, **Swahili and English**.
- We evaluate modern LLMs in both zero-shot and few-shot settings, alongside human evaluation, to assess the accuracy literal and figurative understanding. We further compare human and **LLM-as-a-judge** evaluations to examine the reliability of automatic judgment for culturally grounded proverb translation and interpretation.
- We introduce a structured **error taxonomy** for understanding model failure modes in African proverb translation and interpretation.

2 Related Work

Numerous studies show that proverbs encode culturally grounded meanings through metaphor, including work on Kipsigis (Soi, 2014), Swahili (Kobia, 2016), Bukusu and Gusii (Barasa and Opande, 2017), Yoruba (Adisa, 2024), English and Spanish (López Rodríguez, 2009), Persian and Turkish (Pourhossein, 2016), and Tunisian Arabic (Faycel, 2012). These studies underscores the central role of metaphor in proverb interpretation and the challenge it poses for machine translation (Shutova, 2015; Özbal et al., 2016; Wang et al., 2025). Recent MT research identifies three persistent challenges in figurative language: the gap between translation and interpretation (Wang et al., 2025), low-resource language barriers (Azime et al., 2025; Magdy et al., 2025), and inadequate evaluation metrics (Wang et al., 2025). Despite work on simile and metaphor detection, generation, and interpretation, MLLMs continue to struggle with proverb understanding, revealing broader weaknesses in figurative reasoning and cultural grounding (Liu et al., 2024; Thapa et al., 2025; Wang et al., 2025). A major limitation remains the lack of standardized evaluation benchmarks. ProverbEval (Azime et al., 2025) established an important baseline for African proverbs evaluation, but is limited to five Ethiopian languages. To address this, we introduce AFRI-PROVERB, a multilingual benchmark covering 44 African languages for evaluating literal translation and figurative understanding in African proverbs.

3 AFRI-PROVERB

Developing multilingual models capable of understanding and translating African proverbs requires high-quality culturally grounded local data from diverse linguistic and geographic contexts. We introduce AFRI-PROVERB, a benchmark for evaluating proverb translation and figurative interpretation across 44 East and Central African languages. The dataset contains 8,415 proverbs, curated from culturally grounded sources and validated by native speakers and human annotators. Each proverb is paired with both literal and figurative interpretations in two target languages, $t \in \{\text{Eng, Swa}\}$. Literal translations $y \in \{\text{eng_lit, swa_lit}\}$ preserve surface meaning, while figurative interpretations $y \in \{\text{eng_fig, swa_fig}\}$ capture the underlying cultural and semantic intent (Figure 1).

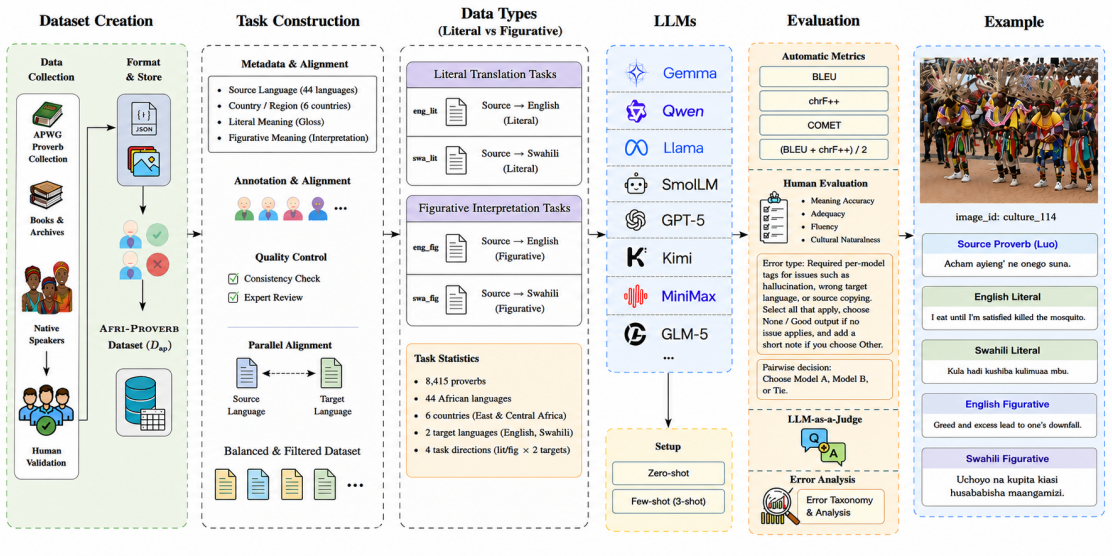

Figure 1: Methodology overview of **AFRI-PROVERB**.

3.1 Data Collection

We constructed **AFRI-PROVERB** by systematically aggregating proverb resources from published and carefully curated written sources. The dataset was compiled from three complementary data sources

(i) Online and published resources. The first step involved searching for existing proverb collections using language-specific keywords across websites, archival repositories, and electronic books. A primary source was African Proverbs Working Group (APWG)[1], which provides proverb collections curated by native speakers and cultural experts. which contained diverse proverb collections curated by native speakers, educators and local cultural experts from communities. We cover diverse communities across Kenya, Uganda, Tanzania, Ethiopia, and Democratic Republic of Congo (DRC).

(ii) Archival books and institutional libraries. To increase coverage for selected languages, we incorporated larger proverb collections from monographs and digitized books obtained through archival [2]and local University library. This expanded coverage forKamba and Meru languages.

(iii) Native speaker involvement. Native speakers played a central role in data collection, verification, and annotation. They confirmed the validity of proverb entries, resolved ambiguous cases, and validated orthographic conventions and culturally appropriate interpretations. Across all sources, annotators preserved proverbs verbatim as in the source to maintain idiomatic structure, dialectal variation, and culturally situated usage. See sources and language coverage in Table 31 and Appendix. G.

3.2 Quality Control

To ensure high data quality and consistency, all proverb entries underwent manual verification by native speakers. Annotators reviewed the source proverbs for linguistic accuracy, orthographic consistency, and clarity of interpretation, while validating both the English and Swahili literal and figurative equivalents across tasks. Ambiguous or inconsistent entries were corrected or removed during preprocessing. Disagreements were resolved with native speakers based on linguistic validity, cultural appropriateness, and fidelity to the source material. To preserve cultural and linguistic fidelity, proverbs were retained as close as possible to their original forms (see § 3.1).Human evaluation was conducted by native speakers sourced from the Masakhane and online community.

3.3 Dataset Structure

Each proverb entry comprises of five aligned fields as described in § 3: (1) the source proverb in its original language, (2) an English literal translation, (3) a Swahili literal translation, (4) an English figurative interpretation, and (5) a Swahili figurative interpretation. The reason for aligning design is threefold: (1) to explicitly separate the proverb's surface wording from its intended semantic meaning, (2) to enable fine-grained analysis of the semantic drift between literal and figurative interpretations, and (3) to support cross-lingual examination of cultur-

[1] https://afriprov.tangaza.ac.ke/
[2] https://archive.org/

ally embedded expressions while providing a controlled setting for evaluating whether LLMs can retain culturally grounded meanings as opposed to surface-level cues (Figure 1).

3.4 Proverb Annotation and Normalization

After data collection (§ 3.1) each proverb entry is subjected to a structured annotation and normalization process prior to evaluation, ensuring linguistic accuracy, semantic consistency, and cross-lingual comparability(§ 3.2). Annotators performed the following: (i) verified the correctness of the source proverb and flagged any unclear or corrupted entries; (ii) validated, or added a *literal translation* in Swahili and/or English if missing from the source; and (iii) validated, or added a *figurative interpretation* in Swahili and/or English if absent from the source. Figurative interpretations in our dataset are expressed at an abstract level that explicitly states the underlying principle conveyed by each proverb, rather than reproducing metaphorical imagery or culturally specific surface forms. Annotators were instructed to remove animals, objects, and physical actions present in the source proverb and to replace them with plain, general statements that capture the intended semantic function. As shown in Fig. 6 and described in Appendix § J.

3.5 Data Preprocessing

We applied a systematic preprocessing pipeline to normalize the raw proverb collection, motivated by the need to ensure valid alignment, prevent evaluation leakage, and disentangle surface translation from cultural reasoning. This included (i) removing exact and near-duplicate entries to prevent redundancy and evaluation bias; (ii) enforcing one-to-one alignment between each source proverb and its Latin transliteration; (iii) standardizing literal translations to surface-level meaning by removing editorial notes and interpretive commentary; and (iv) abstracting figurative interpretations into context-independent meanings. as discussed in § 3.4.

3.6 Data statistics

As described in § 3, **AFRI-PROVERB** comprises 8,415 curated proverbs collected across 6 African countries (Kenya, Tanzania, DRC, Uganda, Ethiopia, and Somalia), covering 44 African languages. Each proverb entry contains four aligned fields (§ 3.3). Eleven languages have 200–500 proverbs; while the rest each have ∼100 proverbs.

Country	Lang.	Prov.	Tok. (Prov)	Tok. (All)	Avg
Kenya	18	4,904	91,044	321,954	18.6
Tanzania	8	1,203	14,782	87,973	12.3
DRC	9	908	13,086	65,235	14.4
Uganda	7	700	9,606	48,169	13.7
Somalia	1	500	7,420	35,084	14.8
Ethiopia	2	200	2,695	15,388	13.5

Table 2: Country-level statistics of AFRI-PROVERB.

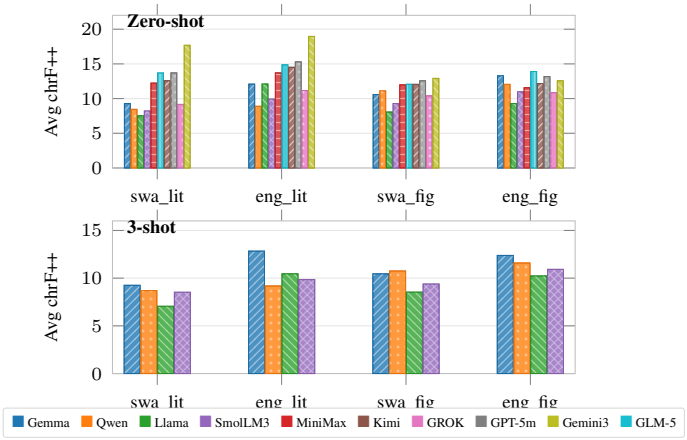

Figure 2: Task-level average performance (zero-shot and few-shot) across models (BLEU + ChrF++)/2.

The corpus contains 573,803 tokens overall, including 138,633 source-language tokens, with an average proverb length of 16.5 tokens. Dataset coverage is uneven due to data availability. Kenya contributes the largest portion of the corpus (4,904 proverbs; 18 languages), followed by Tanzania (1,203; 8), DRC (908; 9), Uganda (700; 7), Somalia (500; 1), and Ethiopia (200; 2). Average proverb length varies across countries (12.3–18.6 tokens), with most languages averaging 10–15 tokens per proverb, although some outliers such as Nubian exhibit longer sequences (Tables 2, 9, and 10). Dataset completeness exceeds 99% across aligned fields, spans across multiple language families (Table 3).

3.7 Data Sources

We discussed data sources in § 3.1. Come in different formats such as ebooks, books, websites, and scholarly publications. Table 8 shows all data sources, with pairs, origin, authorship, and format.

[2] https://www.ethnologue.com/
[2] https://glottolog.org/

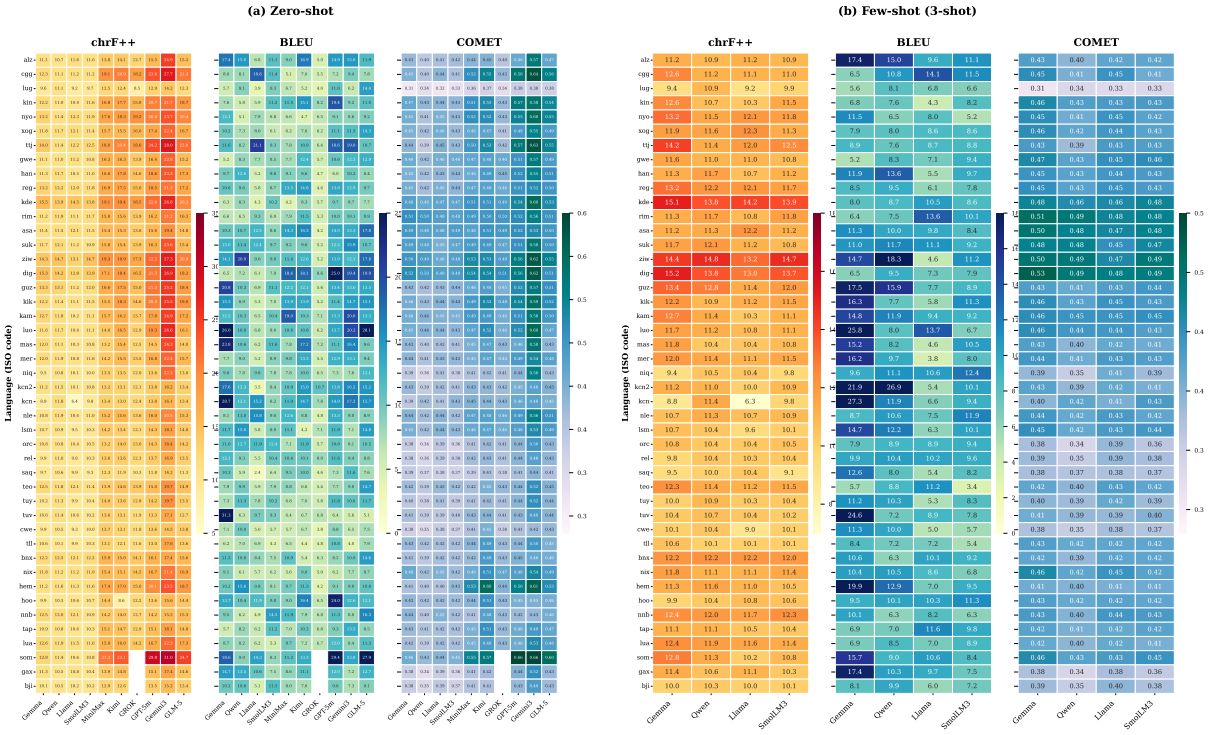

Figure 3: Average performance (zero-shot and few-shot) across languages and models. Few-shot generally improves performance. However, not uniform all across languages. COMET scores for literal tasks only.

Category	Count	Percent
Language Family (N=44)		
Niger–Congo	30	65.9%
Nilo–Saharan	8	18.2%
Afro–Asiatic	5	9.1%
Creole	1	2.3%
Vitality Status		
Stable	25	56.8%
Institutional	15	34.1%
Endangered	4	9.1%
Digital Support		
Ascending	25	56.8%
Emerging	15	34.1%
Vital	3	6.8%
Still	1	2.3%

Table 3: **Distribution of the 44 languages in the dataset.**

4 Tasks Overview

We design our tasks to evaluate proverb understanding as a structured progression from surface-level translation to deeper figurative interpretation. Because proverbs rely on imagery, metaphor, and culturally grounded inference, we explicitly separate these stages rather than treating proverb translation as a single task. We evaluate four tasks defined by two target languages $t \in \{\text{English, Swahili}\}$ and two output types $y \in \{\text{literal, figurative}\}$.

Literal Translation (LMT). Given a source proverb p in a native language, the model generates a literal translation $\hat{y}_{\text{lit}} = f(p, t)$ in the target language t, preserving wording, imagery, and surface structure without interpretation, even if the output is unnatural. This setting isolates the model's ability to perform direct lexical and structural mapping.

Figurative Interpretation (FMI). Given a source proverb p, the model generates its figurative meaning $\hat{y}_{\text{fig}} = f(p, t)$ in the target language t, requiring abstraction beyond surface form to express the underlying meaning. This evaluates cultural and inferential understanding, as correct outputs often diverge significantly from the literal form.

4.1 Evaluation Metrics

We evaluate translation quality of the proverb using BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and COMET(Rei et al., 2020). We report BLEU, ChrF++, and COMET for the `swa_lit` and `eng_lit` tasks, and BLEU and ChrF++ for the `swa_fig` and `eng_fig` tasks. We also computed *Row Avg* as the mean of $(\text{BLEU} + \text{ChrF++})/2$ for each direction, excluding COMET.

Human Evaluation Framework: Automatic metrics alone are insufficient for evaluating proverb translation. To capture the nuances missed by automated metrics, we employed expert human annota-

Model	Rows	Avg	Win (%)	Loss (%)	Tie (%)	EN-L	SW-L	EN-F	SW-F
gemma3-4b-it	1786	2.23	50.5%	49.5%	46.7%	2.14	2.26	2.62	1.94
llama-3.2-3b	1780	2.01	41.9%	58.1%	46.9%	2.16	2.05	2.17	1.51
minmax	1775	**2.42**	**68.0%**	32.0%	**37.2%**	**2.34**	2.26	**2.78**	**2.46**
qwen3-4b-it	1768	2.15	43.5%	56.5%	47.2%	2.05	**2.41**	2.44	1.58
smollm3-3b	1771	2.08	42.2%	57.8%	50.8%	2.06	2.30	2.30	1.51

Table 4: Mean scores of **human evaluation** overall (0-shot) performance across models, with Likert-scale ratings.

Model	n	Accuracy	Adequacy	Fluency	Cultural	Avg
EN-Literal						
minmax	563	**2.13**	**2.15**	**2.68**	**2.39**	**2.34**
qwen3-4b-it	555	1.98	2.01	2.14	2.08	2.05
smollm3-3b	562	1.91	1.93	2.26	2.14	2.06
gemma3-4b-it	557	1.89	1.92	2.54	2.20	2.14
llama-3.2-3b	565	1.86	1.87	2.63	2.27	2.16
SW-Literal						
qwen3-4b-it	552	**2.38**	**2.34**	2.44	**2.47**	**2.41**
smollm3-3b	550	2.25	2.30	2.35	2.31	2.30
gemma3-4b-it	553	2.19	2.20	2.33	2.31	2.26
minmax	550	2.11	2.13	**2.57**	2.24	2.26
llama-3.2-3b	551	1.99	1.98	2.15	2.10	2.05
EN-Figurative						
minmax	337	**2.47**	**2.46**	**3.43**	2.76	**2.78**
gemma3-4b-it	350	2.13	2.12	3.40	**2.81**	2.62
qwen3-4b-it	340	1.96	1.96	3.31	2.51	2.44
smollm3-3b	337	1.88	1.83	2.99	2.51	2.30
llama-3.2-3b	344	1.87	1.87	2.68	2.25	2.17
SW-Figurative						
minmax	325	**2.13**	**2.15**	**3.16**	**2.41**	**2.46**
gemma3-4b-it	326	1.68	1.60	2.48	1.99	1.94
qwen3-4b-it	321	1.49	1.44	1.87	1.52	1.58
smollm3-3b	322	1.46	1.41	1.69	1.49	1.51
llama-3.2-3b	320	1.43	1.41	1.73	1.48	1.51

Table 5: **Human evaluation analysis across tasks on Likert ratings averages for four dimensions.** No.of evaluated instances (n) vary due to annotator availability.

tors utilizing four-axis evaluation rubric:(i) **Meaning Accuracy** measures whether the generated otput correctly conveys the intended interpretation of the source proverb. (ii) **Adequacy** assesses whether the semantic content of the proverb is preserved. (iii) **Fluency** evaluates grammatical correctness and readability in the target language, while (iv) **Cultural Naturalness(Fidelity)** measures whether the translation sounds culturally appropriate as a proverb or proverb interpretation. Annotators assign five-point Likert scores (1 (poor), 5 (excellent)) for each dimension and performed pairwise comparisons between two model outputs to determine which translation better captures the proverb meaning (see Fig 8). Each annotator evaluated 50 proverbs per task, about 200 total across all four tasks. Language coverage was limited by native speakers availability (See Fig 11 & Fig 12).

5 Experimental Setup

Our evaluation pipeline is designed for consistent and scalable comparison across both locally deployed and API-based models. All models are evaluated under a unified setup with standardized prompting, decoding, and evaluation protocols. Experiments are executed using a combination of local GPUs and cloud infrastructure, while API-based systems are evaluated asynchronously for efficient large-scale inference. Outputs are generated independently for each task. organized by model, task, and language.

Model	Rows	Accuracy	Adequacy	Fluency	Culture	Avg
gemma3-4b-it	1786	1.99	1.99	2.63	2.31	2.23
llama-3.2-3b	1780	1.83	1.82	2.33	2.07	2.01
minmax	1775	**2.19**	**2.20**	**2.87**	**2.42**	**2.42**
qwen3-4b-it	1768	2.01	2.00	2.41	2.18	2.15
smollm3-3b	1771	1.93	1.93	2.32	2.14	2.08

Table 6: **Human evaluation analysis per model**.Mean average of four dimensions per instance shown by row.

5.1 Model Used

We used both open weight and proprietary models. Open models include Gemma-3-4B-It (Team, 2025a), Qwen-3-4B-It (Team, 2025b), Llama-3.2-3B-Instruct (Grattafiori et al., 2024) and smollm3-3b (Bakouch et al., 2025). and proprietary models accessed via API[3], including MiniMax-01 (Li et al., 2025), GLM-5 (Zeng et al., 2026), Kimi-K2.5 (Team et al., 2026), and GPT-5 mini (Singh et al., 2025), grok-4.1-fast models [4],and GPT-5.4 (Singh et al., 2025) as a judge model.

5.2 Prompting Strategy

We employ a unified prompting framework for consistent evaluation across tasks and languages. All inputs follow a shared chat-style template, with LMT and FMI differing only in task-specific instructions. Our design isolates literal translation from figurative interpretation (Appendix K).

Zero-shot. We evaluate models generalization in a zero-shot setting (no in-context example). For each configuration, models are assigned the role of a professional translator and prompted to translate from the source to the target language. Evaluation is conducted separately for each task configuration.

Few-Shot. We use a 3-shot setting, prompts include three in-context examples with the dataset annotation guidelines. Examples demonstrate expected output and task constraints. Evaluation is conducted separately for each task configuration.

5.3 Inference Setup

Models are evaluated under a standardized inference setup using task-specific prompts that distinguish LMT from FMI generation. Local models are evaluated using Hugging Face Accelerate with two parallel processes and bfloat16 mixed precision on

[3] https://openrouter.ai/
[4] https://openrouter.ai/x-ai/grok-4.1-fast

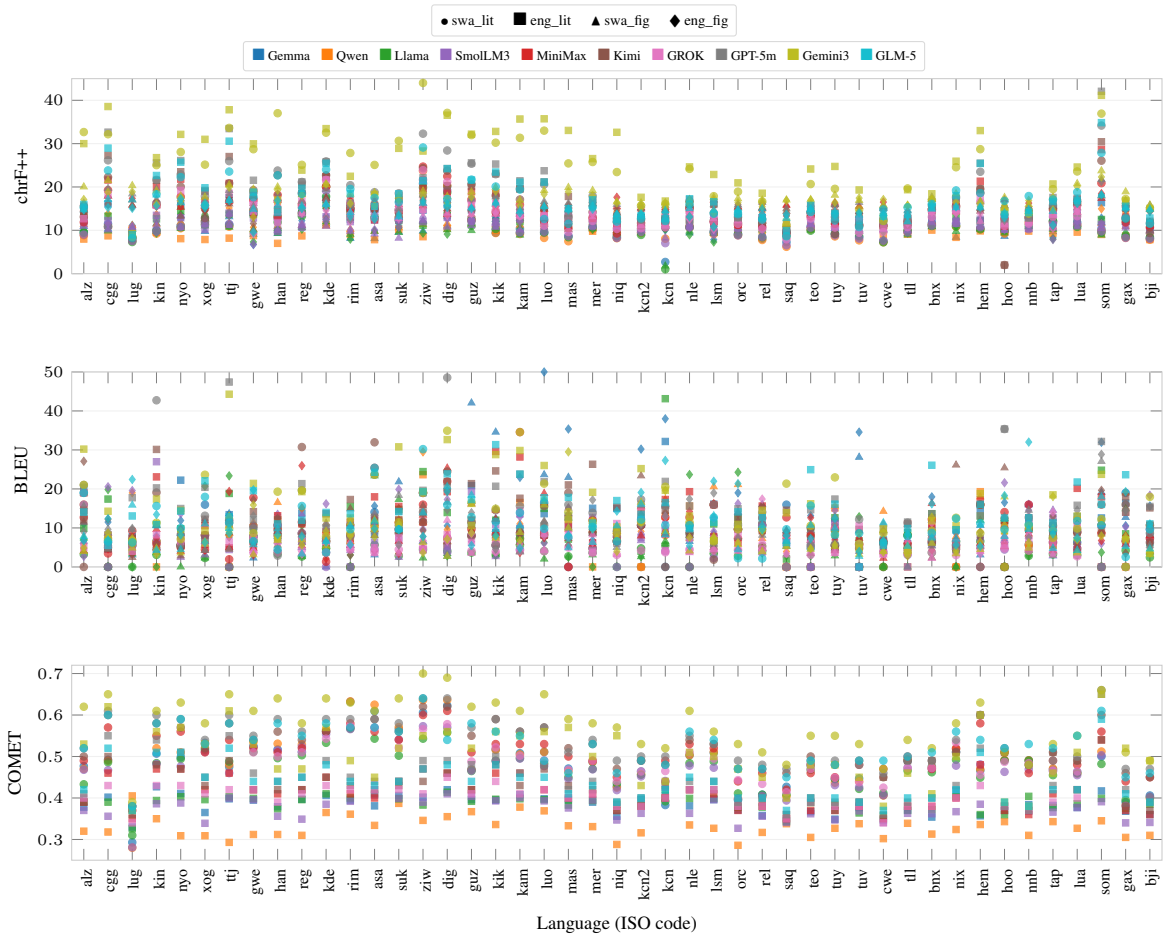

Figure 4: Performance (0-shot) for all 10 models across four tasks. Colors denote models and marker shapes indicate task types. COMET is for literal tasks only. GROK-4.1-fast has partial language coverage.

A100 GPUs. We use deterministic decoding for local models (greedy decoding; temp = 0) to ensure reproducible outputs. API-based models are evaluated through the OpenRouter interface using default decoding settings, which may introduce minor variability due to limited parameter control. Local experiments use a standardized Modal pipeline for scalable evaluation. No model-specific prompt templates are applied, and all models use identical input ordering and dataset splits. API requests are executed asynchronously with controlled concurrency (`semaphore = 8`) and a fixed 1.0-second delay between requests. Intermediate outputs are periodically checkpointed for fault tolerance.

6 Results and Discussion

6.1 Figurative & Cultural Understanding Remains Challenging

Figures 4 & 8 show that literal translation consistently outperforms figurative interpretation across nearly all models, languages, and evaluation metrics. Figurative tasks also exhibit substantially greater variance and frequent low-score outliers, particularly under BLEU, indicating that current MLLMs are considerably stronger at surface lexical mapping than culturally grounded semantic reasoning. While models often produce fluent literal translations, they struggle to preserve cultural abstraction, implicit meaning, and proverb-specific intent. African proverbs frequently encode moral lessons through metaphor and symbolism, requiring models to infer meaning beyond direct lexical translation. **Human evaluation** further reveals a strong disconnect between fluency and semantic accuracy. Many outputs appear grammatically natural and proverb-like while failing to preserve the intended figurative meaning. For example, for the Sukuma proverb "Chala chimwenga tachibanda tsaha" ("One finger does not crush a louse"), one model generates the fluent but semantically unrelated sentence "The monkey eats the fruit from the tree that is leaning on the ground." **Error analysis** reveals recurring failure patterns across mod-

els, including hallucination, semantic drift, source copying, truncated generations, and culturally unrelated moral abstractions. Models frequently replace culturally grounded meanings with generic motivational statements that sound plausible in English while failing to preserve the intended semantic and cultural content. Suggesting, MLLMs often rely on shallow statistical associations and surface fluency rather than grounded figurative understanding.

6.2 Performance Variation Across Languages, Models, Prompting, and Tasks

The `swa_lit` task consistently achieves the highest COMET scores across languages as illustrated in Figure 4 and 8. However, this trend is less consistent for chrF++ or BLEU, where English-target directions generally perform better. This divergence likely reflects two factors: (i) the typological proximity between Swahili and many predominantly Bantu source language, which benefits COMET's embedding-based evaluation; and (ii) the stronger representation of English in pretraining corpora, which favors surface-level lexical overlap metrics. Rather than functioning as a universal pivot language, Swahili appears to facilitate semantic transfer for linguistically related African languages. Figure 3 further shows substantial variation across languages, models, metrics, and prompting settings. While few-shot prompting occasionally improves performance, gains are generally vary across languages and tasks. Some languages such as *gik* and *mer* benefit more consistently, whereas others including *som* and *tur* show limited improvement. In several cases, zero-shot performance remains comparable to or stronger than few-shot performance (See Fig 2 & Fig 3), particularly for literal translation tasks. Figurative tasks remain substantially more difficult under both prompting settings. Across MLLMs, proprietary MLLMs consistently outperform open-weight MLLMs in both automatic and human evaluation. Gemini 3 Flash achieves the strongest overall performance, followed by GLM-5, GPT-5 mini, Kimi K2.5, and MiniMax-01. Open-weight MLLMs including **Gemma, Qwen, Llama, and SmolLM3** generally obtain lower scores overall, although relative performance varies across languages. Table 4 show similar trends: MiniMax achieves the highest average score (2.68) and strongest win rate (62.1%), while SmolLM3 records the lowest average score (2.36) and highest tie rate (45.7%) in human evaluation.

6.3 Automatic Metrics and LLM-based Evaluation

Figures 4 & 7 show clear differences in how evaluation metrics respond to figurative tasks and linguistic variability. BLEU exhibits substantially higher variance than chrF++ and COMET, often collapsing on figurative tasks due to its reliance on exact n-gram overlap, which penalizes valid paraphrases and alternative interpretations. In contrast, chrF++ and COMET remain more stable across languages and prompting settings. We report COMET only for literal translation tasks, since figurative interpretation often admits multiple valid meanings, limiting the reliability of single-reference evaluation. Human evaluation further highlights limitations of automatic metrics. Several fluent outputs receiving favorable automatic scores are still judged semantically incorrect or culturally unnatural. Furthermore, our LLM-as-a-judge evaluation (Section E) also shows stronger semantic agreement than lexical metrics, but still rewards plausible hallucinations in some cases. This corroborates our findings regarding the limitations of current LLMs which are biased towards structurally fluent outputs, but lacks grounding within the analysed cultural context.

7 Conclusion

We introduce AFRI-PROVERB, the first proverb benchmark for evaluating literal translation (LMT) and figurative interpretation (FMI) of East and Central African languages. The dataset is from culturally grounded sources, manually curated by native speakers and provides aligned proverb pairs in two target languages: English & Swahili, evaluated across 10 models. Our results reveal limited overall performance across languages, with even the strongest models achieving only moderate human evaluation scores. Human evaluation further exposes a clear gap between fluency and true semantic understanding, as models frequently generate fluent but culturally or semantically incorrect interpretations. We additionally observe strong language-dependent variation, consistent performance gaps between proprietary and open-weight models, and only marginal improvements from few-shot prompting. Standard automatic metrics such as BLEU also prove unreliable for culturally grounded figurative evaluation. Our findings underscores the need for culturally grounded resources, more reliable evaluation methods, and MLLMs capable of deeper semantic and cultural reasoning.

Limitations

This work has several limitations:

- First, our evaluation focuses primarily on zero-shot and few-shot prompting, which highlights the inherent multilingual capabilities of current LLMs, but may underestimate performance achievable through task-specific fine-tuning or culturally grounded adaptations. Future work can explore regional fine-tuning or targeted cultural adaptation tasks.
- Second, the benchmark covers a limited set of languages, due to the resource constraints and dependence on native speaker annotation. Expanding coverage to additional languages and cultures remains important for future work
- Third, many proverbs reflect historically situated social norms such as gender roles, authority, family structure. Model outputs should therefore be interpreted as reflections of cultural familiarity rather than endorsements of such views.
- Finally, the dataset remains relatively small due to limited proverb availability and the high cost of expert human annotation. Data expansion or proverb collections and evaluation is critical.

Ethical Statement

Cultural Bias and Representation: African proverbs encode culturally situated beliefs, values, and social norms, including historically specific views related to gender, authority, and family structure. Consequently, model outputs may reproduce biases present due to under-representation of pre-training data and embedded in proverbs themselves. To mitigate this, we rely on native speakers familiar with the relevant cultural context to carefully review data to reduce misrepresentation and reinforcement of harmful stereotypes.

Data Collection and Annotation Process: AFRIPROVERB was curated from publicly available datasets such as the African Proverbs Working Group (APWG), books, ebooks, oral traditions, and publicly available proverb collections with support from native speakers and Masakhane community annotators familiar with the respective cultural contexts. All translations and figurative interpretations were manually verified for linguistic accuracy, semantic consistency, and cultural fidelity. Harmful or sensitive content was carefully reviewed during annotation, and the dataset contains no personal or identifiable information.

Impact and use of research: This benchmark is designed to advance research on culturally grounded evaluation for multilingual LLMs, with a particular focus on low-resource African languages. The dataset and associated artifacts will be publicly released and made freely available to the research community under the Creative Commons Attribution (CC BY 4.0). By explicitly separating literal translation from figurative interpretation, the benchmark exposes limitations of current models and evaluation methods while promoting the development of more culturally aware NLP systems.

Use of AI Assistants

We use Claude Code (Anthropic) and ChatGPT for code debugging, editing assistance, and brainstorming visualization or diagram ideas. All scientific decisions, experimental design, analysis, interpretation of results, and final manuscript content were reviewed, validated, and approved by the authors

References

David Ifeoluwa Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, and 1 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070.

Oluwadamilare Adisa. 2024. Speaking of animals: A conceptual analysis of animal metaphors in yoruba. *Yoruba studies review*.

Franco Aixelá. 1996. Culture-specific items in translation. *(No Title)*, page 52.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, and 1 others. 2025. Proverbeval: Exploring llm evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6250–6266.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3.

Margaret Nasambu Barasa and Isaac Nilson Opande. 2017. Proverbs in lubukusu and ekegusii in kenya: Empowering or disempowering for women and girls?

Matthieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: multiword expression processing: a survey. *Computational Linguistics*, 43(4):837–892.

Naome Etori and Maria Gini. 2024. Rideke: Leveraging low-resource twitter user-generated content for sentiment and emotion detection on code-switched rhs dataset. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 234–249.

Dahklaoui Faycel. 2012. Food metaphors in tunisian arabic proverbs. *Working Papers in Linguistics*, 3.

Sayan Ghosh and Shashank Srivastava. 2022. epic: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

John M Kobia. 2016. A conceptual metaphorical analysis of swahili proverbs with reference to chicken metaphor. *International Journal of Education and Research*, 4(2):217–228.

Katerina Korre and John Pavlopoulos. 2025. Are proverbs the new pythian oracles? exploring sentiment in greek sayings. *arXiv preprint arXiv:2510.13341*.

George Lakoff and Marl Turner. 2024. *More than cool reason: A field guide to poetic metaphor*. University of Chicago press.

Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, and 1 others. 2025. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37:84799–84838.

Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2023. On the cultural gap in text-to-image generation. *arXiv preprint arXiv:2307.02971*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039.

Irene López Rodríguez. 2009. Of women, bitches, chickens, and vixens: Animal metaphors for women in english and spanish.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. Jawaher: A multidialectal dataset of arabic proverbs for llm benchmarking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12320–12341.

SITUMA PhD Martin. 2024. A pragmatic analysis of the significance of turkana proverbs. *International Journal of Research and Innovation in Social Science*, 8(1):2584–2593.

Charles Nimo, Tobi Olatunji, Abraham Toluwase Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Ezinwanne C Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, and 1 others. 2025. Afrimed-qa: a pan-african, multi-specialty, medical question-answering benchmark dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1973.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. Prometheus: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (Lrec'16)*, pages 3787–3793.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Shahrooz Pourhossein. 2016. Animal metaphors in persian and turkish roverbs: a cognitive linguistic study.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing, pages 2685–2702.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Huichan Seo, Sieun Choi, Minki Hong, Yi Zhou, Junseo Kim, Lukman Ismaila, Naome Etori, Mehul Agarwal, Zhixuan Liu, Jihie Kim, and 1 others. 2025. Exposing blindspots: Cultural bias evaluation in generative image models. *arXiv preprint arXiv:2510.20042*.

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Fancy C Soi. 2014. *A Stylistic and pragmatic study of Kipsigis proverbs*. Ph.D. thesis, University of Nairobi.

Gemma Team. 2025a. Gemma 3.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 307 others. 2026. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

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

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Surabhi Adhikari, Imran Razzak, and Usman Naseem. 2025. Probing the limits of multilingual language understanding: Low-resource language proverbs as llm benchmark for ai wisdom. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 120–129.

Lawrence Venuti. 2017. *The translator's invisibility: A history of translation*. Routledge.

Emilio Villa-Cueva, Sholpan Bolatzhanova, Diana Turmakhan, Kareem Elzeky, Henok Biadglign Ademtew, Alham Fikri Aji, Israel Abebe Azime, Jinheon Baek, Frederico Belcavello, Fermin Cristobal, and 1 others. 2025. Cammt: Benchmarking culturally aware multimodal machine translation. *arXiv preprint arXiv:2505.24456*.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, and 1 others. 2024. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023.

Minghan Wang, Viet Thanh Pham, Farhad Moghimifar, and Thuy Vu. 2025. Proverbs run in pairs: evaluating proverb translation capability of large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1646–1662.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chengxing Xie, Cunxiang Wang, and 1 others. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

A Literature Review

A.1 Metaphorical nature of Proverbs

Proverbs are inherently metaphorical, interpretable through the Great Chain Metaphor Theory framework (Lakoff and Turner, 2024), making them a gold standard for metaphor identification (Özbal et al., 2016). A multilingual proverb collection annotated for metaphors is essential for studying cross-cultural differences (Özbal et al., 2016). Numerous studies examine how proverbs encode culturally grounded meanings: stylistic analyzes of Kipsigis proverbs (Soi, 2014), chicken metaphors reflecting socio-cultural values in Swahili (Kobia, 2016), gender representations in Bukusu and Gusii (Barasa and Opande, 2017), animal metaphors mapping to human behavior in Yoruba (Adisa, 2024), English and Spanish (López Rodríguez, 2009), Persian and Turkish (Pourhossein, 2016), and food metaphors encoding human traits in Tunisian Arabic (Faycel, 2012). Collectively, these demonstrate that metaphors contribute substantially to the figurative meaning of proverbs (Faycel, 2012), posing a compelling challenge for machine translation, which must address the identification of metaphors given their language and culture-specific nature (Shutova, 2015; Özbal et al., 2016; Wang et al., 2025).

A.2 Cultural Language in MT

In translation studies,Culturally-Specific Items (CSIs), (Aixelá, 1996) refer to words or concepts that lack direct equivalents or carry different connotations in the target culture. These often arise when cultural references embedded in the source language do not directly exist or are understood differently in the target (Villa-Cueva et al., 2025). A common translation strategy is conforming the source text to the target culture by substituting unknown elements with familiar while this approach can improve surface-level comprehensibility, yet it simultaneously erases traces of the original culture(Venuti, 2017). This dilemma is inherited by modern text-only MT systems(Villa-Cueva et al., 2025). Recent work in multimodal MT and culturally aware NLP (Romero et al., 2024; Villa-Cueva et al., 2025; Seo et al., 2025) has incorporated the use of images as a source of supplementary context,allowing images to function as cultural proxy for knowledge and disambiguation. In line with (Wang et al., 2025; Ghosh and Srivastava, 2022; Thapa et al., 2025), we view proverbs as an alternative and complementary form of cultural grounding that is inherently textual. Rather than avoiding CSIs through normalization, we foreground them, using proverbs to systematically assess how machines preserve, distort, or fail to represent cultural meaning in translation.

B Impact of Few-shot Prompting

Minimal Impact of Few-shot Prompting. Comparing zero-shot and 3-shot results for the four small open-source models reveals that in-context examples provide negligible improvement across all three metrics and 45 languages (Figures 3, **??**). On chrF++, the average differences are marginal: Gemma shifts from 11.64 to 11.23 (zero-shot to 3-shot averaged across tasks), Qwen from 11.54 to 10.55, Llama from 10.46 to 9.56, and SmolLM3 from 10.81 to 10.17. in all four cases, 3-shot performance is slightly lower than zero-shot rather than higher. BLEU and COMET scores show similarly flat or slightly degraded patterns. This counterintuitive finding that few-shot examples fail to help and may marginally hurt suggests several explanations. First, the 3–4B parameter models likely lack sufficient capacity to effectively leverage in-context examples for languages absent from their pretraining data; the examples cannot teach linguistic knowledge that the model fundamentally lacks. Second, proverb translation requires deep cultural and metaphorical understanding that cannot be induced from three examples alone. Third, the in-context examples may actually constrain model outputs by encouraging overly literal copying patterns rather than flexible generation. Few-shot prompting typically improves translation quality, highlighting that for truly low-resource languages.

C Human Evaluation Analysis

Human evaluation analyzes five language models (*MinMax*, *Qwen3-4B-It*, *SmolLM3-3B*, *Gemma3-4B-It*, and *Llama-3.2-3B*) across 16 African languages using four evaluation dimensions: Meaning Accuracy, Adequacy, Fluency, and Cultural Naturalness. The study covers 80 language–model observations and approximately 3,100 evaluated samples, with a mean overall score of 1.93 ± 0.63 (range: 1.01–3.64). Results indicate that language-specific characteristics are the dominant source of variation, accounting for 69.5% of the observed variance, while model architecture contributes substantially less (6.1%). This suggests that language

properties play a larger role than model choice in determining performance (Figure 15). Quality metrics reveal a pronounced semantic–fluency asymmetry (Figure 16). Fluency achieves the highest score (2.28 ± 0.73), substantially exceeding Meaning Accuracy (1.73 ± 0.69) and Adequacy (1.72 ± 0.68), producing a gap of 0.55 points (31.5%). This pattern reflects an ***illusion of fluency***, where models generate grammatically natural and proverb-like outputs that fail to preserve intended meaning. Inter-metric analysis (Figure 5) further reveals strong relationships among semantic dimensions. Meaning Accuracy and Adequacy exhibit an almost perfect correlation ($r = 0.996$), indicating substantial overlap between semantic preservation measures, while Meaning Accuracy also correlates strongly with Cultural Naturalness ($r = 0.87$). In contrast, Fluency shows weaker correlation with semantic metrics ($r \approx 0.62$), suggesting partial decoupling between linguistic fluency and semantic fidelity. This pattern indicates that current models can produce fluent outputs without consistently preserving underlying meaning.

Table 12 reveals several consistent patterns across languages and models. First, performance differences are driven primarily by **meaning accuracy and adequacy** rather than **fluency**. For example, in Sukuma, all models achieve very high fluency scores **(> 4.3)**, yet meaning accuracy remains low **(1.17–2.38)**, indicating that models generate grammatically fluent translations but often fail to capture the intended figurative meaning of the source proverbs. Somali also, shows relatively high fluency **(3.54)** and low meaning accuracy **(1.16)**. Second, cultural naturalness closely tracks meaning accuracy, shows that most In particular, models often produce generic moral statements that resemble common English proverbs such, given the proverb *"Cha mfupi chiriwa ni mure"* **(literally, "The short person's thing is eaten by the tall person")**, one model generates the fluent but unrelated statement "The wolf devours its own tail," replacing the original meaning with a generic proverb-like expression. A similar failure for the proverb *"Chala chimwenga tachibanda tsaha"* **(literally, "One finger does not crush a louse,")**. One model produces a fluent but incorrect sentence, "The monkey eats the fruit from the tree that is leaning on the ground," as shown in Figure 12. In other cases, models produce prompt artifacts or incomplete outputs, suggesting instruction-following failures during generation. It is evident that current MLLMs

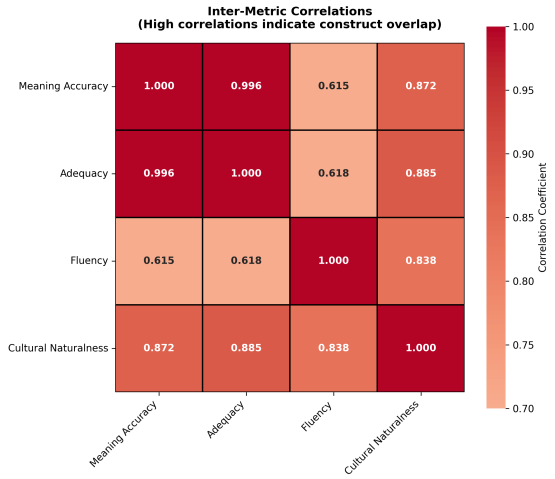

Figure 5: Inter-metric correlations across human evaluation dimensions. Meaning Accuracy and Adequacy exhibit near-perfect correlation, Fluency shows weaker association with semantic measures, supporting observed semantic–fluency gap.

possess limited grounded understanding of culturally specific figurative language. When faced with unfamiliar African proverbs, models often rely on memorized patterns from high-resource languages, producing generic moral statements or proverb-like templates instead of recovering the intended meaning.

C.1 Model Performance Rankings

Figure 14 indicates; although differences across models are modest, a consistent ranking emerges in mean performance: MinMax achieves the highest score (2.20 ± 0.60), followed by Gemma3-4B-It (1.96 ± 0.55), Qwen3-4B-It (1.90 ± 0.74), SmolLM3-3B (1.81 ± 0.65), and Llama-3.2-3B (1.76 ± 0.56). MinMax also exhibits the lowest coefficient of variation (CV = 27.4%), indicating greater consistency, whereas Qwen3-4B-It shows the highest variability (CV = 39.1%). However, the 0.45-point gap between the highest- and lowest-performing models is substantially smaller than the 3.34-point spread across languages (Ekegusii: 3.43 vs. Borana: 1.09), underscores language-level dominate over model-level(see Figure 17).

C.2 Key Findings and Implications

Finding 1: Illusion of fluency. Models consistently achieve higher Fluency scores (2.28 ± 0.73) than Meaning Accuracy scores (1.73 ± 0.69), producing a 31.5% gap between linguistic naturalness and semantic preservation. Models therefore often generate fluent and proverb-like outputs while

failing to preserve intended figurative meaning.

Finding 2: Implications for multilingual systems. Our results indicate that progress in multilingual systems may depend less on generic architectural improvements and more on language-specific adaptation, targeted fine-tuning, and mechanisms that explicitly preserve semantic meaning.

Recruitment and Payment Participants were recruited through an open volunteer call shared through community social media platforms and Slack channels. Participation was voluntary, and individuals self-selected by contacting the research team after reviewing the study description and requirements. To compensate participants for their time and effort, gift cards valued between $20 and $50 were provided, with compensation varying according to task scope and expected workload

D Qualitative Error analysis

We analyze 2,770 human-annotated outputs to characterize the failure modes of five MLLMs. Outputs are categorized in nine error categories (Table 44); categories are not mutually exclusive, so percentages do not sum to 100%. Only 13 outputs (0.5%) received a "none" label, indicating that near-total failure is consistent across all models and conditions. As shown in Table 44, hallucination (i) is the dominant error type, occurring in 1,270 outputs (45.8%), followed by (ii) source copying (879; 31.7%) and (iii) wrong target language (842; 30.4%). These three high-frequency errors together account for the majority of translation failures and reflect qualitatively distinct breakdown modes: **hallucination** indicates the model generates fluent but semantically unrelated content; **source copying** indicates a failure to engage with the translation task at all; and **wrong target language** indicates a failure of instruction following at the output level. Lower-frequency errors reveal additional failure modes. **Truncated or incomplete outputs** (382; 13.8%) indicate generation instability. **Semantic distortion/drift** (256; 9.2%), captures cases where outputs appear structurally plausible yet alter, weaken, or reverse the intended proverb meaning. **Meta/instruction leakage** (87; 3.1%) occurs when models describe the task instead of performing it (e.g., "The figurative meaning of this proverb is..."). Less frequent errors include **incoherent or nonsensical outputs** (69; 2.5%), **proverb substitution** (46; 1.7%), and **code-switching** (9; 0.3%). Figure 18 show error type by model. Table 40 summarizes common failure modes examples. Such as **(i) Source Copying (Input Echo:)** For example, Digo proverb *"Akiba si mbaya"*, some models output *"A kibba si mbaya"*, which remains in the source language. **(ii) Hallucinating:** For instance, "The short person's thing is eaten by the tall person") is translated as *"A wolf eats its own tail."* While grammatically correct and stylistically plausible, the output is semantically disconnected from the source meaning. Models rely on memorized or high-probability proverb templates rather than grounded translation. **(iii) Prompt Artifacts (Meta) and Instruction Leakage:** such as *"The literal translation of the proverb in English is..."*. These artifacts indicate incomplete instruction following and decoding instability, where the model fails to separate task description from generation.

E LLM-as-a-Judge Evaluation

Given the observed limitations of standard automated metrics, we investigated whether advanced LLMs could serve as reliable judges for culturally embedded translations alongside human evaluators. We utilised the zero-shot outputs from the five models evaluated in our human study (Gemma-3-4B-IT, Llama-3.2-3B, MiniMax-01, Qwen-3-4B-IT, and SmolLM3-3B). We selected GPT-5.4 as our judge model, motivated by its large parameter scale and superior baseline performance. The LLM judge was provided with the exact same four rubrics as the human evaluators (Meaning Accuracy, Adequacy, Fluency, and Cultural Naturalness) and instructed to rate each model response on an ordinal 1–5 scale. To examine the alignment between the LLM judge and human native speakers, we paired their respective rubric scores. Because the evaluation relies on ordinal Likert-scale rankings, we report the Spearman's rank correlation coefficient ρ to measure the relationship between the human and LLM scores.

The results 7 demonstrate a severe misalignment between human and LLM judgments. The overall Spearman correlation across all tasks and rubrics was exceptionally weak (ρ=0.224,p<0.001), indicating only a marginal relationship with human native speakers' scores. Breaking this down by specific metrics reveals a distinct "fluency bias" where the LLM aligned best with humans on grammatical Fluency (ρ=0.290,p<0.001), but struggled

Task	Meaning	Adequacy	Fluency	Cultural	Overall
All Tasks	**0.161**	**0.165**	**0.290**	**0.211**	**0.224**
ENG-FIGURATIVE	0.239	0.244	0.285	0.207	0.255
ENG-LITERAL	0.003†	0.006†	0.066	0.005†	-0.006†
SWA-FIGURATIVE	0.127	0.142	0.310	0.245	0.287
SWA-LITERAL	0.223	0.222	0.215	0.217	0.207

Table 7: Spearman's rank correlation (ρ) between human native speaker and LLM judge (GPT-5.4) scores, broken down by evaluation rubric and translation task. All reported correlations are statistically significant ($p < 0.001$) except those marked with † ($p > 0.05$).

significantly to evaluate deep semantic preservation, showing its weakest correlations on Meaning Accuracy (ρ=0.161,p<0.001) and Adequacy (ρ=0.165,p<0.001). Cultural Naturalness similarly showed weak alignment (ρ=0.211,p<0.001). This mirrors our generative findings that the models prioritise surface-level linguistic fluency over true semantic preservation, the LLM judge also relied heavily on surface fluency to score translations but fails to reliably assess culturally grounded meaning. Table 7 presents these correlation results further broken down by language task and rubric.

Model and Budget Since GPT-5.4 is proprietary, model size information is not publicly available. The total API cost for LLM-based evaluation was approximately $480. And $100 for inference 0-shot for all other proprietary models used.

F Language Family, Digital Support and Vitality Status

Table 3 shows the distribution of 44 languages: Niger–Congo dominates (30; 65.9%), followed by Nilo–Saharan (8; 18.2%), Afro–Asiatic (5; 9.1%), and one Creole (2.3%). Most languages are Stable (56.8%) or Institutional (34.1%), with few Endangered (9.1%). Digital support is mainly Ascending (56.8%) and Emerging (34.1%), with only 6.8% Vital and none Thriving (Table 32). Despite strong inter-generational transmission, most languages remain under-resourced digitally.

G Languages Covered

1. **Nyala (nle):** Also known as The Abanyala, AvaNyala, AbaNyala, WaNyala or *Nyala* are a stable indigenous Kenyan language, from Niger-Congo family and part of the Oluluyia macrolanguage. They are one of the eighteen subgroups of the Luhya (Abaluyia) community, a major Bantu group in western Kenya located in Busia County along the Kenya–Uganda border With an estimated population of 10k–1M speakers. The language has Emerging digital support.

2. **Samburu (saq):** This is a stable indigenous Nilo-Saharan language spoken in northern Kenya by a pastoralist community related to the Maasai. They migrated from present-day South Sudan and settled north of Mount Kenya and south of Lake Turkana. Estimated 10k–1M speakers. It's digital support is Ascending. Although historically part of the Loikop community, they became known as the Samburu, a name widely believed to derive either from a term meaning "butterflies" or from the Maasai word samburr, referring to a traditional leather bag.

3. **Turkana (tuv):** Turkana is a stable indigenous Nilotic language of Kenya belonging to the Nilo-Saharan family. Spoken as a first language by the Turkana community in Turkana County in northwestern Kenya. The 2009 Kenyan census reports approx 1.02 M speakers (2.14% of the population), making Turkana the third-largest Nilotic group after Kalenjin and Luo. Digital language support for Turkana is currently ascending.

4. **Teso (teo):** Ateso is a stable indigenous language of Uganda and Kenya. It belongs to the Nilo-Saharan language family. They are part of the Plains Nilotic group and speak related languages within the Nilo-Saharan family including Turkana, Karamojong, Toposa, Maasai, and Samburu.

5. **Rendille (rel):** Rendille is a stable indigenous Kenyan language belonging to the Cushitic branch, spoken primarily in northern Kenya with historical origins in southern Ethiopia. It has an estimated 96,000 native speakers, and its digital support is currently ascending.

6. **Kamba (kam):** Kamba is a stable indigenous language of Kenya and Tanzania. It belongs to the Niger-Congo language family. The language is used as a first language by all in the ethnic community.Ethnologue estimates 4.6M Speakers, *Institutional* vitality and *Ascending* digital support.

7. **Digo (dig)**: Chidigo is a stable indigenous language of Kenya and Tanzania belonging to Niger-Congo family. Their territory extends inland from the Indian Ocean toward the Tanzania border. Chiduruma (language spoken), belongs to Niger–Congo variety. Ethnologue estimates 10k–1M speakers, reflecting limited documentation, and classifies the language as having *Institutional* vitality and *Ascending* digital support.

8. **Maasai (mas)**: Also known as Maa is a stable Nilotic language of the Nilo-Saharan language family spoken in Kenya and northern Tanzania, particularly in regions near the African Great Lakes. It is closely related to other Nilotic languages such as Dinka, Kalenjin, and Nuer. The language is spoken by an estimated 1.5 million native speakers, and digital language support is currently ascending.

9. **Nubi (kcn)**: Nubi is an endangered indigenous language of Uganda and Kenya. It is an Arabic-based creole. The language is used as a first language by all adults in the ethnic community, but not all young people. The language is spoken by an estimated 50K native speakers, and digital language support is ascending.

10. **Dholuo (luo)**: Dholuo is a stable indigenous language of Kenya and Tanzania. It belongs to the Nilo-Saharan language family. The language is used as a first language by all in the ethnic community. The language is spoken by an estimated 4.2M native speakers, and digital language support is ascending.

11. **Gikuyu (kik)**: Gikuyu is a stable indigenous language of Kenya. It belongs to the Niger-Congo language family. The language is used as a first language by all in the ethnic community. The language is spoken by an estimated 6.6M native speakers, and digital language support is ascending.

12. **Ekegusii (guz)**: Ekegusii is a stable indigenous language of Kenya and Tanzania. It belongs to the Niger-Congo language family. The language is used as a first language by all in the ethnic community. The language is spoken by an estimated L1:2.2M native speakers and L2:500K. Digital language support is ascending.

13. **Ganda (lug):** Also known as Luganda, Ganda is a Bantu language of Uganda spoken by the Baganda people. It has over 5.5 million native speakers, with millions more using it as a second language. Ganda belongs to the Niger–Congo language family and is used both as a first language within the ethnic community and as a language of instruction in education.

14. **Zigula (ziw)**: Also known as Chizigua or Kizigua, is a Bantu language spoken by the Zigua people mainly in northern and coastal Tanzania. The dilect is closely related to Mushunguli spoken by the Somali Bantu minority in southern Somalia. Zigula is an endangered and belongs to Niger-Congo language family.

15. **Hangaza (han)**: Also known as Kihangaza, is a Bantu language spoken by the Hangaza people, residing in the Kagera region of Tanzania. The language is stable with approximately 500,003 native speakers, and digital language support is currently emerging.It belongs to the Niger-Congo family.

16. **Kara (reg)**: Kara is an endangered indigenous Niger–Congo Bantu language spoken in northwestern Tanzania. It is used by the Kara people, primarily on Kara Island, a small islet within the Ukerewe Islands group in Lake Victoria, which forms part of the Mwanza Region. The language is spoken by an estimated 86,000 native speakers (based on available census estimates), and digital language support is currently emerging.

17. **Alur (alz)**: Alur is a stable Nilo-Saharan language spoken in the Democratic Republic of the Congo and northwestern Uganda. It is used by the Alur people. The language is spoken by an estimated 1.7 million native speakers, and digital language support is currently ascending.

18. **Borana (gax)**: Borana–Arsi–Guji Oromo is a stable Afro-Asiatic (Cushitic) language variety spoken across Ethiopia, Kenya, and southern Somalia. It is part of the Oromo macrolanguage and is spoken by approximately 3–6 million native speakers, with digital language support currently ascending.

19. **Hema (nix)**: Hema is a stable indigenous lan-

guage of the Democratic Republic of the Congo, belongs to the Niger–Congo language family. Estimated estimated 10K–1M native speakers according, and digital language support is currently emerging.

20. **Tshiluba (lua)**: Also known as Cilubà, Luba-Kasai, or Luba-Lulua—is a Bantu language of the Niger–Congo family spoken mainly in Kasai-Oriental and Kasai-Central in the DRC. It is one of the country's four national languages. Estimates indicate 15M native speakers, and digital language support is currently ascending.

21. **Orma (orc)**: Orma is a stable indigenous language of Kenya. It belongs to the Afro-Asiatic language family and is part of the Oromo macrolanguage. The language has an estimated 10K–1M native speakers, and digital language support is currently emerging.

22. **Holoholo (hoo)**: Holoholo is a stable indigenous language of the Democratic Republic of the Congo. They are also known as Horohoro or Kalunga ; Bantu ethnic group that live around Kalemie on the opposite shore of Lake Tnganyika. Estimated to have about 10K–1M native speakers, and digital language support is currently emerging.

23. **Burji (bji)**: Burji is an endangered indigenous language of Ethiopia. It belongs to the Afro-Asiatic language family. They are also known as Horohoro or Kalunga ; Bantu ethnic group that live around Kalemie on the opposite shore of Lake Tnganyika. Estimated to have about 10K–1M native speakers, and digital language support is currently emerging.

24. **Tetela (tll)**: Also called Otetela, Kitetela, or Sungu is a stable indigenous language of the Democratic Republic of the Congo. It belongs to the Niger-Congo language family. The Tetela or Batetela; Bantu language of northern Kasai-Oriental Province. Estimated to have about 1M–1B native speakers, and digital language support is currently ascending.

25. **Nyaturu (rim)**: Nyaturu, also known as Kinyaturu or Rimi (Kirimi), is a stable Niger–Congo Bantu language spoken in central Tanzania, primarily in the Singida Region. It is spoken by the Nyaturu (Wanyaturu), also referred to as the Arimi. The language has an estimated 1M-1B native speakers, and digital language support is currently ascending.

26. **Makonde (kde)**: Makonde is a stable Niger–Congo Bantu language spoken primarily in southeastern Tanzania and northern Mozambique. In Tanzania, speakers are concentrated in the Mtwara and Lindi Regions, while in Mozambique they are mainly found in Cabo Delgado Province, including the Mueda Plateau. Smaller Makonde-speaking communities are also present in Kenya, in Kwale, Msambweni Sub-county and Taita Taveta County near the Kenya–Tanzania border. The language, also known as ChiMakonde, is spoken by an estimated 1.3 million native speakers, and digital language support is currently ascending.

27. **Hemba (hem)**: Hemba, also known as Eastern Luba (Kiluba-Hemba), is a stable Niger–Congo Bantu language spoken primarily in the Democratic Republic of the Congo. It is used by the Hemba people, with closely related varieties understood in parts of northern Zambia and Tanzania. The language is spoken by an estimated 1.3 million native speakers, and digital language support is currently emerging.

28. **Kwele or Bakwele (bkw)**: Kwele, also known as Bakwele or Bekwel, is a stable indigenous Niger–Congo Bantu language spoken in parts of the Republic of the Congo, Cameroon, and Gabon. The language is spoken by an estimated 120,000 native speakers, and digital language support is currently emerging.

29. **Tooro or Rutooro(ttj)**: Tooro, also known as Rutooro (related to Runyoro), is a stable Niger–Congo Bantu language spoken in western Uganda. It is used by the Tooro (Abatooro) people, primarily in the Tooro Kingdom, including districts such as Kabarole, Kyenjojo, and Kyegegwa. The language is spoken by an estimated 500,000 native speakers, and digital language support is currently ascending.

30. **Soga (oxg)**: Soga, also known as Lusoga or Olusoga, is a stable Niger–Congo Bantu language

spoken in eastern Uganda. It is used by the Basoga people, primarily in the Busoga region, including districts such as Bugiri, Iganga, Jinja, Kamuli, and Mayuge. The language is spoken by an estimated 1 million native speakers, and digital language support is currently ascending.

31. **Rufumbira (kin)**: Rufumbira is a stable Niger–Congo Bantu language variety spoken in southwestern Uganda, primarily in Kisoro District. It is used by the Bafumbira people, who are culturally associated with the Bahutu, Batutsi, and Batwa groups. Rufumbira is commonly described as a regional variety or dialect of Kinyarwanda, which is also spoken across the Uganda–Rwanda border. The language is spoken by an estimated 1 million native speakers, and digital language support is currently vital.

32. **Pare(asa)**: Pare, also known as Kipare or Asu (with alternative names Chasu, Casu, Athu, and Chathu), is a stable Niger–Congo Bantu language of the Northeast Coast Bantu subgroup spoken in northern Tanzania. It is used by the Pare people. The language is spoken by an estimated 500,000 native speakers, and digital language support is currently ascending.

33. **Gweno (gwe)**: Gweno, also known as Kighonu, is an endangered (moribund) Niger–Congo Bantu language spoken in the North Pare Mountains of the Kilimanjaro Region in Tanzania. It is used primarily by older adults within the Gweno community, who are sometimes identified as part of the broader Asu/Chaga ethnolinguistic grouping. The language has not been transmitted to younger generations since the 1970s. Gweno is spoken by an estimated 2,500 native speakers, and digital language support is currently emerging.

34. **Bangubangu (bnx)**:Bangubangu is a stable indigenous Niger–Congo Bantu language spoken primarily in the eastern Democratic Republic of the Congo, especially in Kabambare Territory. Smaller Bangubangu-speaking communities are also present in neighboring Tanzania. The language is used by the Bangubangu people and is spoken by an estimated 250,000–700,000 native speakers, and digital language support is currently emerging.

35. **Olusamia (lsm)**:Olusamia, also known as Saamya-Gwe, is a stable Niger–Congo Bantu language spoken in eastern Uganda and western Kenya. It forms part of the Luhya (Oluluyia) macrolanguage and is used by the Samia (Abasamia) people. The language is spoken by an estimated 480,000 native speakers, and digital language support is currently emerging.

36. **Runyoro (nyo)**:Runyoro, also known as Orunyoro, nyoro, is a stable Niger–Congo Bantu language spoken in western Uganda, primarily in the Bunyoro Kingdom region. It is used by the Banyoro people. The language is spoken by an estimated 670,000–1.4 million native speakers, and digital language support is currently ascending.

37. **Chiga (cgg)**: Chiga,also known as Rukiga, Kiga, Ciga, Oluchiga, and Nkore-Kiga, is a stable Niger–Congo Bantu language spoken in southwestern Uganda. The language is spoken by an estimated 1.6–3.5 million native speakers, and digital language support is currently ascending.

38. **Taabwa (tap)**: Nande is a stable indigenous language of the Democratic Republic of the Congo. It belongs to the Niger-Congo language family. The language is spoken by an estimated 500,000 native speakers, and digital language support is currently emerging.

39. **Nande (nnb)**: Nande, also known as Kinande, Yira, Banande, is a stable Niger–Congo Bantu language spoken primarily in North Kivu Province of the Democratic Republic of the Congo, with additional speaker communities in neighboring Uganda. The language is spoken by an estimated 900,000–1.1 million native speakers, and digital language support is currently ascending.

40. **Nandi (niq)**: Nandi, also known as Naandi, is a stable Nilotic language of the Nilo-Saharan language family and forms part of the Kalenjin macrolanguage. It is spoken by the Nandi people in the Rift Valley region of western Kenya. The language is spoken by an estimated 950,000–1.3 million native speakers, and digital language support is currently ascending.

41. **Tachoni (lts)**: Tachoni, also known as Lutachoni, is a stable Niger–Congo Bantu language of Kenya and forms part of the Luhya (Oluluyia) macrolanguage. It is spoken primarily in western Kenya, including areas around Webuye, Ndivisi, Tongaren, Lugari, and Matete. The language is spoken by an estimated 94,000–118,000 native speakers, and digital language support is currently emerging.

42. **Sukuma (suk)**: is a stable indigenous Bantu language of Tanzania belonging to the Niger–Congo family. Spoken primarily in northwestern Tanzania, southeast of Lake Victoria, including Mwanza and Shinyanga and Lake Eyasi. It is the first language and used in everyday communication across sukuma generations, although not formally taught in schools. It has an estimated **8.1 million** native speakers. Digital language support for Sukuma is ascending.

43. **Somali (som)**: Somali is the official national language of Somalia. It belongs to the Afro-Asiatic language family. It is used as a language of instruction in education. The language is spoken by an estimated 40,000–217,000 native speakers, and digital language support is currently limited.

44. **meru (mer)**: Kimîîru is a stable indigenous language of Kenya. It belongs to the Niger-Congo language family. The language is used as a first language by all in the ethnic community. It is taught as a subject of instruction in education. The language is spoken by an estimated 2M native speakers, and digital language support is ascending.

Language	# Pairs	Source	Author	Year	Format
Luo	700	Archive.org	Asenath Bole Odaga	2005	ebooks
Turkana	100	African Proverbs Working Group	Gabriel Odhiambo	2015	ebooks
Teso	100	African Proverbs Working Group	Gabriel Odhiambo	2016	ebooks
Samburu	100	African Proverbs Working Group	Margaret Wambere Ireri	2016	ebooks
Rendille	100	African Proverbs Working Group	Benon Benjamin Chelo	2017	ebooks
Nubian	500	Archive.org	Maher Habbob	2022	ebooks
Kamba	2000	JSTOR	Jeremia M. Kitunda	2021	publications
Meru	1200	Book Collection	Edward Steven Mwiti	2004	book
Gusii	1000	Web Collection	–	–	webpages
Duruma	100	African Proverbs Working Group	Margaret Wambere Ireri	2017	ebooks
Digo	100	African Proverbs Working Group, BTL E. Africa	Margaret Wambere Ireri	2016	ebooks
Maasai	100	African Proverbs Working Group	David Kesier	2024	ebooks
Maasai	105	African Proverbs Working Group	Joseph Kariuki	2014	ebooks
Ganda_Kooki	100	African Proverbs Working Group	Atemu Ishiabwe	2019	ebooks
Zigula	100	African Proverbs Working Group	Joseph Kariuki	2019	ebooks
Hangaza	100	African Proverbs Working Group	Joseph Nkumbulwa	2011	ebooks
Swahili	143	Web Source	Hassan O. Ali	2004	website
Kara	100	African Proverbs Working Group	–	–	ebooks
Alur	100	African Proverbs Working Group	George Pirwoth Atido	2011	ebooks
Borana	100	African Proverbs Working Group	Angelica Chelo & Benon Chelo	2013	ebooks
Kikuyu	100	African Proverbs Working Group	Margaret Wambere Ireri	2014	ebooks
Hema	100	African Proverbs Working Group	Calvin Katabarwa	2010	ebooks
Mbeere	100	African Proverbs Working Group	Margaret Wambere Ireri	2014	ebooks
Tshiluba	100	African Proverbs Working Group	Tyty Mukendi	2011	ebooks
Orma	100	African Proverbs Working Group	Calvin Katabarwa & Angelique Chelo	2012	ebooks
Holoholo	100	African Proverbs Working Group	Elias Bushiri	2016	ebooks
Burji	100	African Proverbs Working Group	Angelique Chelo	2016	ebooks
Tetela	100	African Proverbs Working Group	Elysee Okubo	2016	ebooks
Nyala	100	African Proverbs Working Group	Kevin Okubo	2016	ebooks
Nyaturu	100	African Proverbs Working Group	Devotha Bishangi	2016	ebooks
Makonde	100	African Proverbs Working Group	Margaret Wambere Ireri	2017	ebooks
Bwisha	100	African Proverbs Working Group	Devotha Bishangi	2017	ebooks
Hemba	100	African Proverbs Working Group	Atemu Ishiabwe	2017	ebooks
Kwele	100	African Proverbs Working Group	Elias Bushiri	2017	ebooks
Tooro	100	African Proverbs Working Group	Kevin Okubo	2019	ebooks
Soga	100	African Proverbs Working Group	Adrien Socrate	2019	ebooks
Rufumbira	100	African Proverbs Working Group	Angela Chelo	2019	ebooks
Pare	100	African Proverbs Working Group	Annastasi Oisebe	2019	ebooks
Nuer	100	African Proverbs Working Group	Margaret Wambere Ireri	–	ebooks
Lopit	100	African Proverbs Working Group	Benon Chelo	2020	ebooks
Gweno	100	African Proverbs Working Group	Etoka Peter	2019	ebooks
Bangubangu	100	African Proverbs Working Group	Elias Bushiri	–	ebooks
Olusamia	100	African Proverbs Working Group	Gabriel Odhiambo	2017	ebooks
Runyoro	100	African Proverbs Working Group	Angela Katabarwa	2017	ebooks
Chiga	100	African Proverbs Working Group	Kevin Okubo	2016	ebooks
Taabwa	100	African Proverbs Working Group	Etoka Peter	2017	ebooks
Nande	100	African Proverbs Working Group	Adrien Socrate	2017	ebooks
Nandi	100	African Proverbs Working Group	Joseph Kariuki	2017	ebooks
Tachoni	100	African Proverbs Working Group	Annastasi Oisebe	2017	ebooks
Tugen	100	African Proverbs Working Group	Kevin Okubo	2017	ebooks
Gikuyu	1000	Scribd	–	–	webpage
Somali	1000	Book Collection	Cabdulqadir Nuur Xuseen	2015	ebooks

Table 8: Sources of African proverb datasets across languages.

H Data Sources

I Annotator guidelines for proverb

Annotator Instructions: Proverb Verification and Annotation

Objective. The objective of this annotation task is to verify, standardize, and enrich a multilingual African proverb dataset for research on figurative language, cultural reasoning, and machine translation. Annotators must ensure that each entry represents an authentic proverb, is correctly written in its conventional form, and is paired with accurate literal and figurative interpretations in both English and Swahili. Accuracy, consistency, and cultural fidelity are essential.

Step 1: Proverb Verification
- Read the original text carefully.
- Determine whether the text is genuinely a proverb.
- If the text is not a proverb such as slogan or ordinary sentence, flag it.
- If it is a proverb, verify spelling, grammar, and conventional usage; correct errors without rewriting.

Example (Correct): Mfalme hasifiwi kwao.
Example (Incorrect): Mfalme hasifiwi kwake sana (modified form)
Example (Flag): Leo ni siku nzuri ya kufanya kazi.

Step 2: Literal Translation (English & Swahili)
- Provide a literal, word-by-word translation.
- Do not include interpretation or explanation.
- Supply translations in both English and Swahili.

Correct: "Hurry hurry has no blessing."
Incorrect: "Doing things too fast leads to failure."

Step 3: Figurative Interpretation (English & Swahili)
- Explain the intended meaning or lesson of the proverb.
- Reflect culturally grounded usage.
- Provide interpretations in both English and Swahili.

Correct: Acting too quickly often leads to mistakes or poor outcomes.
Incorrect: "Hurry hurry has no blessing."

Step 4: Proverb Topic Labeling
- Some proverbs are provided with an existing **topic label** such as *Abuse, Precaution, gender norms*.
- If the topic label is appropriate, **do not change it**.
- If the topic label is present but **not the most appropriate**, replace it with a more suitable topic.
- If no topic label is provided, assign a topic that best captures the core theme of the proverb.
- Topic labels must be **one word, or at most two words**.

Example (Keep Topic): **Topic:** *Abuse*
Example (Change Topic): **Topic:** *Women/Men* $\rightarrow$ *gender norms*
Example (Add Topic): **Topic:** *Death*

Key Guideline. Literal translation describes what the proverb says; figurative interpretation explains what the proverb means. When uncertain, flag the entry rather than guessing.

Figure 6: Annotator guidelines for proverb verification, literal translation, and figurative interpretation.

J Annotator Template: Literal vs. Figurative Translations

The template below was provided to annotators for producing literal (imagery-preserving) and figurative (abstract) translations of proverbs. The example illustrates correct and incorrect literal renderings as well as proper abstract figurative paraphrases in English and Swahili.

Annotator Template: Literal vs. Figurative Translations Requirements.

A. Literal Translation (Required)

A literal translation MUST:

- Preserve the imagery and entities (animals, objects, actions).
- Avoid interpretation, explanation, or moralization.
- Sound odd or ambiguous on purpose.
- Be word- or phrase-aligned as much as possible.

Do NOT:

- Replace the proverb with an equivalent proverb.
- Explain the lesson.
- Use abstract language.

B. Figurative Translation (Required)

A figurative translation MUST:

- Explicitly state the underlying principle.
- Remove metaphor, imagery, and animals.
- Use plain, abstract language.
- Be culturally faithful, not proverb-matching.

Do NOT:

- Use another proverb.
- Retain imagery.
- Be poetic or idiomatic.

Worked Example

Source (ekegusii_prov): *Nyang'au tiboyera nyama ye'nyomba.*
Literal (`eng_literal`): *A dog does not eat the meat of the home.*
Literal (`swa_literal`): *Mbwa hali nyama ya nyumbani.*

Correct

Preserves *dog*, *meat*, and *home*; stays literal and aligned.

Incorrect

People protect their own.
(This is figurative/abstract, not literal.)

Figurative (`eng_figurative`: *Even those who behave badly tend to protect their own group.*
Figurative(`swa_figurative`): *Hata watu wasiotii hulinda walio wao.*

Correct

Explicit and abstract; removes imagery; states the underlying principle.

Incorrect

A dog never bites its owner.
(This is a different proverb, not an abstract paraphrase.)

K Prompting Strategy

K.1 Prompt 1: Zero Shot Prompting

LMT prompt (Literal Translation – Zero-Shot)

Task:
You are a professional translator specializing in proverbs. Your task is to translate the following proverb **literally** from {source_language} to {target_language}.
A literal translation MUST:
- Preserve the imagery and entities (people, objects, animals, actions).
- Avoid interpretation, explanation, or moralization.
- Keep the original structure as closely as possible.
- Sound unusual or ambiguous if necessary.

Do NOT:
- Replace the proverb with an equivalent proverb.
- Explain the meaning or lesson.
- Use abstract interpretation.

Instruction:
Respond only with the literal translation.
Do not add explanations or additional text.

Now translate the following proverb

Input ({source_language}):
{proverb}

Output ({target_language}):

FMT prompt (Figurative Meaning Translation – Zero-Shot)

Task:
You are a professional translator specializing in proverbs. Your task is to infer the figurative meaning of the following proverb from {source_language} to {target_language}.
A figurative translation MUST:
- Explicitly state the underlying principle.
- Remove metaphor, imagery, and references to animals or objects.
- Use plain, abstract language.
- Remain culturally faithful to the meaning of the proverb.

Do NOT:
- Use another proverb.
- Retain imagery or metaphor.
- Produce poetic or idiomatic wording.

Instruction:
Respond only with the figurative translation.
Do not add explanations or additional text.

Expected format

Input ({source_language}):
A proverb in {source_language}

Output ({target_language}):
The figurative meaning of the proverb in {target_language}.

Now translate the following proverb

Input ({source_language}):
{proverb}

Output ({target_language}):

K.2 Prompt 2: Few-Shot Prompting (3-shot)

LMT prompt (Literal Translation - Digo)

Task:
You are a professional translator specializing in proverbs. Translate the proverb from {source_language} to {target_language} literally.
A literal translation MUST:
- Preserve the imagery and entities (people, objects, actions).
- Avoid interpretation, explanation, or moralization.
- Keep the original structure as closely as possible.
- Sound unusual or ambiguous if necessary.

You should NOT:
- Replace the proverb with an equivalent proverb.
- Explain the meaning or lesson.
- Use abstract interpretation.

Instruction:
Respond only with the literal translation.
Do not add explanations or additional text.

Examples

Example 1
Input ({source_language}):
Achili ni nyere, chila mmwenga ana zakwake

Output ({target_language}):
Akili ni kama nywele; kila mtu ana zake.

Example 2
Input ({source_language}):
Achiyeshiba kamanya mndzala

Output ({target_language}):
Aliyeshiba hamjui mwenye njaa.

Example 3
Input ({source_language}):
Achiyetsupa mbere, siye myao

Output ({target_language}):
Anayekupita si mwenzako.

Now translate the following proverb

Input ({source_language}):
{proverb}

Output ({target_language}):

LMT prompt (Literal Translation - Luo)

Task:
You are a professional translator specializing in proverbs. Your task is to translate the following proverb **literally** from {source_language} to {target_language}.
A literal translation MUST:
- Preserve the imagery and entities (people, objects, animals, actions).
- Avoid interpretation, explanation, or moralization.
- Keep the original structure as closely as possible.
- Sound unusual or ambiguous if necessary.

Do NOT:
- Replace the proverb with an equivalent proverb.
- Explain the meaning or lesson.
- Use abstract interpretation.

Instruction:
Respond only with the literal translation.
Do not add explanations or additional text.

Examples

Example 1
Input ({source_language}):
Acham ayieng' ne onego suna.

Output ({target_language}):
I eat until I am satisfied killed the mosquito.
Output ({target_language}):
Kula hadi kushiba kulimuua mbu.

Example 2
Input ({source_language}):
Achang' chacha ngeta mopadore.

Output ({target_language}):
Achang' chacha the flat thin ngeta seed.
Output ({target_language}):
Achang' chacha ni mbegu ya ngeta iliyo nyembamba na tambarare.

Example 3
Input ({source_language}):
Achi ngure god kabondo.

Output ({target_language}):
The hip shaper Kabondo's hill.
Output ({target_language}):
Mlima wa Kabondo huunda viuno.

Now translate the following proverb

Input ({source_language}):
{proverb}

Output ({target_language}):

K.3 Prompt 2: Figurative Meaning Interpretation (3-shot)

FMT prompt (Figurative Meaning Translation -Digo)

Task:
You are a professional translator specializing in proverbs. Your task is to infer the figurative meaning of the following proverb from {source_language} to {target_language}.
A figurative translation MUST:
- Explicitly state the underlying principle or lesson.
- Remove metaphor, imagery, and references to animals or objects.
- Use clear, plain, abstract language.
- Remain culturally faithful to the meaning of the proverb.

Do NOT:
- Use another proverb.
- Retain metaphorical imagery.
- Produce poetic or idiomatic wording.

Instruction:
Respond only with the figurative translation.
Do not add explanations or additional text.

Examples

Example 1
Input ({source_language}):
Achili ni nyere, chila mmwenga ana zakwake

Output ({target_language}):
People think differently; no two minds are the same.

Example 2
Input ({source_language}):
Achiyeshiba kamanya mndzala

Output ({target_language}):
People who are comfortable often fail to understand the suffering of others.

Example 3
Input ({source_language}):
Achiyetsupa mbere, siye myao

Output ({target_language}):
Competition weakens friendship.

Now translate the following proverb

Input ({source_language}):
{proverb}

Output ({target_language}):

FMT prompt (Figurative Meaning Translation - Luo)

Task:
You are a professional translator specializing in proverbs. Your task is to infer the figurative meaning of the following proverb from {source_language} to {target_language}.
A figurative translation MUST:
- Explicitly state the underlying principle or lesson.
- Remove metaphor, imagery, and references to animals or objects.
- Use clear, plain, abstract language.
- Remain culturally faithful to the meaning of the proverb.

Do NOT:
- Use another proverb.
- Retain metaphorical imagery.
- Produce poetic or idiomatic wording.

Instruction:
Respond only with the figurative translation.
Do not add explanations or additional text.

Examples

Example 1
Input ({source_language}):
Acham ayieng' ne onego suna.

Output ({target_language}):
Greed and excess lead to one's downfall.
Output ({target_language}):
Uchoyo na kupita kiasi husababisha maangamizi.

Example 2
Input ({source_language}):
Achang' chacha ngeta mopadore

Output ({target_language}):
An exception to what is expected.
Output ({target_language}):
Tofauti na ilivyotarajiwa.

Example 3
Input ({source_language}):
Achi ngure god kabondo.

Output ({target_language}):
A burden that causes lasting strain.
Output ({target_language}):
Mzigo unaosababisha madhara ya muda mrefu.

Now translate the following proverb

Input ({source_language}):
{proverb}

Output ({target_language}):

L Human Evaluation

Human Evaluation Instructions (Afri-Proverb)

Task:
You will evaluate translations of African proverbs produced by different models. Your goal is to assess the quality of each translation according to four evaluation criteria and also compare model outputs directly.
These criteria evaluate whether the translation:
- Correctly conveys the meaning of the proverb
- Preserves the semantic content of the original proverb
- Is grammatically fluent
- Sounds culturally natural in the target language

Please read the **source proverb** and the **model translations** carefully before assigning scores.
Evaluation Metrics
1. Meaning Accuracy
Assess whether the translation correctly conveys the **intended meaning or interpretation** of the proverb.
- 5 – Meaning completely correct and fully captures the message.
- 4 – Mostly correct; minor nuance missing.
- 3 – Partially correct; meaning preserved but incomplete.
- 2 – Mostly incorrect interpretation.
- 1 – Completely incorrect meaning.

2. Adequacy
Measures whether the **meaning of the source proverb is preserved**.
- For **literal translation**: meaning and imagery preserved.
- For **figurative translation**: interpretation or moral preserved.
- 5 – All meaning preserved.
- 4 – Minor meaning loss.
- 3 – Partial preservation.
- 2 – Major meaning loss.
- 1 – Meaning missing or incorrect.

3. Fluency
Evaluate whether the translation is **grammatically correct and natural**.
- 5 – Perfectly fluent.
- 4 – Minor issues.
- 3 – Understandable but awkward.
- 2 – Difficult to read.
- 1 – Unintelligible.

4. Cultural Naturalness
Assess whether the translation sounds **culturally appropriate as a proverb**.
- 5 – Very natural proverb expression.
- 4 – Mostly natural.
- 3 – Neutral.
- 2 – Somewhat unnatural.
- 1 – Completely unnatural.

Pairwise Model Comparison
Select which translation is better overall.
- Model A is better
- Model B is better
- Tie

Choose A or B if one translation is clearly better in meaning, adequacy, fluency, or cultural naturalness. Choose **Tie** if both are equally good or poor.
Open-Ended Comment (Optional)
You may comment on:
- Translation errors
- Cultural issues
- Loss of metaphor or imagery
- Strengths or weaknesses

Comment:
(Write your comment here)
Summary
- Correctness → Meaning Accuracy
- Semantic Preservation → Adequacy
- Grammatical Quality → Fluency
- Cultural Appropriateness → Cultural Naturalness

Please evaluate translations **independently and objectively**.

Language	Country	# Proverbs	Tokens(Prov Col)	Avg./Prov.
Gikuyu	Kenya	554	5,667	10.2
Meru	Kenya	520	5,901	11.3
Luo	Kenya	519	4,343	8.4
Ekegusii	Kenya	514	7,141	13.9
Nubian 2	Kenya	500	5,915	11.8
Nubian	Kenya	500	39,481	79.0
Somali	Somalia	500	7,420	14.8
Sukuma	Tanzania	500	5,850	11.7
Digo	Kenya	389	4,179	10.7
Kamba	Kenya	303	4,898	16.2
Maasai	Kenya	205	2,284	11.1
Tshiluba	DRC	110	1,657	15.1
Kihangaza	Tanzania	102	1,255	12.3
Kihara	Tanzania	101	1,249	12.4
Kwele	DRC	100	1,490	14.9
Tetela	DRC	100	1,340	13.4
Hema	DRC	100	1,421	14.2
Hemba	DRC	100	1,384	13.8
Holoholo	DRC	100	1,285	12.8
Nande	DRC	100	1,587	15.9
Taabwa	DRC	100	1,433	14.3
Borana	Ethiopia	100	1,436	14.4
Burji	Ethiopia	100	1,259	12.6
Nandi	Kenya	100	1,062	10.6
Nyala	Kenya	100	1,267	12.7
Olusamia	Kenya	100	1,148	11.5
Orma	Kenya	100	1,337	13.4
Rendille	Kenya	100	1,223	12.2
Samburu	Kenya	100	935	9.4
Teso	Kenya	100	1,713	17.1
Tugen	Kenya	100	1,419	14.2
Turkana	Kenya	100	1,131	11.3
Gweno	Tanzania	100	1,370	13.7
Makonde	Tanzania	100	1,468	14.7
Nyaturu	Tanzania	100	1,187	11.9
Pare	Tanzania	100	1,220	12.2
Zigula	Tanzania	100	1,183	11.8
Alur	Uganda	100	960	9.6
Chiga	Uganda	100	1,645	16.4
Ganda	Uganda	100	1,340	13.4
Rufumbira	Uganda	100	1,270	12.7
Runyoro	Uganda	100	1,401	14.0
Soga	Uganda	100	1,329	13.3
Tooro	Uganda	100	1,661	16.6
Bangubangu	DRC	98	1,489	15.2

Table 9: **Total languages covered (44).** Number of proverbs, total tokens in the proverb column, and average tokens per proverb for each language in AFRI-PROVERB.

M Detailed Dataset Statistics

N Results and Analysis

For the bottom row, for each of the four *Row Avg* columns we take the mean over all languages:

$$\text{BottomAvg}_{\text{direction}} = \frac{1}{N} \sum_{\ell=1}^{N} \text{RowAvg}_{\ell,\ \text{direction}},$$

where N is the number of languages (45). As before, these averages use only the BLEU and ChrF++ row–average values; COMET score is NOT included in any of the four left and bottom-row aggregates.

Language	Country	Total	Avg	Med	Min	Max	Std
Nubian	Kenya	39,481	79.0	74.0	8	172	41.0
Somali	Somalia	7,420	14.8	14.0	5	45	10.0
Ekegusii	Kenya	7,141	13.9	14.0	7	22	3.8
Nubian 2	Kenya	5,915	11.8	11.0	6	26	5.0
Meru	Kenya	5,901	11.3	10.0	4	42	9.5
Sukuma	Tanzania	5,850	11.7	11.0	4	47	10.8
Gikuyu	Kenya	5,667	10.2	10.0	4	23	4.8
Kamba	Kenya	4,898	16.2	15.0	6	48	10.5
Luo	Kenya	4,343	8.4	8.0	2	28	6.5
Digo	Kenya	4,179	10.7	10.0	4	34	7.5
Maasai	Kenya	2,284	11.1	10.0	6	31	6.2
Teso	Kenya	1,713	17.1	16.5	6	40	8.5
Tooro	Uganda	1,661	16.6	15.5	4	37	8.2
Tshiluba	DRC	1,657	15.1	14.0	5	30	6.2
Chiga	Uganda	1,645	16.4	16.0	4	37	8.2
Nande	DRC	1,587	15.9	15.0	6	28	5.5
Kwele	DRC	1,490	14.9	14.0	4	33	7.2
Bangubangu	DRC	1,489	15.2	15.0	5	28	5.8
Makonde	Tanzania	1,468	14.7	14.0	6	36	7.5
Borana	Ethiopia	1,436	14.4	14.0	6	30	6.0
Taabwa	DRC	1,433	14.3	14.0	5	27	5.5
Hema	DRC	1,421	14.2	13.0	5	31	6.5
Tugen	Kenya	1,419	14.2	14.0	6	31	6.2
Runyoro	Uganda	1,401	14.0	13.0	5	32	6.8
Hemba Pov	DRC	1,384	13.8	13.0	5	27	5.5
Gweno	Tanzania	1,370	13.7	13.0	6	29	5.8
Tetela	DRC	1,340	13.4	13.0	7	29	5.5
Ganda	Uganda	1,340	13.4	13.0	6	25	4.8
Orma	Kenya	1,337	13.4	13.0	7	34	6.8
Soga	Uganda	1,329	13.3	13.0	5	26	5.2
Holoholo	DRC	1,285	12.8	13.0	6	26	5.0
Rufumbira	Uganda	1,270	12.7	12.0	6	22	4.0
Nyala	Kenya	1,267	12.7	12.0	5	23	4.5
Burji	Ethiopia	1,259	12.6	12.0	4	32	7.0
Kihangaza	Tanzania	1,255	12.3	12.0	7	19	3.0
Kihara	Tanzania	1,249	12.4	12.0	6	22	4.0
Rendille	Kenya	1,223	12.2	11.5	5	28	5.8
Pare	Tanzania	1,220	12.2	12.0	6	21	3.8
Nyaturu	Tanzania	1,187	11.9	11.5	6	21	3.8
Zigula	Tanzania	1,183	11.8	12.0	5	29	6.0
Olusamia	Kenya	1,148	11.5	11.0	4	23	4.8
Turkana	Kenya	1,131	11.3	10.0	5	29	6.0
Nandi	Kenya	1,062	10.6	10.0	4	26	5.5
Alur	Uganda	960	9.6	9.0	4	16	3.0
Samburu	Kenya	935	9.4	9.0	5	25	5.0

Table 10: **Token statistics by language (proverb column).** We report total token count, average length, median length, minimum, maximum, and standard deviation of proverb length per language.

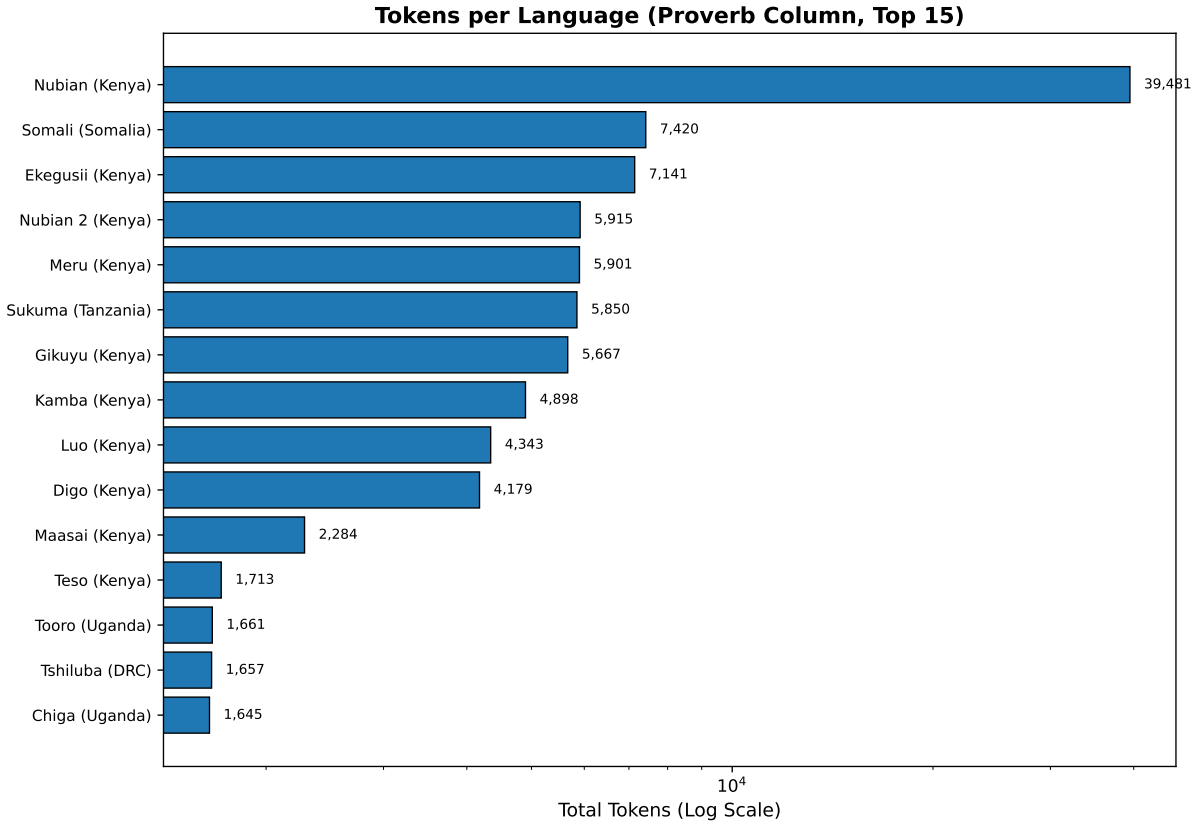

Figure 7: **Total proverb tokens per language (Top 15; log scale)**

N.1 Gemma3-4b-it(zero-shot)

N.2 Qwen3-4b-it(zero-shot)

N.3 Llama-3.2-3B (Zero-shot)

N.4 SmolLM3-3B(zero-shot)

N.5 Minimax-01(zero-shot)

N.6 GLM-5(zero-shot)

N.7 Kimi K2.5 (zero-shot)

N.8 GPT-5 mini (zero-shot)

N.9 Grok 4.1 Fast(zero-shot)

N.10 Gemini 3 Flash Preview(zero-shot)

N.11 Qwen-3-4B-It(3-shots)

N.12 Llama-3.2-3B (3-shots)

N.13 SmolLM3-3B (3-shots)

N.14 Gemma3-4b-it (3-shot)

O Results and Analysis Visualization

P Human Evaluation

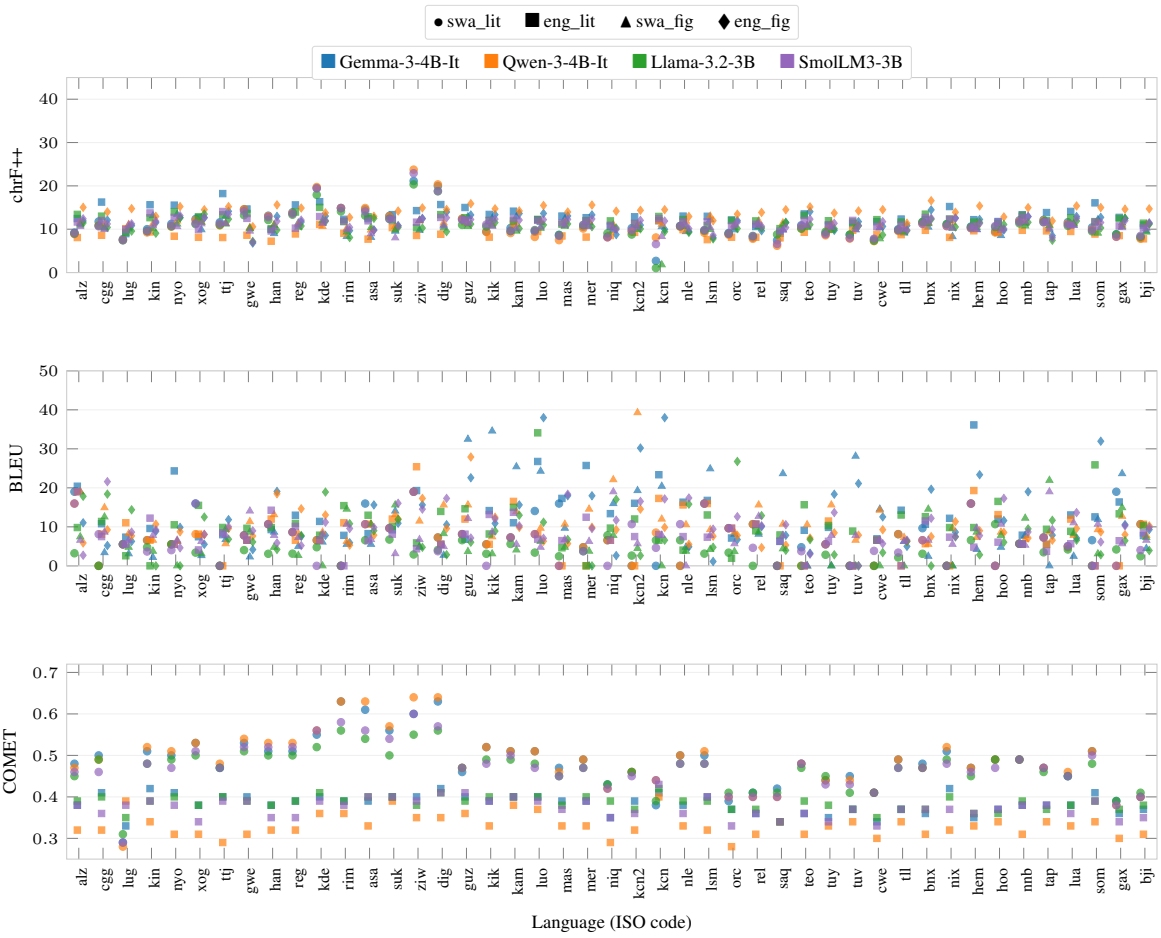

Figure 8: Few-shot (3-shot) results for four models (Gemma, Llama, Qwen, and SmolLM3) across four tasks on Afri-Proverb (45 languages), evaluated using chrF++, BLEU, and COMET. Colors denote models and marker shapes indicate task types. COMET is reported for swa_lit and eng_lit only.

Language	Column	Avg Chars	Avg Words	Total Tokens	Avg Tokens	Min	Max
Turkana	turkana_prov	29.8	4.1	1,131	11.3	5	29
	eng_literal	34.1	6.6	686	6.9	2	16
	swa_literal	31.6	5.0	1,052	10.5	4	22
	eng_figurative	52.6	7.2	782	7.8	4	14
	swa_figurative	47.9	7.1	1,542	15.4	9	24
Tugen	tugen_prov	38.3	5.8	1,419	14.2	6	31
	eng_literal	51.9	10.4	1,170	11.7	3	38
	swa_literal	46.4	7.3	1,573	15.7	6	46
	eng_figurative	97.1	17.6	1,969	19.7	7	47
	swa_figurative	52.9	7.8	1,823	18.2	7	33
Teso	teso_prov	47.7	7.2	1,713	17.1	6	40
	eng_literal	48.9	9.5	1,101	11.0	4	26
	swa_literal	41.6	6.4	1,485	14.8	7	33
	eng_figurative	38.8	5.7	718	7.2	4	13
	swa_figurative	37.0	5.2	1,299	13.0	7	21
Samburu	samburu_prov	24.8	3.5	935	9.4	5	25
	eng_literal	46.2	8.8	947	9.5	3	21
	swa_literal	40.3	6.1	1,337	13.4	6	26
	eng_figurative	56.6	7.4	794	7.9	5	13
	swa_figurative	52.4	7.5	1,696	17.0	9	27
Rendille	rendille_prov	32.1	5.1	1,223	12.2	5	28
	eng_literal	48.5	9.5	1,051	10.5	4	27
	swa_literal	46.6	7.1	1,589	15.9	6	28
	eng_figurative	78.7	14.0	1,537	15.4	4	40
	swa_figurative	39.2	5.8	1,360	13.6	7	20
Orma	orma_prov	37.4	6.3	1,337	13.4	7	34
	eng_literal	66.0	12.9	1,483	14.8	6	39
	swa_literal	59.4	9.0	2,053	20.5	9	39
	eng_figurative	60.6	11.7	1,332	13.3	4	31
	swa_figurative	41.3	6.0	1,447	14.5	7	27
Olusamia	olusamia_prov	30.7	4.0	1,148	11.5	4	23
	eng_literal	49.4	9.6	1,117	11.2	3	24
	swa_literal	44.2	6.6	1,562	15.6	6	32
	eng_figurative	169.0	31.1	3,360	33.6	16	66
	swa_figurative	55.6	7.8	1,875	18.8	12	33
Nyala	abanyala_prov	35.9	4.5	1,267	12.7	5	23
	eng_literal	47.8	9.2	1,019	10.2	4	21
	swa_literal	41.5	6.2	1,438	14.4	4	30
	eng_figurative	120.8	23.0	2,506	25.1	12	41
	swa_figurative	41.6	6.2	1,454	14.5	9	26
Nubian	nubian_prov	33.6	4.7	39,481	79.0	8	172
	eng_literal	43.5	8.6	5,203	10.4	4	25
	swa_literal	38.4	5.7	6,854	13.7	5	28
	eng_figurative	52.0	8.1	5,382	10.8	4	27
	swa_figurative	47.5	6.7	8,424	16.8	4	42
Nubian 2	nubian_2_prov	32.0	4.6	5,915	11.8	6	26
	eng_literal	43.5	8.6	5,203	10.4	4	25
	swa_literal	38.4	5.7	6,854	13.7	5	28
	eng_figurative	52.0	8.1	5,382	10.8	4	27
	swa_figurative	47.5	6.7	8,424	16.8	4	42
Digo	digo_prov	29.4	4.4	4,179	10.7	4	34
	eng_literal	39.6	7.7	3,793	9.8	3	19
	swa_literal	31.2	4.8	4,500	11.6	5	23
	eng_figurative	31.9	4.6	2,372	6.1	3	13
	swa_figurative	29.1	4.1	4,069	10.5	5	24
Ekegusii	ekegusii_prov	35.4	4.9	7,141	13.9	7	22
	eng_literal	29.8	5.5	3,579	7.0	4	11
	swa_literal	26.5	4.6	4,986	9.7	5	17
	eng_figurative	42.6	6.1	3,994	7.8	4	14
	swa_figurative	38.9	5.9	6,898	13.4	6	24
Gikuyu	gikuyu_prov	28.6	4.5	5,667	10.2	4	23
	eng_literal	39.6	7.5	5,153	9.3	3	22
	swa_literal	35.5	5.2	7,093	12.8	4	39
	eng_figurative	39.5	5.8	4,157	7.5	4	17
	swa_figurative	35.1	5.0	6,870	12.4	6	26
Kamba	kamba_prov	31.4	5.2	4,898	16.2	6	48
	eng_literal	43.4	8.2	3,072	10.1	3	22
	swa_literal	34.9	5.3	3,864	12.8	5	27
	eng_figurative	32.7	4.7	1,882	6.2	4	16
	swa_figurative	30.5	4.3	3,293	10.9	5	18
Luo	luo_prov	22.6	4.3	4,343	8.4	2	28
	eng_literal	37.3	7.1	4,689	9.0	3	25
	swa_literal	32.2	4.8	6,075	11.7	4	26
	eng_figurative	40.4	5.9	3,890	7.5	3	17
	swa_figurative	37.2	5.2	6,762	13.0	6	24
Maasai	maasai_prov	30.4	4.2	2,284	11.1	6	31
	eng_literal	48.2	9.6	2,315	11.3	4	29
	swa_literal	41.6	6.3	2,958	14.4	5	33
	eng_figurative	42.6	6.4	1,660	8.1	4	17
	swa_figurative	38.1	5.6	2,692	13.1	6	23
Meru	meru_prov	31.2	4.9	5,901	11.3	4	42
	eng_literal	49.9	9.6	6,128	11.8	4	45
	swa_literal	42.5	6.4	7,771	14.9	5	35
	eng_figurative	45.6	6.5	4,148	8.0	3	15
	swa_figurative	42.8	6.0	7,643	14.7	3	27
Nandi	nandi_prov	28.1	4.3	1,062	10.6	4	26
	eng_literal	48.2	9.7	1,149	11.5	4	28
	swa_literal	42.9	6.6	1,458	14.6	5	34
	eng_figurative	120.8	21.8	2,431	24.3	12	46
	swa_figurative	48.9	7.1	1,683	16.8	9	28

Table 11: Per-language token and length statistics across Kenyan proverb and translation columns.

Language	Column	Avg Chars	Avg Words	Total Tokens	Avg Tokens	Min	Max
Gweno	gweno_prov	35.9	5.6	1,370	13.7	6	29
	eng_literal	47.3	9.4	1,156	11.6	5	23
	swa_literal	39.0	6.1	1,399	14.0	6	28
	eng_figurative	237.6	42.6	4,845	48.4	18	109
	swa_figurative	76.6	11.2	2,585	25.8	11	49
Kihangaza	kihangaza_prov	31.9	4.2	1,255	12.3	7	19
	eng_literal	110.9	21.0	2,463	24.1	5	50
	swa_literal	36.1	5.1	1,316	12.9	4	24
	eng_figurative	62.5	9.8	1,186	11.6	6	19
	swa_figurative	57.1	8.7	2,005	19.7	8	48
Kihara	kihara_prov	31.9	4.1	1,249	12.4	6	22
	eng_literal	46.4	9.1	1,074	10.6	4	19
	swa_literal	35.9	5.5	1,299	12.9	5	24
	eng_figurative	71.9	12.2	1,475	14.6	8	22
	swa_figurative	46.8	6.9	1,683	16.7	10	25
Makonde	makonde_prov	41.4	6.2	1,468	14.7	6	36
	eng_literal	45.4	9.1	1,058	10.6	4	23
	swa_literal	36.9	5.6	1,339	13.4	6	25
	eng_figurative	73.4	12.1	1,450	14.5	6	44
	swa_figurative	53.5	7.9	1,834	18.3	10	31
Nyaturu	nyaturu_prov	32.2	4.4	1,187	11.9	6	21
	eng_literal	40.3	7.6	907	9.1	3	21
	swa_literal	28.6	4.4	1,063	10.6	4	19
	eng_figurative	118.4	20.9	2,361	23.6	10	46
	swa_figurative	71.8	10.2	2,414	24.1	10	44
Pare	pare_prov	32.6	4.4	1,220	12.2	6	21
	eng_literal	40.4	7.7	910	9.1	3	21
	swa_literal	28.6	4.4	1,064	10.6	4	19
	eng_figurative	118.0	20.9	2,341	23.4	10	46
	swa_figurative	89.5	12.8	2,930	29.3	12	47
Sukuma	sukuma_prov	30.9	4.3	5,850	11.7	4	47
	eng_literal	39.6	7.8	4,472	8.9	2	27
	swa_literal	31.8	4.7	5,540	11.1	3	37
	eng_figurative	66.3	10.0	5,928	11.9	5	29
	swa_figurative	60.3	8.7	9,995	20.0	9	48
Zigula	zigula_prov	30.9	4.4	1,183	11.8	5	29
	eng_literal	53.3	10.3	1,346	13.5	4	30
	swa_literal	29.9	4.6	1,110	11.1	6	21
	eng_figurative	62.2	11.3	1,324	13.2	4	28
	swa_figurative	37.5	5.6	1,319	13.2	6	29

Table 12: Per-language token and length statistics for Tanzanian proverb and translation columns.

Language	Column	Avg Chars	Avg Words	Total Tokens	Avg Tokens	Min	Max
Alur	alur_prov	23.5	4.5	960	9.6	4	16
	eng_literal	37.9	6.9	896	9.0	4	17
	swa_literal	32.8	4.9	1,187	11.9	5	25
	eng_figurative	47.2	8.0	964	9.6	4	18
	swa_figurative	36.6	5.3	1,263	12.6	7	27
Chiga	chiga_prov	46.0	5.6	1,645	16.4	4	37
	eng_literal	47.5	9.5	1,008	10.1	4	20
	swa_literal	45.8	7.1	1,545	15.4	6	31
	eng_figurative	102.7	18.9	2,086	20.9	9	33
	swa_figurative	67.5	9.8	2,291	22.9	11	38
Ganda	ganda_prov	35.2	4.8	1,340	13.4	6	25
	swa_literal	52.1	10.3	1,225	12.2	5	25
	eng_literal	46.4	7.3	1,601	16.0	6	33
	eng_figurative	89.9	15.9	1,829	18.3	8	37
	swa_figurative	42.8	6.1	1,491	14.9	7	37
Rufumbira	rufumbira_prov	34.4	4.8	1,270	12.7	6	22
	eng_literal	53.9	10.8	1,150	11.5	3	24
	swa_literal	42.0	6.5	1,449	14.5	6	29
	eng_figurative	61.2	11.9	1,279	12.8	4	44
	swa_figurative	38.6	5.4	1,360	13.6	8	22
Runyoro	runyoro_prov	39.3	5.3	1,401	14.0	5	32
	swa_literal	45.7	7.0	1,549	15.5	6	27
	eng_literal	54.7	11.0	1,192	11.9	4	25
	eng_figurative	50.7	9.0	1,038	10.4	4	25
	swa_figurative	36.1	5.1	1,288	12.9	6	24
Soga	soga_prov	34.3	3.8	1,329	13.3	5	26
	eng_literal	53.4	10.4	1,208	12.1	4	21
	swa_literal	41.1	5.5	1,477	14.8	7	32
	eng_figurative	45.4	8.5	1,000	10.0	4	20
	swa_figurative	37.7	5.6	1,292	12.9	5	23
Tooro	tooro_prov	46.1	5.6	1,661	16.6	4	37
	swa_literal	45.5	6.9	1,594	15.9	7	32
	eng_literal	48.7	9.5	1,108	11.1	5	22
	eng_figurative	73.9	12.3	1,435	14.4	8	24
	swa_figurative	50.7	7.2	1,758	17.6	11	28

Table 13: Per-language token and length statistics for Ugandan proverb and translation columns.

Language	Column	Avg Chars	Avg Words	Total Tokens	Avg Tokens	Min	Max
Bangubangu	bangubangu_prov	41.0	6.0	1,489	15.2	5	28
	eng_literal	46.9	9.3	1,062	10.8	4	19
	swa_literal	41.7	6.6	1,454	14.8	8	30
	eng_figurative	52.6	8.3	983	10.0	4	25
	swa_figurative	45.1	6.8	1,509	15.4	8	29
Hema	hema_prov	39.9	5.4	1,421	14.2	5	31
	eng_literal	53.1	10.7	1,156	11.6	4	26
	swa_literal	44.0	6.6	1,494	14.9	6	28
	eng_figurative	47.1	8.5	890	8.9	3	21
	swa_figurative	40.1	6.0	1,314	13.1	5	24
Hemba	hemba_prov	37.4	5.4	1,384	13.8	5	27
	eng_literal	45.6	8.9	1,081	10.8	3	24
	swa_literal	39.5	6.2	1,368	13.7	5	27
	eng_figurative	74.0	11.8	1,433	14.3	8	26
	swa_figurative	46.8	6.8	1,660	16.6	10	32
Holoholo	holoholo_prov	36.3	5.3	1,285	12.8	6	26
	eng_literal	57.6	11.3	1,321	13.2	6	24
	swa_literal	48.7	7.7	1,694	16.9	6	45
	eng_figurative	125.0	23.2	2,612	26.1	8	42
	swa_figurative	66.2	9.6	2,261	22.6	10	37
Kwele	kwele_prov	32.4	6.1	1,490	14.9	4	33
	eng_literal	46.8	9.2	1,092	10.9	5	18
	swa_literal	39.0	5.9	1,354	13.5	6	26
	eng_figurative	76.4	12.5	1,475	14.8	8	25
	swa_figurative	46.8	6.7	1,618	16.2	9	30
Nande	nande_prov	41.0	5.2	1,587	15.9	6	28
	eng_literal	51.7	10.5	1,191	11.9	5	24
	swa_literal	40.3	6.1	1,445	14.4	6	26
	eng_figurative	52.4	9.7	1,113	11.1	4	24
	swa_figurative	39.1	5.7	1,341	13.4	5	22
Taabwa	taabwa_prov	38.0	5.6	1,433	14.3	5	27
	eng_literal	52.5	10.2	1,238	12.4	4	26
	swa_literal	46.0	7.4	1,630	16.3	5	34
	eng_figurative	183.6	33.5	3,753	37.5	10	74
	swa_figurative	52.0	7.5	1,780	17.8	8	37
Tetela	tetela_prov	37.8	6.0	1,340	13.4	7	29
	eng_literal	41.2	7.8	870	8.7	4	20
	swa_literal	32.5	5.2	1,107	11.1	5	19
	eng_figurative	46.9	6.6	839	8.4	3	19
	swa_figurative	48.5	6.9	1,352	13.5	5	24
Tshiluba	tshiluba_prov	41.9	6.0	1,657	15.1	5	30
	eng_literal	48.9	9.2	1,208	11.0	3	21
	swa_literal	40.8	6.2	1,570	14.3	8	27
	eng_figurative	58.2	9.1	1,236	11.2	6	23
	swa_figurative	43.4	6.4	1,645	15.0	8	30

Table 14: Per-language token and length statistics for DRC proverb and translation columns.

Language	Column	Avg Chars	Avg Words	Total Tokens	Avg Tokens	Min	Max
Borana	borana_prov	37.2	5.6	1,436	14.4	6	30
	eng_literal	65.4	12.6	1,469	14.7	5	33
	swa_literal	67.2	10.4	2,288	22.9	9	48
	eng_figurative	59.2	11.1	1,287	12.9	5	27
	swa_figurative	53.2	7.9	1,858	18.6	7	34
Burji	burji_prov	31.4	4.7	1,259	12.6	4	32
	eng_literal	56.8	11.1	1,331	13.3	6	35
	swa_literal	54.4	8.4	1,862	18.6	9	40
	eng_figurative	57.1	10.6	1,248	12.5	5	29
	swa_figurative	38.6	5.6	1,350	13.5	8	22

Table 15: Per-language token and length statistics for Ethiopian proverb and translation columns.

Language	Column	Avg Chars	Avg Words	Total Tokens	Avg Tokens	Min	Max
Somali	somali_prov	39.5	6.7	7,420	14.8	5	45
	eng_literal	47.8	9.2	5,765	11.5	4	33
	swa_literal	41.2	6.0	7,426	14.9	6	32
	eng_figurative	59.5	9.0	5,537	11.1	4	24
	swa_figurative	52.3	7.7	8,936	17.9	8	32

Table 16: Per-language token and length statistics for Somali proverb and translation columns.

lang	Gemma-3-4B-It (Zero-shot)										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	**19.00**	9.54	0.477	**20.41**	12.16	0.378	**19.00**	10.93	11.04	**11.83**	14.27	16.29	14.96	11.44
chiga	0.00	**11.48**	0.492	4.55	**16.11**	0.402	0.00	**10.27**	**11.50**	11.40	5.74	10.33	5.13	11.45
ganda	5.52	**7.75**	0.293	7.77	**9.04**	0.330	3.09	**10.39**	6.27	**11.13**	6.64	8.40	6.74	8.70
rufumbira	6.57	**10.01**	0.507	5.20	**15.97**	0.427	10.68	**11.53**	7.97	**11.18**	8.29	10.58	11.11	9.57
runyoro	5.52	**11.13**	0.498	**22.25**	15.73	0.411	8.64	**13.44**	11.90	**12.46**	8.33	18.99	11.04	12.18
soga	**15.97**	12.10	0.533	6.27	**12.40**	0.365	6.57	**9.73**	11.99	**12.85**	14.04	9.34	8.15	12.42
tooro	0.00	**11.36**	0.475	13.40	**17.51**	0.398	**13.65**	13.12	7.64	**14.11**	5.68	15.46	13.38	10.88
gweno	7.81	**14.63**	0.525	6.57	**14.11**	0.395	2.29	**9.09**	4.17	**6.71**	11.22	10.34	5.69	5.44
kihangaza	10.68	**13.14**	0.517	9.29	**10.05**	0.379	7.81	**9.26**	11.04	**12.64**	11.91	9.67	8.54	11.84
kihara	8.64	**14.15**	0.511	13.81	**15.22**	0.397	4.99	**11.05**	**15.09**	12.26	11.40	14.52	8.02	13.68
makonde	6.57	**20.06**	0.561	6.84	**17.21**	0.406	0.00	**11.49**	5.52	**13.30**	13.31	12.03	5.75	9.41
nyaturu	0.00	**14.82**	0.632	6.64	**12.59**	0.395	5.52	**8.14**	7.77	**9.28**	7.41	9.62	6.83	8.52
pare	14.06	**14.86**	0.609	8.26	**12.88**	0.381	3.09	**8.09**	**15.62**	9.88	14.46	10.57	5.59	12.75
sukuma	9.65	**12.95**	0.566	11.48	**13.13**	0.406	**21.83**	9.60	8.98	**11.03**	11.30	12.31	15.72	10.01
zigula	**19.00**	20.79	0.606	**19.00**	14.65	0.402	6.57	**9.83**	11.99	**11.94**	19.89	16.83	8.20	11.97
digo	7.27	**20.15**	0.623	5.52	**15.89**	0.420	2.72	**11.01**	10.55	**14.05**	13.71	10.70	6.87	12.30
ekegusii	8.12	**12.51**	0.469	**20.86**	14.71	0.396	**42.04**	12.58	11.99	**13.26**	10.32	17.79	27.31	12.63
gikuyu	5.52	**9.44**	0.517	8.82	**13.46**	0.395	**34.57**	12.51	12.33	**13.20**	7.48	11.14	23.54	12.76
kamba	7.27	**10.33**	0.508	9.84	**14.08**	0.408	8.64	**12.74**	**22.96**	13.47	8.80	11.96	10.69	18.22
luo	**16.52**	9.39	0.511	13.65	**12.33**	0.398	23.64	**11.40**	**100.00**	13.28	12.96	13.99	17.52	56.64
maasai	0.00	**8.70**	0.471	13.04	**13.53**	0.395	22.96	**12.51**	**35.36**	13.31	4.35	13.28	17.73	24.33
meru	4.77	**10.49**	0.487	15.04	**12.51**	0.390	4.03	**11.46**	6.83	**13.42**	7.63	13.78	7.75	10.13
nandi	6.57	**8.27**	0.428	14.99	**11.17**	0.356	**9.65**	9.46	7.10	**8.94**	7.42	13.08	9.55	8.02
nubian_2	0.00	**10.13**	0.464	11.99	**12.97**	0.396	10.68	**11.00**	**30.21**	10.90	5.07	12.48	10.84	20.55
nubian	0.00	**2.69**	0.383	32.17	**12.07**	0.405	15.97	**10.93**	**37.99**	10.08	1.35	22.12	13.45	24.04
nyala	0.00	**10.78**	0.503	9.29	**12.58**	0.381	6.57	**9.42**	8.52	**10.29**	5.39	10.94	8.00	9.40
olusamia	**15.97**	9.47	0.509	**16.09**	13.10	0.396	8.64	**12.37**	5.99	**7.83**	12.72	14.60	10.51	6.91
orma	9.65	**9.18**	0.395	10.68	**12.24**	0.371	4.77	**9.85**	19.00	**12.02**	9.42	11.46	7.31	15.51
rendille	**10.68**	8.04	0.408	**10.68**	10.88	0.358	**15.97**	10.84	11.04	**9.85**	9.36	10.78	13.40	10.44
samburu	**15.97**	7.27	0.427	5.52	**9.44**	0.345	9.29	**10.08**	0.00	**11.95**	11.62	7.48	9.69	5.98
teso	4.77	**10.90**	0.478	5.88	**13.78**	0.367	0.00	**10.55**	13.13	**14.70**	7.84	9.83	5.27	13.92
tugen	5.52	**9.18**	0.445	11.45	**11.03**	0.349	4.77	**10.47**	7.46	**10.20**	7.35	11.24	7.62	8.83
turkana	0.00	**8.39**	0.447	0.00	**11.89**	0.377	**28.12**	10.24	**34.57**	11.93	4.19	5.94	19.18	23.25
kwele	0.00	**7.62**	0.407	6.92	**11.53**	0.347	7.13	**10.10**	7.13	**10.23**	3.81	9.23	8.62	8.68
tetela	8.12	**9.80**	0.489	5.93	**12.42**	0.371	4.20	**9.29**	6.74	**10.91**	8.96	9.18	6.75	8.83
bangubangu	5.93	**11.65**	0.476	9.29	**12.20**	0.354	11.99	**10.81**	17.97	**14.22**	8.79	10.75	11.40	16.10
hema	0.00	**11.36**	0.511	9.58	**14.89**	0.418	0.00	**8.38**	8.64	**12.58**	5.68	12.24	4.19	10.61
hemba	**15.97**	10.48	0.466	**14.92**	12.12	0.357	3.09	**10.16**	6.99	**12.17**	13.23	13.52	6.63	9.58
holoholo	0.00	**9.64**	0.491	11.99	**11.56**	0.370	**12.44**	8.61	**16.52**	9.78	4.82	11.78	10.53	13.15
nande	5.67	**12.33**	0.491	5.70	**13.25**	0.384	**15.97**	11.93	10.55	**12.67**	9.00	9.48	13.95	11.61
taabwa	7.27	**12.15**	0.469	5.52	**13.63**	0.382	2.84	**9.67**	7.06	**8.29**	9.71	9.57	6.25	7.68
tshiluba	6.27	**11.82**	0.463	12.93	**14.03**	0.377	9.29	**11.32**	6.20	**13.40**	9.05	13.48	10.30	9.80
somali	10.68	**10.27**	0.500	12.55	**16.39**	0.417	**19.13**	11.86	31.95	**12.81**	10.48	14.47	15.50	22.38
borana	**19.00**	8.76	0.390	**16.37**	12.65	0.368	12.70	**11.80**	**10.57**	12.01	13.88	14.51	12.25	11.29
burji	**10.68**	8.16	0.406	6.26	**10.66**	0.360	8.12	**9.88**	**15.62**	11.82	9.42	8.46	8.99	13.72
Avg	7.49	11.07	0.46	11.00	13.15	0.38	10.53	10.65	14.88	11.68	**9.28**	**12.08**	**10.59**	**13.28**

Table 17: Zero-shot **literal and figurative** generation tasks across Afri-Proverb languages using **Gemma-3-4B-It**. For each language direction we report **BLEU**, **ChrF++**, and **COMET** where applicable. For each language and task, the higher score is in **bold**. The last four columns give row-wise averages per language direction (averaging BLEU and ChrF++ only; COMET is *not* included). The bottom row gives column-wise averages

lang	Qwen-3-4B-It (Zero-shot)										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	**15.97**	8.86	0.470	**19.00**	7.98	0.320	**19.00**	11.35	5.87	**14.70**	12.41	13.49	**15.17**	10.28
chiga	0.00	**10.93**	0.490	**8.12**	8.72	0.318	8.64	**10.94**	7.46	**13.82**	5.47	8.42	9.79	**10.64**
ganda	5.52	**7.54**	0.280	3.98	**10.28**	0.405	7.81	**11.08**	**19.01**	15.44	6.53	7.13	9.44	**17.23**
rufumbira	6.57	**9.23**	0.520	0.00	**9.53**	0.350	5.52	**11.42**	5.30	**13.63**	7.90	4.76	8.47	**9.47**
runyoro	5.52	**10.47**	0.510	**5.52**	8.07	0.309	5.67	**11.85**	3.72	**15.34**	7.99	6.80	8.76	**9.53**
soga	8.12	**12.06**	0.530	**8.12**	7.87	0.309	7.81	**11.86**	5.30	**14.82**	**10.09**	7.99	9.84	10.06
tooro	0.00	**10.82**	0.480	0.00	**8.19**	0.293	8.64	**11.51**	7.86	**15.16**	5.41	4.10	10.08	**11.51**
gweno	7.81	**14.53**	0.535	**6.57**	8.78	0.312	**11.34**	10.19	7.65	**10.69**	**11.17**	7.67	10.77	9.17
kihangaza	10.68	**13.07**	0.531	**11.99**	6.99	0.312	**16.52**	10.90	11.39	**15.98**	**11.87**	9.49	13.71	13.68
kihara	8.64	**13.90**	0.529	**6.57**	8.71	0.310	8.26	**11.12**	**14.98**	15.04	**11.27**	7.64	9.69	15.01
makonde	6.57	**20.08**	0.565	**6.57**	11.31	0.365	9.29	**11.04**	**10.83**	13.01	**13.33**	8.94	10.16	11.92
nyaturu	0.00	**15.06**	0.633	**6.57**	9.10	0.361	6.57	**10.25**	6.34	**13.37**	7.53	7.83	8.41	**9.86**
pare	10.68	**15.10**	0.625	**10.68**	7.84	0.334	10.68	**9.77**	**10.60**	13.16	**12.89**	9.26	10.22	11.88
sukuma	9.65	**13.00**	0.571	**11.99**	10.24	0.387	**11.99**	10.59	12.03	**14.41**	11.33	11.11	11.29	**13.22**
zigula	**19.00**	23.70	0.636	**23.64**	8.52	0.346	11.48	**11.34**	**48.11**	15.27	21.35	16.08	11.41	**31.69**
digo	7.27	**20.31**	0.636	**5.52**	10.47	0.355	6.57	**11.55**	9.57	**14.49**	13.79	7.99	9.06	**12.03**
ekegusii	8.12	**12.41**	0.468	11.99	**12.58**	0.367	11.99	**11.44**	9.24	**15.78**	10.27	12.29	11.72	**12.51**
gikuyu	5.52	**9.37**	0.516	6.57	**10.00**	0.336	**14.54**	11.32	8.91	**14.78**	7.45	8.29	12.93	**11.84**
kamba	7.27	**9.14**	0.505	**16.52**	12.47	0.378	6.27	**11.18**	11.04	**14.24**	8.21	**14.50**	8.73	12.64
luo	8.12	**8.24**	0.511	**8.12**	10.94	0.369	16.52	**11.98**	10.99	**15.66**	8.18	9.53	14.25	**13.33**
maasai	0.00	**7.47**	0.464	**12.55**	10.52	0.333	9.65	**12.12**	9.58	**14.35**	3.74	11.54	10.88	**11.96**
meru	4.77	**10.15**	0.487	0.00	**9.78**	0.331	9.65	**11.87**	12.45	**15.71**	7.46	4.89	10.76	**14.08**
nandi	6.57	**8.18**	0.425	**6.57**	9.55	0.288	9.29	**11.07**	8.74	**14.30**	7.38	8.06	10.18	**11.52**
nubian_2	0.00	**9.36**	0.457	0.00	**10.67**	0.316	15.51	**11.04**	9.10	**14.77**	4.68	5.34	13.28	**11.94**
nubian	8.13	**8.06**	0.439	12.98	**13.37**	0.401	**17.29**	11.31	9.86	**14.61**	8.09	13.18	14.30	**12.24**
nyala	0.00	**10.62**	0.503	**15.62**	11.85	0.335	7.27	**11.55**	9.99	**13.54**	5.31	**13.74**	9.41	11.77
olusamia	**15.97**	9.13	0.511	**15.97**	9.53	0.327	**20.56**	12.56	7.60	**12.39**	12.55	12.75	**16.56**	9.99
orma	**9.65**	8.86	0.399	**9.65**	9.25	0.286	**21.11**	11.16	10.25	**13.93**	9.26	9.45	16.14	**12.09**
rendille	**10.68**	7.77	0.403	**10.68**	9.94	0.317	7.27	**12.06**	8.69	**14.19**	9.23	10.31	9.67	**11.44**
samburu	0.00	**6.19**	0.404	0.00	**10.24**	0.338	7.81	**11.35**	3.93	**14.63**	3.10	5.12	9.58	**9.28**
teso	0.00	**10.21**	0.480	0.00	**10.82**	0.305	9.29	**11.13**	10.55	**14.90**	5.11	5.41	10.21	**12.73**
tugen	5.52	**8.58**	0.437	**15.11**	11.56	0.327	**15.62**	11.36	8.91	**13.72**	7.05	13.34	13.49	**11.32**
turkana	0.00	**7.68**	0.440	0.00	**12.47**	0.338	6.27	**11.51**	6.27	**14.13**	3.84	6.23	8.89	**10.20**
kwele	0.00	**7.25**	0.406	6.57	**8.99**	0.302	**14.26**	10.88	9.26	**14.87**	3.63	7.78	12.57	**12.06**
tetela	**8.12**	9.54	0.493	0.00	**9.08**	0.339	6.57	**10.54**	6.27	**11.34**	**8.83**	4.54	8.55	8.80
bangubangu	6.57	**11.29**	0.474	**12.61**	10.05	0.313	9.98	**11.92**	10.70	**16.66**	8.93	11.33	10.95	**13.68**
hema	0.00	**10.82**	0.518	0.00	**8.37**	0.324	7.81	**11.60**	4.42	**13.92**	5.41	4.18	9.71	**9.17**
hemba	**15.97**	10.22	0.469	**19.30**	9.86	0.336	**17.97**	10.79	8.96	**15.44**	**13.10**	14.58	14.38	12.20
holoholo	0.00	**9.27**	0.492	**11.99**	9.59	0.343	10.68	**10.38**	8.55	**12.06**	4.63	10.79	10.53	**10.31**
nande	5.67	**11.82**	0.491	**5.52**	9.82	0.310	9.86	**11.87**	3.69	**14.58**	8.75	7.67	10.86	**9.13**
taabwa	7.27	**11.94**	0.471	**5.52**	9.34	0.343	8.64	**10.82**	**11.27**	11.14	9.61	7.43	9.73	**11.21**
tshiluba	4.77	**11.30**	0.459	**11.99**	9.58	0.327	6.27	**11.48**	9.58	**15.10**	8.03	10.79	8.88	**12.34**
somali	0.00	**9.90**	0.512	0.00	**9.04**	0.345	9.29	**11.58**	8.70	**14.96**	4.95	4.52	10.43	**11.83**
borana	0.00	**8.26**	0.382	0.00	**8.67**	0.305	**19.00**	10.91	7.99	**14.15**	4.13	4.33	14.96	**11.07**
burji	**10.68**	7.83	0.399	**10.68**	7.85	0.310	**10.68**	11.22	10.25	**14.92**	9.26	9.26	10.95	**12.59**
Avg	**6.03**	**10.88**	**0.488**	**8.03**	**9.74**	**0.333**	**10.95**	**11.26**	**9.82**	**14.29**	**8.45**	**8.89**	**11.11**	**12.05**

Table 18: Zero-shot **literal and figurative** generation task across Afri-Proverb languages using **Qwen-3-4B-It**. For each language direction we report **BLEU**, **ChrF++**, and **COMET** where applicable. The best score for each language are in **BOLD**. The last four columns give row-wise averages per language direction; the bottom row gives column-wise averages.

lang	Llama-3.2-3B (Zero-shot)										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	3.22	**9.44**	0.433	9.91	**13.60**	0.391	7.46	**10.33**	6.74	**12.93**	6.33	**11.75**	8.90	9.84
chiga	0.00	**10.83**	0.486	17.40	**13.27**	0.390	0.00	**9.85**	**19.84**	10.77	5.41	**15.33**	4.92	15.30
ganda	0.00	**7.30**	0.310	4.37	**8.93**	0.326	3.39	**10.22**	0.00	**10.23**	3.65	**6.65**	6.80	5.11
rufumbira	3.09	**9.87**	0.484	10.92	**12.64**	0.394	3.75	**9.77**	0.00	**9.36**	6.48	**11.78**	6.76	4.68
runyoro	5.52	**10.70**	0.494	9.78	**14.40**	0.403	0.00	**10.44**	8.52	**13.69**	8.11	**12.09**	5.22	11.10
soga	2.29	**11.24**	0.516	6.83	**12.61**	0.396	2.29	**11.02**	20.69	**13.67**	6.76	9.72	6.65	**17.18**
tooro	0.00	**11.10**	0.488	18.84	**13.75**	0.403	0.00	**10.73**	23.36	**13.42**	5.55	**16.30**	5.37	18.39
gweno	6.57	**13.67**	0.511	9.63	**13.68**	0.399	8.64	**10.10**	6.13	**7.31**	10.12	**11.65**	**9.37**	6.72
kihangaza	3.39	**11.86**	0.496	10.28	**9.60**	0.374	4.20	**10.39**	4.03	**10.02**	7.62	9.94	7.30	**7.03**
kihara	2.66	**13.25**	0.503	7.46	**13.87**	0.396	2.66	**10.66**	10.59	**10.25**	7.96	**10.66**	6.66	10.42
makonde	4.77	**17.81**	0.533	3.84	**15.71**	0.410	5.30	**11.95**	3.36	**12.47**	11.29	9.78	8.63	**7.92**
nyaturu	0.00	**14.38**	0.567	15.51	**13.08**	0.400	0.00	**9.09**	3.09	**7.96**	7.19	**14.30**	4.55	5.52
pare	5.67	**12.88**	0.543	11.48	**12.96**	0.410	**23.36**	9.76	9.36	**12.96**	9.28	12.22	**16.56**	11.16
sukuma	2.83	**12.31**	0.502	14.75	**12.25**	0.398	**17.29**	9.50	14.92	**10.61**	7.57	13.50	**13.39**	12.77
zigula	4.46	**19.87**	0.543	**24.45**	11.17	0.383	2.29	**12.10**	4.75	**9.35**	**12.17**	17.81	7.19	7.05
digo	4.07	**18.73**	0.558	13.95	**13.41**	0.414	3.46	**10.14**	2.84	**9.12**	11.40	**13.68**	6.80	5.98
ekegusii	3.67	**11.04**	0.465	11.04	**12.44**	0.392	3.67	**9.94**	9.10	**11.21**	7.36	11.74	6.80	**10.15**
gikuyu	3.09	**10.32**	0.490	11.98	**12.05**	0.397	3.09	**10.52**	3.09	**11.44**	6.70	**12.02**	6.80	7.27
kamba	5.60	**9.44**	0.496	7.12	**11.98**	0.398	4.07	**8.89**	9.29	**10.32**	7.52	9.55	6.48	**9.81**
luo	4.07	**9.52**	0.483	15.01	**12.20**	0.396	2.02	**9.66**	6.29	**10.35**	6.80	**13.61**	5.84	8.32
maasai	2.72	**8.41**	0.450	7.13	**11.47**	0.383	12.55	**10.21**	2.52	**11.13**	5.56	9.30	**11.38**	6.82
meru	3.67	**10.72**	0.470	8.81	**11.73**	0.394	3.09	**10.36**	0.00	**10.22**	7.20	**10.27**	6.73	5.11
nandi	0.00	**8.66**	0.435	9.43	**11.94**	0.385	0.00	**9.99**	10.13	**9.68**	4.33	**10.68**	5.00	9.91
nubian_2	4.77	**8.94**	0.456	4.02	**11.69**	0.382	2.63	**9.42**	2.63	**10.25**	6.85	**7.86**	6.02	6.44
nubian	5.52	**1.00**	0.392	**43.14**	12.73	0.420	6.57	**1.76**	5.52	**10.25**	3.26	**27.93**	4.16	7.89
nyala	0.00	**10.69**	0.484	7.86	**12.17**	0.399	0.00	**9.82**	**23.68**	8.97	5.34	10.02	4.91	**16.33**
olusamia	3.39	**9.66**	0.488	11.02	**11.62**	0.399	4.07	**9.32**	4.75	**7.30**	6.53	**11.32**	6.70	6.03
orma	2.83	**9.01**	0.411	12.55	**12.46**	0.378	7.81	**10.61**	**24.27**	9.66	5.92	12.51	9.21	**16.97**
rendille	4.03	**8.42**	0.408	10.68	**11.61**	0.378	3.75	**10.08**	3.58	**8.89**	6.22	**11.14**	6.91	6.23
samburu	0.00	**8.73**	0.408	2.02	**10.18**	0.343	0.00	**9.82**	2.83	**11.05**	4.37	6.10	4.91	**6.94**
teso	0.00	**9.92**	0.468	8.96	**13.65**	0.389	0.00	**10.06**	2.84	**11.44**	4.96	**11.31**	5.03	7.14
tugen	5.52	**9.21**	0.446	12.61	**12.09**	0.383	2.84	**9.00**	10.05	**9.27**	7.37	**12.35**	5.92	9.66
turkana	0.00	**8.67**	0.422	6.61	**11.44**	0.364	0.00	**9.88**	12.87	**11.42**	4.34	9.03	4.94	**12.15**
kwele	0.00	**7.23**	0.412	6.57	**12.65**	0.357	0.00	**7.74**	3.39	**9.65**	3.62	**9.61**	3.87	6.52
tetela	8.12	**9.97**	0.466	8.74	**11.44**	0.370	0.00	**8.96**	3.93	**9.40**	9.04	**10.09**	4.48	6.66
bangubangu	3.67	**10.99**	0.462	13.65	**13.38**	0.372	8.64	**10.86**	7.79	**13.02**	7.33	**13.52**	9.75	10.40
hema	6.57	**10.93**	0.492	4.92	**12.94**	0.403	0.00	**9.75**	0.00	**11.05**	8.75	**8.93**	4.87	5.52
hemba	6.57	**10.24**	0.449	13.04	**12.15**	0.360	11.99	**10.37**	7.66	**12.59**	8.41	**12.60**	11.18	10.13
holoholo	0.00	**10.01**	0.490	11.87	**11.42**	0.360	0.00	**9.39**	11.99	**11.40**	5.01	**11.65**	4.70	11.69
nande	2.84	**10.87**	0.490	9.63	**13.87**	0.404	4.20	**10.89**	2.84	**12.92**	6.85	**11.75**	7.54	7.88
taabwa	3.22	**12.13**	0.455	8.45	**11.97**	0.377	4.07	**10.26**	9.10	**7.97**	7.68	**10.21**	7.16	8.53
tshiluba	3.09	**10.79**	0.460	6.58	**12.70**	0.386	4.20	**10.94**	11.04	**11.67**	6.94	**9.64**	7.57	11.36
somali	0.00	**9.48**	0.482	**24.80**	12.49	0.399	0.00	**8.91**	3.74	**11.32**	4.74	**18.64**	4.45	7.53
borana	4.20	**8.49**	0.395	8.48	**12.33**	0.375	**14.92**	9.58	14.61	**12.80**	6.34	**10.40**	12.25	13.71
burji	2.45	**8.28**	0.402	8.51	**11.71**	0.385	3.75	**9.11**	5.58	**11.77**	5.36	**10.11**	6.43	8.67
Avg	**3.95**	**11.09**	**0.468**	**11.86**	**12.33**	**0.387**	**6.02**	**10.11**	**8.14**	**11.73**	**7.52**	**12.10**	**8.06**	**9.29**

Table 19: Zero-shot **literal and figurative** generation task results across 45 Afri-Proverb languages using **Llama-3.2-3B**. For each language direction we report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**, and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. The best score (BLEU or ChrF++) for each language and direction, including row-wise averages, is in **BOLD**. The bottom row gives column-wise averages.

lang	SmolLM3-3B (Zero-shot)										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	**15.97**	9.00	0.468	**19.00**	11.60	0.370	6.57	**10.94**	2.71	**12.43**	12.49	**15.30**	8.76	7.57
chiga	0.00	**10.80**	0.467	4.77	**12.24**	0.356	8.75	**10.09**	**20.56**	11.56	5.40	8.50	9.42	**16.06**
ganda	5.52	**7.52**	0.281	5.52	**9.69**	0.376	6.57	**10.69**	**19.58**	10.87	6.52	7.61	8.63	**15.22**
rufumbira	4.37	**9.70**	0.482	**26.98**	13.46	0.386	6.74	**11.02**	6.84	**12.18**	7.04	**20.22**	8.88	9.51
runyoro	5.52	**10.82**	0.476	5.52	**12.90**	0.388	2.45	**10.64**	**13.65**	13.19	8.17	**9.21**	6.55	13.42
soga	8.12	**11.63**	0.511	8.12	**10.23**	0.339	3.09	**11.45**	5.20	**12.48**	9.88	9.18	7.27	**8.84**
tooro	0.00	**11.46**	0.460	4.46	**13.97**	0.400	**12.44**	11.08	11.04	**13.45**	5.73	**9.22**	11.76	12.24
gweno	7.81	**14.08**	0.524	6.57	**12.85**	0.396	**10.68**	9.66	**8.97**	6.80	**10.95**	9.71	10.17	7.88
kihangaza	10.68	**12.63**	0.515	8.30	**9.80**	0.356	6.57	**9.93**	**13.49**	11.59	**11.65**	9.05	8.25	12.54
kihara	8.64	**13.79**	0.524	6.57	**10.66**	0.349	6.27	**11.08**	**13.13**	11.58	**11.22**	8.62	8.68	12.35
makonde	0.00	**19.09**	0.554	6.61	**13.12**	0.385	7.81	**10.92**	**16.19**	12.09	**9.54**	9.86	9.37	14.14
nyaturu	8.12	**14.91**	0.566	0.00	**12.00**	0.392	4.20	**9.81**	8.33	**10.11**	**11.52**	6.00	7.01	9.22
pare	4.46	**14.34**	0.570	6.26	**12.35**	0.407	**13.13**	9.15	**10.73**	10.14	**9.40**	9.31	11.14	10.44
sukuma	9.65	**12.50**	0.541	6.92	**12.35**	0.399	2.45	**8.18**	**19.88**	10.69	**11.08**	9.64	5.32	15.29
zigula	18.99	**21.68**	0.573	4.37	**12.02**	0.390	6.57	**12.25**	9.65	**12.99**	**20.34**	8.19	9.41	11.32
digo	3.74	**19.21**	0.570	5.52	**12.97**	0.410	4.93	**10.74**	**17.29**	12.64	11.48	9.24	7.84	**14.96**
ekegusii	8.12	**12.19**	0.466	**19.30**	12.10	0.404	10.68	**11.07**	6.47	**12.54**	10.15	**15.70**	10.88	9.50
gikuyu	5.52	**11.21**	0.489	10.83	**11.90**	0.393	2.63	**10.67**	9.00	**12.32**	8.37	11.36	6.65	**10.66**
kamba	4.46	**9.53**	0.495	**13.59**	12.68	0.401	**10.68**	10.09	**12.87**	11.98	7.00	**13.14**	10.38	12.42
luo	8.12	**9.69**	0.477	8.12	**12.17**	0.393	9.82	**10.77**	6.10	**11.96**	8.91	**10.14**	10.29	9.03
maasai	0.00	**8.68**	0.446	10.34	**11.64**	0.376	**15.11**	10.79	9.29	**12.05**	4.34	10.99	**12.95**	10.67
meru	4.77	**10.93**	0.469	**12.67**	12.17	0.395	8.64	**10.79**	9.46	**12.32**	7.85	**12.42**	9.72	10.89
nandi	0.00	**8.25**	0.419	6.57	**11.07**	0.347	8.12	**10.41**	8.70	**10.21**	4.12	8.82	**9.26**	9.46
nubian_2	**10.68**	9.65	0.452	7.50	**11.49**	0.363	6.74	**10.06**	8.49	**11.84**	10.17	**9.50**	8.40	10.17
nubian	4.07	**7.04**	0.441	9.27	**12.23**	0.426	8.13	**8.23**	11.48	**11.59**	5.56	**10.75**	8.18	11.54
nyala	0.00	**10.72**	0.478	7.50	**11.79**	0.363	8.12	**10.79**	**11.45**	10.57	5.36	**9.64**	9.46	11.01
olusamia	2.29	**9.77**	0.473	**15.97**	11.53	0.396	8.12	**10.82**	**9.10**	8.90	6.03	**13.75**	9.47	9.00
orma	**9.65**	8.83	0.397	9.65	**10.46**	0.327	**14.06**	10.37	**16.23**	12.21	9.24	10.06	12.21	**14.22**
rendille	**10.68**	8.49	0.405	10.68	**11.36**	0.356	9.65	**11.19**	**10.63**	10.18	9.58	11.02	**10.42**	10.40
samburu	0.00	**6.55**	0.394	6.29	**10.00**	0.341	0.00	**9.83**	6.57	**10.63**	3.28	8.15	4.92	**8.60**
teso	0.00	**10.62**	0.482	0.00	**11.74**	0.362	9.29	**11.28**	6.57	**11.84**	5.31	5.87	**10.28**	9.21
tugen	5.52	**8.84**	0.431	**11.15**	11.00	0.347	9.65	**10.38**	**14.54**	11.34	7.18	11.08	10.02	**12.94**
turkana	0.00	**8.10**	0.426	0.00	**12.12**	0.369	9.29	**9.79**	0.00	**10.60**	4.05	**6.06**	9.54	5.30
kwele	3.83	**7.62**	0.407	6.57	**10.60**	0.341	6.57	**10.13**	5.69	**11.51**	5.73	8.58	8.35	**8.60**
tetela	3.09	**9.55**	0.473	0.00	**11.15**	0.363	6.57	**9.88**	3.36	**10.71**	6.32	**5.57**	8.22	7.04
bangubangu	6.57	**11.12**	0.471	9.70	**12.44**	0.362	5.52	**11.76**	8.27	**13.60**	8.84	11.07	8.64	**10.93**
hema	0.00	**10.73**	0.481	7.35	**12.54**	0.385	10.68	**10.75**	6.26	**12.50**	5.37	**9.94**	10.71	9.38
hemba	**15.97**	10.26	0.469	11.48	**11.49**	0.368	**15.97**	10.17	**14.54**	11.51	**13.12**	11.49	13.07	13.03
holoholo	4.46	**10.34**	0.463	6.15	**12.25**	0.381	6.83	**9.97**	**21.59**	10.14	7.40	9.20	8.40	**15.87**
nande	5.67	**11.63**	0.478	5.67	**12.06**	0.367	7.27	**11.37**	6.34	**12.60**	8.65	8.87	9.32	**9.47**
taabwa	6.44	**11.68**	0.465	**12.67**	11.13	0.366	**14.54**	10.50	**10.99**	7.88	**9.06**	11.90	12.52	9.44
tshiluba	4.77	**11.31**	0.455	5.67	**12.18**	0.377	4.77	**10.78**	6.01	**12.26**	8.04	**8.92**	7.78	9.13
somali	0.00	**9.74**	0.503	0.00	**11.36**	0.391	**10.68**	10.42	5.99	**11.55**	4.87	5.68	10.55	**8.77**
borana	0.00	**8.25**	0.381	6.44	**11.19**	0.340	5.67	**10.53**	10.28	**11.82**	4.12	8.81	8.10	**11.05**
burji	**9.65**	8.04	0.403	7.77	**10.13**	0.341	10.25	**10.86**	**18.36**	11.75	8.84	8.95	10.55	**15.06**
Avg	**5.46**	**10.94**	**0.47**	**8.10**	**11.74**	**0.37**	**8.07**	**10.49**	**10.49**	**11.46**	**8.20**	**9.92**	**9.28**	**10.97**

Table 20: Zero-shot **literal and figurative** generation results across Afri-Proverb languages using **SmolLM3-3B**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is bolded; Row Avg is the mean of (**BLEU+ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages (all bold).

lang	Minimax-01 (Zero-shot)										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	**12.70**	11.86	0.49	**14.54**	13.39	0.39	4.11	**14.94**	4.70	**15.10**	12.28	**13.96**	9.53	9.90
chiga	3.46	**19.27**	0.57	4.83	**21.61**	0.46	7.48	**17.92**	4.61	**17.68**	11.36	**13.22**	12.70	11.14
ganda	**9.29**	8.27	0.37	7.35	**8.78**	0.35	6.90	**15.76**	3.31	**17.36**	8.78	8.06	**11.33**	10.33
rufumbira	5.87	**15.73**	0.55	**23.09**	20.03	0.47	9.61	**15.67**	7.44	**15.60**	10.80	**21.56**	12.64	11.52
runyoro	4.99	**17.75**	0.56	7.69	**21.81**	0.47	3.98	**15.29**	9.66	**15.59**	11.37	**14.75**	9.63	12.62
soga	4.46	**15.14**	0.51	10.60	**16.74**	0.42	5.60	**16.27**	4.27	**14.55**	9.80	**13.67**	10.94	9.41
tooro	1.96	**16.60**	0.54	4.92	**20.82**	0.46	4.96	**16.77**	**19.34**	17.69	9.28	12.87	10.87	**18.52**
gweno	**17.75**	17.15	0.52	5.20	**17.39**	0.42	4.78	**17.21**	3.10	**13.35**	**17.45**	11.29	11.00	8.23
kihangaza	9.65	**18.18**	0.51	13.13	**13.26**	0.41	9.10	**16.89**	4.35	**18.24**	**13.91**	13.20	13.00	11.30
kihara	11.48	**16.07**	0.52	11.04	**18.53**	0.42	5.46	**16.28**	**25.97**	16.71	13.77	14.79	10.87	**21.34**
makonde	1.34	**20.06**	0.56	9.00	**22.30**	0.45	4.22	**17.42**	2.18	**16.79**	10.70	**15.65**	10.82	9.48
nyaturu	8.12	**15.11**	0.58	11.48	**13.77**	0.41	5.51	**16.79**	6.61	**17.46**	11.61	**12.62**	11.15	12.03
pare	**25.41**	15.02	0.56	**17.97**	12.45	0.40	9.63	**16.73**	4.11	**17.25**	**20.21**	15.21	13.18	10.68
sukuma	11.34	**14.10**	0.54	15.51	**17.08**	0.44	6.57	**15.95**	3.49	**15.90**	12.72	**16.30**	11.26	9.70
zigula	15.97	**24.70**	0.60	11.48	**19.72**	0.47	13.26	**15.67**	4.79	**17.22**	**20.34**	15.60	14.47	11.00
digo	21.93	**22.47**	0.61	**19.30**	18.58	0.46	**25.34**	13.63	7.99	**13.81**	**22.20**	18.94	19.48	10.90
ekegusii	9.65	**17.11**	0.51	11.99	**18.88**	0.42	**16.19**	14.93	10.92	**15.29**	13.38	15.43	**15.56**	13.11
gikuyu	8.26	**14.12**	0.53	**29.62**	18.48	0.46	12.22	**13.80**	5.66	**15.40**	11.19	**24.05**	13.01	10.53
kamba	**34.57**	15.89	0.53	**28.19**	19.62	0.46	9.46	**13.66**	3.71	**13.56**	**25.23**	23.90	11.56	8.64
luo	8.64	**12.23**	0.53	8.26	**14.13**	0.42	**18.85**	14.90	7.52	**14.60**	10.44	11.19	**16.87**	11.06
maasai	0.00	**10.50**	0.50	10.68	**11.65**	0.40	7.27	**15.63**	5.42	**14.91**	5.25	11.17	**11.45**	10.16
meru	11.99	**12.87**	0.49	6.17	**12.72**	0.41	9.36	**16.00**	11.54	**15.15**	12.43	9.44	12.68	**13.34**
nandi	6.74	**9.28**	0.45	9.00	**11.54**	0.37	9.10	**15.46**	6.45	**17.61**	8.01	10.27	**12.28**	12.03
nubian_2	10.68	**11.68**	0.49	**14.54**	11.53	0.38	7.95	**15.24**	10.47	**14.40**	11.18	**13.03**	11.60	12.44
nubian	**12.70**	11.76	0.49	**17.29**	12.04	0.40	8.49	**15.31**	9.10	**14.52**	12.23	**14.66**	11.90	11.81
nyala	12.70	**15.05**	0.53	**19.30**	15.09	0.42	4.11	**14.71**	14.22	**16.06**	13.88	**17.20**	9.41	15.14
olusamia	**15.97**	11.90	0.52	7.81	**12.07**	0.41	16.23	**16.98**	4.41	**15.67**	13.94	9.94	**16.61**	10.04
orma	7.27	**11.25**	0.44	6.27	**11.25**	0.37	10.25	**15.41**	4.44	**14.97**	9.26	8.76	**12.83**	9.71
rendille	**12.55**	10.86	0.46	**15.62**	12.37	0.38	4.78	**15.07**	8.82	**16.08**	11.71	**13.99**	9.93	12.45
samburu	**12.70**	9.33	0.42	7.81	**8.86**	0.35	8.64	**15.79**	8.84	**15.40**	11.02	8.34	**12.21**	12.12
teso	8.04	**13.21**	0.46	7.66	**13.68**	0.37	6.84	**14.43**	12.61	**14.16**	10.62	10.67	10.64	**13.38**
tugen	9.82	**11.49**	0.45	4.42	**12.11**	0.37	6.05	**16.27**	6.92	**16.21**	10.66	8.27	11.16	**11.57**
turkana	5.67	**11.20**	0.48	5.30	**12.26**	0.38	7.32	**15.54**	7.47	**15.20**	8.44	8.78	**11.43**	11.34
kwele	6.57	**11.09**	0.47	4.78	**12.08**	0.35	2.23	**15.05**	9.14	**16.51**	8.83	8.43	8.64	**12.82**
tetela	5.93	**11.96**	0.50	11.04	**12.45**	0.38	5.30	**13.84**	3.93	**14.23**	8.95	**11.75**	9.57	9.08
bangubangu	12.22	**14.81**	0.47	10.99	**14.68**	0.38	16.26	**17.12**	4.33	**16.91**	13.52	12.83	**16.69**	10.62
hema	0.00	**18.63**	0.58	17.29	**21.33**	0.48	6.10	**15.10**	5.73	**14.61**	9.31	**19.31**	10.60	10.17
hemba	**15.97**	12.93	0.46	12.61	**12.65**	0.36	8.45	**15.09**	10.60	**16.27**	**14.45**	12.63	11.77	13.43
holoholo	**14.06**	12.68	0.51	10.68	**10.98**	0.37	5.19	**17.04**	5.99	**16.73**	**13.37**	10.83	11.12	11.36
nande	5.30	**15.64**	0.52	5.99	**16.42**	0.40	6.96	**14.29**	4.16	**15.26**	10.47	**11.21**	10.63	9.71
taabwa	6.27	**14.80**	0.49	8.74	**16.25**	0.41	4.18	**15.18**	8.70	**14.11**	10.54	**12.50**	9.68	11.40
tshiluba	8.91	**15.82**	0.48	**20.15**	17.13	0.42	4.50	**14.21**	5.16	**16.37**	12.37	**18.64**	9.36	10.76
somali	8.64	**20.88**	0.56	17.97	**28.64**	0.54	11.48	**17.37**	6.76	**18.47**	14.76	**23.30**	14.42	12.62
borana	6.61	**11.83**	0.44	**14.06**	13.06	0.37	5.83	**15.90**	7.92	**14.70**	9.22	**13.56**	10.87	11.31
burji	7.27	**10.70**	0.45	10.68	**11.23**	0.36	8.03	**15.17**	6.14	**14.61**	8.98	10.96	**11.60**	10.38
Avg	**10.03**	**14.42**	**0.51**	**12.04**	**15.36**	**0.41**	**8.31**	**15.64**	**7.40**	**15.74**	**12.23**	**13.70**	**11.98**	**11.57**

Table 21: Zero-shot **literal and figurative** generation results across Afri-Proverb languages using **Minimax-01**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is bolded; Row Avg is the mean of (**BLEU+ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages, all in **bold**.

lang	GLM-5 (Zero-shot)										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	**19.00**	15.24	0.52	14.06	**15.97**	0.42	7.81	**14.73**	6.76	**14.70**	**17.12**	15.01	11.27	10.73
chiga	6.57	**23.82**	0.60	5.51	**28.97**	0.52	6.92	**17.14**	12.38	**15.21**	15.19	**17.24**	12.03	13.80
ganda	6.27	**8.21**	0.38	**13.07**	8.96	0.37	15.85	**17.04**	**22.42**	15.07	7.24	11.01	16.45	**18.75**
rufumbira	15.62	**18.31**	0.58	10.83	**22.65**	0.51	4.27	**18.17**	13.45	**15.62**	16.97	**16.74**	11.22	14.54
runyoro	9.98	**22.27**	0.59	14.32	**25.72**	0.51	4.44	**16.37**	8.20	**17.26**	16.12	**20.02**	10.41	12.73
soga	**17.97**	15.74	0.53	**22.09**	19.81	0.45	11.04	**16.26**	6.27	**15.18**	16.85	**20.95**	13.65	10.73
tooro	11.34	**23.55**	0.58	12.87	**30.57**	0.52	9.84	**16.88**	8.70	**16.84**	17.45	**21.72**	13.36	12.77
gweno	6.57	**17.64**	0.54	**19.54**	18.81	0.44	5.70	**14.86**	**19.78**	9.52	12.10	**19.17**	10.28	14.65
kihangaza	4.46	**22.68**	0.58	11.02	**14.11**	0.44	9.82	**15.63**	8.26	**16.84**	**13.57**	12.56	12.73	12.55
kihara	15.62	**18.41**	0.55	11.99	**19.69**	0.45	3.32	**15.62**	7.81	**15.03**	17.02	**15.84**	9.47	11.42
makonde	3.74	**24.02**	0.58	13.97	**25.47**	0.48	6.57	**16.01**	6.42	**15.80**	13.88	**19.72**	11.29	11.11
nyaturu	9.65	**19.55**	0.57	13.65	**18.43**	0.44	10.55	**14.11**	5.70	**13.23**	**14.60**	16.04	12.33	9.46
pare	**25.41**	17.48	0.57	**24.27**	15.53	0.43	11.73	**13.56**	9.86	**12.64**	**21.44**	19.90	12.64	11.25
sukuma	9.65	**16.50**	0.56	14.15	**18.43**	0.44	9.58	**13.09**	9.55	**13.50**	13.08	**16.29**	11.33	11.52
zigula	**30.21**	29.11	0.64	19.30	**21.41**	0.47	13.89	**16.26**	7.81	**16.81**	**29.66**	20.35	15.07	12.31
digo	15.90	**17.20**	0.54	23.87	**24.24**	0.48	**20.56**	16.06	**19.30**	15.26	16.55	**24.05**	18.31	17.28
ekegusii	12.80	**21.70**	0.58	17.29	**21.60**	0.44	**18.39**	15.15	5.49	**14.97**	**17.25**	19.45	16.77	10.23
gikuyu	11.60	**20.10**	0.56	**31.36**	23.45	0.48	9.63	**16.31**	7.99	**16.21**	15.85	**27.41**	12.97	12.10
kamba	13.90	**19.40**	0.55	**23.79**	21.11	0.45	10.55	**14.49**	4.78	**13.97**	16.65	**22.45**	12.52	9.37
luo	9.80	**12.60**	0.49	17.75	**21.14**	0.45	**16.82**	15.50	**67.87**	15.26	11.20	19.44	16.16	**41.57**
maasai	7.60	**11.50**	0.46	**17.82**	14.99	0.41	4.93	**14.92**	8.16	**14.68**	9.55	**16.40**	9.93	11.42
meru	10.25	**15.44**	0.53	11.25	**17.03**	0.43	6.30	**15.61**	9.70	**14.86**	12.84	**14.14**	10.96	12.28
nandi	6.57	**12.01**	0.46	**17.02**	12.59	0.39	10.55	**13.56**	**14.32**	13.71	9.29	**14.81**	12.06	14.02
nubian_2	12.14	**12.76**	0.49	**13.49**	13.37	0.40	**16.09**	13.70	**19.08**	13.84	12.45	13.43	14.90	**16.46**
nubian	10.68	**12.15**	0.48	**16.52**	12.69	0.42	8.33	**14.56**	**27.30**	14.02	11.42	**14.60**	11.44	20.66
nyala	10.68	**15.95**	0.56	9.84	**17.28**	0.45	9.82	**15.02**	5.40	**13.11**	13.32	**13.56**	12.42	9.26
olusamia	12.70	**13.82**	0.54	11.48	**16.57**	0.44	13.13	**14.64**	**21.96**	10.78	13.26	14.03	13.88	**16.37**
orma	2.23	**13.72**	0.47	**14.11**	13.91	0.40	4.42	**14.67**	**21.36**	14.58	7.97	14.01	9.55	**17.97**
rendille	2.16	**12.51**	0.47	12.61	**12.91**	0.40	10.68	**14.99**	9.57	**13.68**	7.33	12.76	**12.84**	11.63
samburu	**9.65**	8.65	0.45	6.27	**9.93**	0.37	6.57	**12.76**	8.03	**13.69**	9.15	8.10	9.66	**10.86**
teso	**15.73**	14.24	0.49	**24.94**	15.75	0.40	10.70	**14.82**	7.31	**14.88**	14.98	**20.35**	12.76	11.09
tugen	8.26	**12.27**	0.48	**16.09**	13.82	0.40	9.86	**14.31**	12.45	**13.56**	10.26	**14.95**	12.09	13.01
turkana	8.64	**10.76**	0.49	0.00	**12.28**	0.39	2.72	**13.58**	3.96	**14.04**	9.70	6.14	8.15	**9.00**
kwele	6.30	**14.80**	0.49	5.51	**13.21**	0.37	11.34	**13.92**	6.72	**13.22**	**10.55**	9.36	12.63	9.97
tetela	7.90	**15.30**	0.50	9.86	**13.62**	0.40	5.30	**12.01**	8.45	**13.37**	**11.60**	11.74	8.66	10.91
bangubangu	8.70	**16.10**	0.51	**26.08**	15.38	0.40	7.46	**14.89**	**16.19**	15.83	12.40	**20.73**	11.17	16.01
hema	11.20	**18.60**	0.54	16.04	**25.43**	0.51	10.68	**15.19**	5.30	**15.40**	14.90	**20.73**	12.94	10.35
hemba	10.90	**17.90**	0.53	**30.74**	14.08	0.38	9.65	**14.40**	**32.01**	14.80	14.40	**22.41**	12.03	23.41
holoholo	9.60	**16.80**	0.52	12.55	**13.69**	0.39	7.81	**14.03**	**18.28**	12.91	13.20	13.12	10.92	**15.59**
nande	12.40	**19.20**	0.56	9.51	**17.98**	0.42	7.13	**15.22**	5.62	**15.38**	**15.80**	13.74	11.17	10.50
taabwa	9.40	**17.10**	0.52	11.25	**16.64**	0.42	4.79	**14.38**	8.49	**11.21**	**13.25**	13.94	9.59	9.85
tshiluba	11.80	**18.80**	0.55	**21.77**	16.97	0.42	4.26	**17.12**	7.37	**16.22**	15.30	**19.37**	10.69	11.79
somali	10.68	**27.87**	0.61	16.04	**34.85**	0.59	17.97	**18.10**	**66.87**	18.17	19.28	**25.44**	18.03	42.52
borana	2.55	**12.74**	0.45	**23.64**	15.55	0.41	5.30	**15.13**	**19.30**	14.89	7.65	**19.60**	10.22	17.10
burji	9.65	**11.50**	0.46	11.04	**12.83**	0.39	4.62	**14.88**	11.02	**14.44**	10.57	11.94	9.75	**12.73**
Avg	**10.29**	**17.10**	**0.53**	**13.17**	**16.57**	**0.43**	**9.19**	**14.93**	**13.26**	**14.54**	**13.69**	**14.87**	**12.06**	**13.90**

Table 22: Zero-shot **literal and figurative** generation results across Afri-Proverb languages using **GLM-5**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is bolded; Row Avg is the mean of (**BLEU+ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages, all in **bold**.

lang	Kimi K2.5										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	0.00	**13.10**	0.50	12.14	**14.75**	0.39	11.48	**14.72**	**27.09**	13.74	6.55	13.45	13.10	**20.41**
chiga	4.99	**22.27**	0.60	4.78	**27.31**	0.50	6.10	**17.91**	12.23	**16.09**	13.63	**16.05**	12.01	14.16
ganda	6.27	**8.21**	0.38	7.35	**8.96**	0.36	2.43	**17.00**	4.86	**15.50**	7.24	8.16	**9.72**	10.18
rufumbira	**19.08**	16.17	0.58	**30.13**	21.55	0.48	5.04	**17.01**	5.97	**16.13**	17.62	**25.84**	11.02	11.05
runyoro	5.67	**20.06**	0.57	5.67	**23.51**	0.47	2.85	**13.26**	4.62	**16.24**	12.87	**14.59**	8.05	10.43
soga	8.12	**12.81**	0.52	12.22	**19.02**	0.43	4.25	**15.50**	5.87	**14.50**	10.47	**15.62**	9.88	10.19
tooro	11.99	**20.85**	0.58	10.98	**27.02**	0.49	8.74	**16.09**	12.02	**17.62**	16.42	**19.00**	12.42	14.82
gweno	10.55	**19.28**	0.54	14.17	**18.01**	0.42	7.63	**16.22**	**17.18**	11.57	**14.91**	16.09	11.93	14.38
kihangaza	8.12	**22.72**	0.56	12.16	**14.78**	0.42	5.60	**17.24**	12.46	**16.66**	**15.42**	13.47	11.42	14.56
kihara	**30.74**	18.64	0.54	11.48	**20.42**	0.42	8.98	**15.87**	7.81	**15.02**	**24.69**	15.95	12.43	11.42
makonde	6.57	**22.70**	0.56	13.62	**22.40**	0.45	5.87	**17.37**	7.16	**14.99**	**14.64**	18.01	11.62	11.07
nyaturu	10.68	**15.88**	0.57	**17.29**	16.30	0.43	14.92	**16.24**	3.04	**14.03**	13.28	**16.79**	15.58	8.54
pare	**31.95**	15.74	0.59	12.70	**15.03**	0.42	8.05	**16.09**	12.45	**14.43**	**23.84**	13.87	12.07	13.44
sukuma	8.64	**14.76**	0.56	12.04	**17.21**	0.43	9.29	**15.17**	8.51	**14.51**	11.70	**14.62**	12.23	11.51
zigula	11.48	**24.34**	0.62	13.89	**18.33**	0.44	15.62	**16.17**	9.29	**16.76**	**17.91**	16.11	15.90	13.03
digo	**24.45**	23.76	0.62	**21.87**	20.26	0.46	6.44	**14.96**	**19.55**	14.50	**24.11**	21.07	10.70	17.03
ekegusii	7.81	**20.59**	0.55	**21.36**	20.12	0.42	10.55	**15.19**	8.79	**14.10**	**14.20**	20.74	12.87	11.45
gikuyu	14.54	**18.01**	0.59	**24.62**	23.03	0.48	7.99	**16.03**	8.31	**16.28**	16.28	**23.83**	12.01	12.29
kamba	8.64	**17.11**	0.56	17.61	**19.52**	0.43	6.41	**14.32**	8.69	**13.73**	12.88	**18.56**	10.37	11.21
luo	**15.62**	15.21	0.57	8.26	**21.16**	0.47	8.07	**14.99**	11.41	**14.82**	15.41	**14.71**	11.53	13.12
maasai	**19.00**	13.22	0.52	**21.02**	17.81	0.43	13.65	**14.73**	15.11	**15.90**	**16.11**	19.42	14.19	15.50
meru	8.74	**14.22**	0.53	**26.33**	16.48	0.42	8.67	**16.19**	9.63	**15.23**	11.48	**21.40**	12.43	12.43
nandi	7.81	**11.91**	0.47	**15.31**	13.28	0.37	6.92	**14.01**	10.06	**14.76**	9.86	**14.30**	10.47	12.41
nubian_2	10.68	**11.04**	0.49	12.70	**12.86**	0.38	**23.36**	15.00	13.33	**13.52**	10.86	12.78	**19.18**	13.43
nubian	0.00	**10.65**	0.48	**14.54**	12.51	0.40	**15.62**	14.80	13.85	**13.95**	5.33	13.53	**15.21**	13.90
nyala	8.64	**15.16**	0.54	10.59	**16.01**	0.42	8.13	**15.97**	7.99	**15.25**	11.90	**13.30**	12.05	11.62
olusamia	1.86	**12.25**	0.53	5.60	**15.97**	0.42	5.02	**12.61**	4.48	**12.70**	7.05	**10.79**	8.81	8.59
orma	9.65	**11.71**	0.47	12.61	**13.02**	0.38	11.73	**15.16**	13.65	**15.92**	10.68	12.81	13.45	**14.78**
rendille	5.30	**11.72**	0.47	9.29	**12.83**	0.39	10.08	**15.24**	**15.54**	14.52	8.51	11.06	12.66	**15.03**
samburu	0.00	**9.63**	0.46	**14.54**	10.24	0.36	9.30	**14.37**	6.27	**13.31**	4.82	**12.39**	11.84	9.79
teso	9.98	**14.12**	0.50	7.15	**15.66**	0.40	3.49	**14.72**	5.90	**13.70**	12.05	**11.40**	9.11	9.80
tugen	**9.43**	9.13	0.48	4.55	**14.09**	0.37	5.93	**15.92**	8.09	**15.40**	9.28	9.32	**10.92**	11.75
turkana	4.77	**11.05**	0.49	9.63	**13.22**	0.39	6.66	**14.35**	5.83	**13.93**	7.91	11.43	**10.51**	9.88
kwele	6.27	**10.98**	0.45	6.27	**10.98**	0.45	9.38	**15.44**	4.92	**14.89**	8.62	8.62	**12.41**	9.91
tetela	4.84	**11.29**	0.49	4.84	**11.29**	0.49	4.77	**12.59**	3.00	**13.07**	8.07	8.07	**8.68**	8.04
bangubangu	6.27	**14.70**	0.49	6.27	**14.70**	0.49	2.34	**15.48**	6.76	**15.14**	10.49	10.49	**8.91**	10.95
hema	15.97	**19.54**	0.60	15.97	**19.54**	0.60	5.97	**14.43**	7.44	**14.45**	**17.76**	17.76	10.20	10.94
hemba	9.65	**12.94**	0.48	9.65	**12.94**	0.48	4.42	**14.84**	7.86	**15.11**	11.29	11.29	9.63	**11.49**
holoholo	0.00	**1.98**	0.51	0.00	**1.98**	0.51	**25.41**	15.61	7.46	**14.83**	0.99	0.99	**20.51**	11.15
nande	4.62	**15.31**	0.51	4.62	**15.31**	0.51	**26.13**	15.00	3.72	**14.74**	9.97	9.97	**20.56**	9.23
taabwa	10.55	**15.08**	0.51	10.55	**15.08**	0.51	6.57	**15.60**	**13.03**	13.00	12.82	12.82	**11.09**	13.01
tshiluba	8.91	**16.41**	0.49	8.91	**16.41**	0.49	4.92	**15.21**	5.93	**15.95**	**12.66**	12.66	10.07	10.94
somalia	15.97	**26.07**	0.60	9.29	**30.43**	0.54	8.49	**17.93**	**19.64**	18.05	21.02	**19.86**	13.21	18.84
borana	**15.97**	12.63	0.46	**17.97**	15.10	0.38	7.13	**16.24**	3.49	**15.21**	14.30	**16.54**	**11.69**	9.35
burji	6.27	**9.38**	0.45	**15.14**	12.07	0.36	3.98	**11.00**	4.82	**14.30**	7.83	**13.61**	7.49	9.56
Avg	**9.93**	**15.21**	**0.53**	**12.39**	**16.63**	**0.44**	**8.76**	**15.33**	**9.45**	**14.85**	**12.57**	**14.51**	**12.05**	**12.15**

Table 23: Zero-shot **literal and figurative** generation results across Afri-Proverb languages using **Kimi K2.5**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is in bold; Row Avg is the mean of (**BLEU+ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages over reported languages only, all in **bold**.

lang	GPT-5 mini										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	**21.02**	15.53	0.52	14.06	**15.58**	0.41	15.51	**16.25**	9.00	**14.79**	**18.28**	14.82	15.88	11.90
chiga	6.27	**26.15**	0.61	8.27	**32.65**	0.55	4.62	**19.66**	9.63	**15.86**	16.21	**20.46**	12.14	12.74
ganda	6.74	**8.44**	0.38	**17.75**	9.16	0.38	9.58	**18.41**	9.86	**15.44**	7.59	13.46	**13.99**	12.65
rufumbira	**42.73**	21.62	0.60	20.26	**25.54**	0.54	10.70	**19.55**	3.96	**16.25**	**32.17**	22.90	15.13	10.11
runyoro	8.26	**22.90**	0.59	14.98	**26.69**	0.51	5.60	**17.31**	7.43	**16.70**	15.58	**20.84**	11.46	12.06
soga	13.13	**17.86**	0.54	**20.41**	19.20	0.45	4.27	**16.94**	6.58	**15.51**	15.50	**19.81**	10.61	11.05
tooro	11.34	**25.97**	0.60	**47.41**	33.47	0.55	4.55	**19.21**	11.02	**18.16**	18.66	**40.44**	11.88	14.59
gweno	9.86	**18.73**	0.55	12.19	**21.55**	0.46	7.12	**15.51**	**13.07**	9.86	14.30	**16.87**	11.32	11.47
kihangaza	2.91	**23.79**	0.59	7.46	**14.78**	0.44	4.62	**17.38**	9.10	**18.32**	**13.35**	11.12	11.00	13.71
kihara	19.64	**21.15**	0.56	**21.36**	21.16	0.45	8.27	**16.71**	6.10	**15.09**	**20.39**	21.26	12.49	10.59
makonde	7.27	**25.91**	0.59	9.86	**25.80**	0.49	11.25	**19.45**	10.39	**16.76**	16.59	**17.83**	15.35	13.58
nyaturu	7.81	**20.40**	0.59	10.68	**16.20**	0.45	9.10	**15.54**	12.22	**12.49**	**14.11**	13.44	12.32	12.36
pare	**23.64**	18.79	0.59	9.29	**15.79**	0.44	13.65	**15.49**	9.26	**12.37**	**21.22**	12.54	14.57	10.81
sukuma	10.68	**18.54**	0.58	14.32	**17.45**	0.44	**18.36**	14.99	5.62	**14.04**	**14.61**	15.89	16.68	9.83
zigula	14.06	**32.30**	0.64	19.64	**22.12**	0.49	13.89	**18.19**	7.81	**16.35**	**23.18**	20.88	16.04	12.08
digo	**48.55**	28.42	0.64	20.32	**24.24**	0.49	16.59	**16.95**	14.69	**15.68**	**38.48**	22.28	16.77	15.19
ekegusii	11.48	**25.36**	0.57	11.48	**25.56**	0.48	16.78	**17.24**	13.91	**17.24**	**18.42**	18.52	17.01	15.58
gikuyu	5.60	**22.95**	0.59	20.69	**25.29**	0.50	14.88	**17.74**	4.55	**17.26**	14.27	**22.99**	16.31	10.91
kamba	15.62	**19.64**	0.56	**23.79**	21.42	0.46	3.83	**15.31**	9.86	**14.64**	17.63	**22.61**	9.57	12.25
luo	10.68	**20.57**	0.57	21.83	**23.74**	0.48	11.73	**16.60**	10.70	**16.12**	15.62	**22.78**	14.17	13.41
maasai	**15.97**	12.74	0.51	12.70	**13.96**	0.42	10.05	**16.22**	5.60	**15.10**	14.36	13.33	**13.14**	10.35
meru	13.89	**17.93**	0.54	10.34	**16.84**	0.44	10.13	**16.87**	**17.19**	15.72	**15.91**	13.59	13.50	16.46
nandi	9.29	**12.63**	0.49	6.10	**12.24**	0.39	5.02	**15.97**	8.74	**13.65**	10.96	9.17	**10.49**	11.20
nubian_2	**17.29**	12.49	0.50	**16.04**	12.45	0.40	11.21	**15.82**	10.63	**14.63**	**14.89**	14.24	13.52	12.63
nubian	**15.62**	12.79	0.50	**21.93**	12.21	0.43	9.07	**15.75**	13.07	**14.45**	**14.21**	17.07	12.41	13.76
nyala	10.68	**16.75**	0.55	15.62	**17.10**	0.44	9.35	**17.08**	**17.40**	13.16	13.71	**16.36**	13.21	15.28
olusamia	**15.97**	14.02	0.53	6.74	**16.48**	0.44	5.87	**15.93**	**18.96**	10.80	14.99	11.61	10.90	**14.88**
orma	10.68	**13.04**	0.49	**14.92**	12.90	0.39	7.49	**15.87**	10.10	**15.33**	11.86	**13.91**	11.68	12.71
rendille	**14.06**	12.57	0.48	**13.04**	12.26	0.40	8.13	**16.49**	11.25	**13.61**	**13.31**	12.65	12.31	12.43
samburu	0.00	**9.77**	0.47	8.64	**8.74**	0.36	5.30	**14.99**	8.03	**13.58**	4.89	**8.69**	10.14	10.81
teso	5.60	**14.27**	0.49	14.67	**14.58**	0.39	5.87	**16.04**	4.83	**15.06**	9.94	**14.62**	10.96	9.95
tugen	5.87	**13.04**	0.49	**17.40**	13.77	0.39	12.98	**15.98**	10.83	**13.97**	9.45	**15.58**	14.48	12.40
turkana	**12.19**	11.35	0.48	5.52	**11.04**	0.39	3.98	**16.43**	3.72	**14.40**	11.77	8.28	**10.20**	9.06
kwele	5.93	**12.29**	0.46	9.58	**12.05**	0.36	11.04	**15.70**	8.49	**14.51**	9.11	10.81	**13.37**	11.50
tetela	11.34	**12.37**	0.49	11.50	**12.15**	0.39	10.88	**14.57**	9.43	**12.94**	11.86	11.82	**12.73**	11.19
bangubangu	7.27	**15.72**	0.49	13.65	**15.25**	0.40	5.35	**17.66**	6.61	**15.78**	11.50	14.45	**11.51**	11.20
hema	0.00	**23.51**	0.60	13.13	**25.50**	0.52	5.30	**16.55**	8.91	**14.89**	11.76	**19.32**	10.93	11.90
hemba	12.70	**13.02**	0.48	10.55	**12.83**	0.37	14.54	**15.39**	7.46	**15.39**	12.86	11.69	**14.97**	11.43
holoholo	**35.36**	13.69	0.52	**35.36**	12.69	0.38	12.98	**15.74**	12.30	**12.32**	**24.53**	24.03	14.36	12.31
nande	5.70	**17.56**	0.54	4.42	**17.30**	0.43	5.60	**16.17**	9.10	**15.68**	**11.63**	10.86	10.89	12.39
taabwa	7.81	**15.74**	0.49	11.45	**17.57**	0.43	9.57	**15.94**	8.45	**10.97**	11.78	**14.51**	12.75	9.71
tshiluba	9.43	**16.88**	0.50	15.81	**17.41**	0.42	8.03	**16.23**	14.54	**16.22**	13.16	**16.61**	12.13	15.38
somali	0.00	**34.17**	0.66	32.17	**42.08**	0.65	**27.05**	21.73	**28.87**	21.31	17.09	**37.13**	24.39	25.09
borana	0.00	**13.01**	0.47	14.06	**16.07**	0.42	7.46	**16.73**	14.54	**14.76**	6.50	15.07	**12.10**	14.65
burji	8.64	**11.72**	0.47	**15.51**	11.67	0.39	5.56	**15.75**	9.58	**14.75**	10.18	**13.59**	10.66	12.17
Avg	**10.24**	**17.17**	**0.53**	**13.12**	**17.46**	**0.44**	**9.12**	**16.01**	**11.42**	**14.91**	**13.70**	**15.29**	**12.57**	**13.17**

Table 24: Zero-shot **literal and figurative** generation results across Afri-Proverb using **GPT-5 mini**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is bolded; Row Avg is the mean of (**BLEU**+**ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages, all in **bold**.

lang	Grok-4.1-Fast										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	–	–	–	4.99	**12.67**	0.40	–	–	–	–	–	**8.83**	–	–
chiga	–	–	–	5.54	**18.22**	0.43	–	–	–	–	–	**11.88**	–	–
ganda	–	–	–	4.03	**8.54**	0.34	–	–	–	–	–	**6.29**	–	–
rufumbira	–	–	–	8.23	**15.77**	0.43	–	–	–	–	–	**12.00**	–	–
runyoro	–	–	–	6.45	**19.21**	0.43	–	–	–	–	–	**12.83**	–	–
soga	–	–	–	6.15	**15.98**	0.41	–	–	–	–	–	**11.07**	–	–
tooro	–	–	–	6.43	**18.56**	0.42	–	–	–	–	–	**12.50**	–	–
gweno	3.67	**16.36**	0.51	7.48	**15.77**	0.42	4.26	**14.18**	7.20	**9.12**	**10.02**	11.63	**9.22**	8.16
kihangaza	3.75	**16.79**	0.50	6.59	**11.80**	0.39	4.72	**14.16**	3.58	**15.69**	**10.27**	9.20	9.44	9.63
kihara	4.37	**15.36**	0.51	7.19	**16.75**	0.41	3.46	**13.80**	3.29	**13.99**	9.87	**11.97**	8.63	8.64
makonde	3.67	**21.64**	0.54	6.63	**21.63**	0.43	6.69	**16.13**	5.98	**14.55**	**12.66**	14.13	11.41	10.27
nyaturu	4.46	**16.25**	0.57	4.93	**13.32**	0.42	5.73	**14.10**	6.26	**11.99**	**10.36**	9.13	9.92	9.13
pare	3.75	**15.57**	0.57	4.93	**12.98**	0.42	3.42	**13.33**	4.62	**12.36**	**9.66**	8.96	8.38	8.49
sukuma	4.46	**14.72**	0.52	7.44	**15.59**	0.42	3.37	**12.72**	5.73	**12.72**	9.59	**11.52**	8.04	9.23
zigula	4.99	**23.86**	0.57	4.46	**14.86**	0.41	7.14	**16.21**	4.02	**14.22**	**14.43**	9.66	11.68	9.12
digo	7.47	**21.17**	0.58	8.33	**17.30**	0.45	11.98	**14.00**	6.61	**13.39**	**14.32**	12.82	12.99	10.00
ekegusii	4.11	**16.73**	0.49	6.29	**15.60**	0.41	3.46	**13.65**	8.51	**13.86**	**10.42**	10.95	8.55	11.19
gikuyu	3.93	**14.58**	0.52	6.44	**16.17**	0.44	6.76	**13.72**	3.83	**14.00**	**9.26**	11.31	10.24	8.92
kamba	4.25	**13.34**	0.50	14.40	**14.93**	0.41	3.93	**13.28**	5.90	**13.30**	8.79	**14.67**	8.61	9.60
luo	4.03	**11.09**	0.49	8.56	**13.27**	0.42	4.25	**13.59**	8.04	**13.70**	7.56	10.92	**8.92**	10.87
maasai	4.99	**9.61**	0.47	8.23	**12.37**	0.40	7.77	**14.10**	7.69	**13.79**	7.30	10.30	**10.94**	10.74
meru	3.60	**12.16**	0.49	3.82	**13.44**	0.41	8.23	**14.54**	6.37	**13.81**	7.88	8.63	**11.39**	10.09
nandi	3.09	**10.32**	0.44	11.12	**11.90**	0.38	3.22	**13.31**	8.77	**12.32**	6.71	11.51	**8.27**	10.55
nubian_2	4.20	**10.49**	0.46	**12.03**	11.38	0.38	**15.31**	13.21	11.15	**13.35**	7.34	11.71	**14.26**	12.25
nubian	3.74	**10.71**	0.47	4.20	**11.22**	0.40	10.52	**13.92**	12.93	**13.56**	7.22	7.71	**12.22**	13.24
nyala	4.46	**13.11**	0.51	4.07	**14.30**	0.42	3.66	**14.36**	7.01	**12.53**	8.79	9.19	**9.01**	9.77
olusamia	4.20	**11.90**	0.50	7.14	**13.00**	0.42	4.46	**13.44**	**12.45**	10.20	8.05	10.07	**8.95**	11.33
orma	4.46	**11.49**	0.44	8.23	**12.47**	0.38	4.07	**14.03**	6.15	**14.17**	7.98	10.35	9.05	**10.16**
rendille	4.46	**10.29**	0.43	3.72	**12.78**	0.38	6.59	**13.93**	**17.40**	12.25	7.38	8.25	10.26	**14.82**
samburu	0.00	**7.92**	0.41	7.43	**9.44**	0.35	3.09	**11.91**	3.42	**12.12**	3.96	8.44	7.50	**7.77**
teso	3.13	**12.73**	0.47	5.86	**15.09**	0.39	6.81	**14.13**	5.62	**13.84**	7.93	**10.48**	10.47	9.73
tugen	3.72	**10.64**	0.45	10.52	**13.12**	0.38	3.46	**14.08**	5.58	**13.35**	7.18	11.82	**8.77**	9.47
turkana	5.82	**9.46**	0.44	10.52	**12.42**	0.38	5.82	**13.09**	4.97	**12.64**	7.64	11.47	**9.46**	8.81
kwele	3.38	**9.59**	0.42	3.46	**10.60**	0.34	3.07	**13.25**	5.43	**13.86**	6.49	7.03	8.16	**9.65**
tetela	3.39	**10.74**	0.47	3.98	**12.21**	0.38	5.09	**11.80**	6.59	**11.63**	7.07	**8.10**	8.45	9.11
bangubangu	3.72	**13.10**	0.47	7.99	**14.39**	0.38	2.17	**14.62**	11.27	**14.17**	8.41	**11.19**	8.40	12.72
hema	4.93	**16.50**	0.53	4.37	**17.49**	0.43	4.03	**11.99**	3.39	**13.85**	10.72	**10.93**	8.01	8.62
hemba	4.20	**11.55**	0.46	6.75	**11.88**	0.37	7.91	**13.57**	8.51	**13.96**	7.88	9.32	10.74	**11.24**
holoholo	4.77	**11.03**	0.49	5.67	**11.56**	0.38	8.05	**13.70**	7.47	**12.52**	7.90	8.62	**10.88**	10.00
nande	3.03	**14.37**	0.48	6.25	**15.03**	0.40	5.03	**14.51**	5.82	**13.96**	8.70	**10.64**	9.77	9.89
taabwa	3.74	**13.15**	0.47	6.26	**13.72**	0.39	3.01	**14.30**	**14.00**	10.45	8.45	9.99	**8.66**	12.23
tshiluba	2.72	**13.26**	0.46	10.58	**14.98**	0.40	2.83	**14.18**	10.58	**14.42**	7.99	**12.78**	8.51	12.50
somali	–	–	–	–	–	–	–	–	–	–	–	–	–	–
borana	–	–	–	–	–	–	–	–	–	–	–	–	–	–
burji	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Avg	**4.87**	**13.44**	**0.48**	**7.78**	**14.53**	**0.40**	**6.45**	**14.34**	**8.30**	**13.39**	**9.16**	**11.15**	**10.40**	**10.85**

Table 25: Zero-Shot Generation results across Afri-Proverb using **Grok-4.1-Fast**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is in bold; Row Avg is the mean of (**BLEU**+**ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages over reported values only, all in **bold**.

lang	Gemini 3 Flash Preview										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	21.02	**32.67**	0.62	30.21	**30.00**	0.53	3.98	**20.02**	4.93	**17.09**	**26.84**	30.10	12.00	11.01
chiga	7.27	**32.17**	0.65	14.28	**38.56**	0.62	7.31	**21.48**	8.69	**18.75**	**19.72**	26.42	14.40	13.72
ganda	6.74	**8.50**	0.39	6.15	**9.25**	0.37	4.86	**20.31**	7.24	**18.87**	7.62	7.70	**12.58**	13.06
rufumbira	11.34	**25.04**	0.61	17.40	**26.77**	0.56	4.78	**19.10**	3.29	**16.06**	18.19	**22.09**	11.94	9.68
runyoro	7.13	**28.07**	0.63	14.50	**32.14**	0.57	3.61	**17.92**	9.03	**16.71**	17.60	**23.32**	10.76	12.87
soga	23.64	**25.15**	0.58	9.24	**31.00**	0.53	5.86	**17.66**	6.45	**15.85**	**24.39**	20.12	11.76	11.15
tooro	9.84	**33.63**	0.65	**44.29**	37.82	0.61	13.04	**20.50**	8.70	**19.93**	21.74	**41.06**	16.77	14.32
gweno	8.64	**28.71**	0.61	21.42	**29.95**	0.53	3.13	**19.50**	**15.88**	13.01	18.67	**25.69**	11.31	14.45
kihangaza	19.30	**37.02**	0.64	7.46	**16.99**	0.47	6.76	**19.99**	7.47	**19.27**	**28.16**	12.22	13.38	13.37
kihara	19.64	**25.11**	0.58	19.30	**23.89**	0.47	4.90	**18.23**	5.59	**18.11**	**22.38**	21.59	11.57	11.85
makonde	7.81	**32.53**	0.64	12.55	**33.47**	0.57	5.35	**20.51**	9.10	**20.70**	20.17	**23.01**	12.93	14.90
nyaturu	8.64	**27.85**	0.63	12.55	**22.47**	0.49	11.37	**18.35**	7.92	**16.20**	**18.25**	17.51	14.86	12.06
pare	23.64	**25.09**	0.61	9.82	**18.29**	0.45	12.04	**18.68**	7.69	**15.66**	**24.37**	14.06	15.36	11.68
sukuma	11.48	**30.65**	0.64	**30.79**	28.91	0.52	11.13	**17.91**	10.28	**17.00**	**21.06**	29.85	14.52	13.64
zigula	19.00	**44.01**	0.70	19.30	**28.29**	0.55	6.27	**19.05**	3.66	**18.02**	**31.50**	23.79	12.66	10.84
digo	34.93	**37.10**	0.69	32.65	**36.56**	0.56	2.72	**17.81**	7.13	**16.22**	**36.01**	34.61	10.27	11.68
ekegusii	16.52	**31.97**	0.62	10.60	**32.19**	0.52	15.99	**18.88**	8.70	**17.76**	**24.25**	21.40	17.44	13.23
gikuyu	14.92	**30.22**	0.63	28.80	**32.82**	0.56	7.40	**19.88**	7.72	**19.04**	**22.57**	30.81	13.64	13.38
kamba	**34.57**	31.35	0.61	29.86	**35.68**	0.55	8.84	**16.95**	6.76	**15.72**	32.96	**32.77**	12.90	11.24
luo	21.36	**33.00**	0.65	26.01	**35.73**	0.56	15.91	**19.19**	17.40	**18.35**	27.18	**30.87**	17.55	17.88
maasai	14.06	**25.44**	0.59	12.14	**33.05**	0.57	9.98	**19.87**	**29.50**	18.90	19.75	**22.59**	14.93	24.20
meru	11.73	**25.76**	0.58	19.14	**26.55**	0.50	11.63	**19.18**	9.84	**17.99**	18.75	**22.84**	15.41	13.92
nandi	8.64	**23.44**	0.57	10.25	**32.62**	0.55	7.43	**16.55**	3.98	**16.36**	16.04	**21.44**	11.99	10.17
nubian_2	11.48	**15.70**	0.53	**25.21**	17.62	0.43	**16.58**	16.32	11.54	**15.22**	13.59	**21.42**	16.45	13.38
nubian	**20.56**	16.08	0.52	**19.64**	16.75	0.44	9.10	**16.35**	**19.58**	15.25	**18.32**	18.20	12.72	17.41
nyala	11.48	**24.18**	0.61	13.04	**24.63**	0.51	4.11	**16.99**	7.54	**16.06**	17.83	**18.84**	10.55	11.80
olusamia	6.57	**17.87**	0.56	6.74	**22.88**	0.50	7.77	**17.57**	7.37	**14.08**	12.22	**14.81**	12.67	10.73
orma	9.65	**18.92**	0.53	10.92	**20.95**	0.47	5.78	**17.29**	6.19	**16.38**	14.28	**15.94**	11.54	11.29
rendille	9.65	**16.68**	0.51	14.54	**18.56**	0.45	3.98	**16.97**	9.24	**15.46**	13.17	**16.55**	10.48	12.35
samburu	**21.36**	11.23	0.48	**13.89**	11.87	0.40	5.30	**17.01**	5.73	**16.84**	**16.29**	12.88	11.16	11.29
teso	10.55	**20.67**	0.55	16.19	**24.15**	0.48	4.55	**17.08**	4.71	**16.92**	15.61	**20.17**	10.82	10.82
tugen	**22.96**	19.57	0.55	8.45	**24.74**	0.48	5.29	**18.10**	5.68	**16.44**	**21.27**	16.59	11.70	11.06
turkana	5.93	**14.83**	0.53	4.83	**19.31**	0.45	3.23	**16.91**	8.59	**17.43**	10.38	**12.07**	10.07	13.01
kwele	8.64	**12.17**	0.47	9.78	**13.09**	0.38	3.98	**15.80**	3.77	**17.09**	10.41	11.44	9.89	**10.43**
tetela	3.67	**19.66**	0.54	4.27	**19.36**	0.46	5.30	**15.77**	2.83	**15.72**	**11.66**	11.82	10.53	9.28
bangubangu	9.29	**17.01**	0.52	11.84	**18.42**	0.41	6.09	**17.15**	12.72	**16.98**	13.15	**15.13**	11.62	14.85
hema	8.64	**28.72**	0.63	18.80	**33.01**	0.60	4.46	**16.70**	6.42	**15.75**	18.68	**25.91**	10.58	11.09
hemba	10.68	**13.87**	0.48	10.25	**14.28**	0.38	4.83	**15.94**	6.20	**16.89**	12.28	**12.27**	10.39	11.54
holoholo	0.00	**14.73**	0.51	9.29	**14.40**	0.39	**17.97**	17.56	10.66	**15.76**	7.37	11.84	**17.77**	13.21
nande	12.61	**24.55**	0.58	10.63	**25.95**	0.50	3.40	**18.44**	4.11	**16.51**	18.58	**18.29**	10.92	10.31
taabwa	7.46	**19.50**	0.53	18.48	**20.69**	0.46	9.86	**17.68**	**18.05**	14.72	13.48	**19.59**	13.77	16.39
tshiluba	9.86	**23.58**	0.55	13.04	**24.60**	0.51	3.60	**20.77**	6.96	**20.08**	16.72	**18.82**	12.19	13.52
somali	12.22	**36.93**	0.66	23.74	**41.14**	0.65	11.37	**23.73**	7.96	**22.31**	24.58	**32.44**	17.55	15.14
borana	0.00	**17.22**	0.51	6.97	**16.79**	0.52	7.81	**18.89**	6.97	**16.79**	8.61	11.88	**13.35**	11.88
burji	**17.97**	14.99	0.49	3.49	**15.05**	0.49	4.24	**15.82**	3.49	**15.05**	**16.48**	9.27	10.03	9.27
Avg	**12.17**	**23.19**	**0.56**	**15.61**	**22.30**	**0.49**	**8.08**	**17.73**	**8.99**	**16.15**	**17.68**	**18.96**	**12.91**	**12.57**

Table 26: Zero-Shot Generation results across Afri-Proverb using **Gemini 3 Flash Preview**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is in bold; Row Avg is the mean of (**BLEU+ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages, all in **bold**.

lang	Qwen-3-4B-It (3-shot)										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	**15.97**	8.87	0.47	**19.00**	8.08	0.32	**19.00**	11.65	5.87	**15.04**	12.42	**13.54**	15.32	10.46
chiga	0.00	**10.93**	0.49	8.12	**8.60**	0.32	**14.92**	11.14	9.24	**13.99**	5.46	8.36	**13.03**	11.61
ganda	5.52	**7.56**	0.28	**11.04**	10.12	0.39	8.26	**11.09**	7.69	**14.75**	6.54	10.58	9.68	**11.22**
rufumbira	6.57	**9.25**	0.52	6.57	**9.37**	0.34	6.57	**11.23**	10.66	**12.96**	7.91	7.97	8.90	**11.81**
runyoro	5.52	**10.58**	0.51	5.52	**8.38**	0.31	6.27	**11.90**	8.74	**15.22**	8.05	6.95	9.09	**11.98**
soga	8.12	**12.06**	0.53	8.12	**8.13**	0.31	7.81	**11.66**	8.03	**14.50**	10.09	8.13	9.74	**11.27**
tooro	0.00	**10.88**	0.48	0.00	**8.08**	0.29	5.67	**11.37**	9.58	**15.21**	5.44	4.04	8.52	**12.40**
gweno	7.81	**14.62**	0.54	6.57	**8.61**	0.31	**11.34**	10.35	7.65	**10.57**	**11.22**	7.59	10.85	9.11
kihangaza	10.68	**13.00**	0.53	**11.99**	7.23	0.32	**13.13**	11.03	**18.48**	15.62	11.84	9.61	12.08	**17.05**
kihara	8.64	**13.73**	0.53	6.57	**8.91**	0.32	8.26	**11.19**	14.63	**14.89**	11.19	7.74	9.73	**14.76**
makonde	6.57	**19.74**	0.56	6.57	**11.10**	0.36	8.64	**10.62**	13.04	**13.78**	**13.16**	8.84	9.63	13.41
nyaturu	0.00	**14.88**	0.63	**11.04**	9.07	0.36	6.27	**10.20**	5.30	**12.72**	7.44	10.05	8.24	**9.01**
pare	10.68	**14.89**	0.63	**10.68**	7.71	0.33	**10.68**	9.80	7.86	**12.97**	**12.79**	9.20	10.24	10.42
sukuma	9.65	**12.99**	0.57	**11.99**	10.37	0.39	**13.13**	10.64	12.03	**14.20**	11.32	11.18	11.89	**13.12**
zigula	19.00	**23.75**	0.64	**25.41**	8.55	0.35	11.48	**11.90**	**17.29**	14.92	21.38	**16.98**	11.69	16.11
digo	7.27	**20.34**	0.64	5.52	**8.84**	0.35	**15.62**	11.47	9.57	**14.55**	**13.81**	7.18	13.55	12.06
ekegusii	8.12	**12.46**	0.47	**11.99**	11.75	0.36	**15.62**	11.22	**27.90**	15.90	10.29	11.87	13.42	**21.90**
gikuyu	5.52	**9.37**	0.52	5.52	**8.17**	0.33	7.46	**11.25**	12.45	**14.71**	7.45	6.85	9.36	**13.58**
kamba	7.27	**9.11**	0.51	**16.52**	11.20	0.38	**13.89**	11.15	9.86	**14.22**	8.19	**13.86**	12.52	12.04
luo	8.12	**8.21**	0.51	8.12	**9.47**	0.37	9.65	**11.81**	6.30	**15.47**	8.17	8.80	10.73	**10.89**
maasai	0.00	**7.47**	0.46	0.00	**8.44**	0.33	10.68	**11.66**	5.70	**13.97**	3.74	4.22	11.17	**9.84**
meru	4.77	**10.15**	0.49	0.00	**8.18**	0.33	**14.54**	11.81	9.88	**15.58**	7.46	4.09	13.18	**12.73**
nandi	6.57	**8.12**	0.42	6.57	**8.54**	0.29	**22.09**	11.02	9.26	**14.16**	7.35	7.56	**16.56**	11.71
nubian_2	0.00	**9.42**	0.46	0.00	**9.27**	0.32	**39.28**	11.05	14.54	**14.36**	4.71	4.64	**25.17**	14.45
nubian	**8.52**	8.09	0.44	**17.29**	12.01	0.40	**11.99**	11.14	9.86	**14.51**	8.31	**14.65**	11.57	12.19
nyala	0.00	**10.55**	0.50	**15.62**	9.89	0.33	10.55	**11.85**	5.51	**12.93**	5.28	**12.76**	11.20	9.22
olusamia	**15.97**	9.16	0.51	**15.97**	7.59	0.32	9.29	**13.02**	7.60	**11.97**	**12.57**	11.78	11.15	9.79
orma	**9.65**	8.72	0.40	**9.65**	8.15	0.28	8.64	**11.12**	7.81	**13.47**	**9.19**	8.90	9.88	10.64
rendille	**10.68**	7.79	0.40	**10.68**	8.14	0.31	**15.62**	11.90	4.67	**13.87**	9.24	9.41	**13.76**	9.27
samburu	0.00	**6.18**	0.40	0.00	**8.00**	0.34	10.68	**11.46**	5.30	**14.50**	3.09	4.00	11.07	**9.90**
teso	0.00	**10.22**	0.48	0.00	**9.29**	0.31	10.55	**10.81**	7.12	**15.15**	5.11	4.65	10.68	**11.14**
tugen	5.52	**8.58**	0.44	**11.48**	9.93	0.33	**15.62**	11.36	8.45	**13.73**	7.05	10.71	**13.49**	11.09
turkana	0.00	**7.83**	0.44	0.00	**9.75**	0.34	6.57	**11.14**	7.86	**14.22**	3.92	4.88	8.86	**11.04**
kwele	0.00	**7.25**	0.41	6.57	**8.82**	0.30	**14.26**	11.02	9.26	**14.51**	3.63	7.70	**12.64**	11.89
tetela	8.12	**9.57**	0.49	0.00	**8.74**	0.34	7.27	**10.63**	6.27	**11.37**	8.84	4.37	8.95	**8.82**
bangubangu	6.57	**11.32**	0.47	5.67	**9.74**	0.31	5.52	**11.34**	7.49	**16.58**	8.95	7.70	8.43	**12.03**
hema	0.00	**10.74**	0.52	0.00	**8.11**	0.32	9.65	**11.51**	11.34	**13.91**	5.37	4.06	10.58	**12.63**
hemba	**15.97**	10.23	0.47	**19.30**	9.63	0.33	7.27	**11.06**	9.00	**15.37**	13.10	**14.47**	9.16	12.19
holoholo	0.00	**9.25**	0.49	**13.13**	9.24	0.34	8.64	**10.25**	8.55	**12.83**	4.63	11.18	9.45	**10.69**
nande	5.67	**11.86**	0.49	5.52	**9.74**	0.31	6.74	**11.62**	7.12	**14.97**	8.76	7.63	9.18	**11.05**
taabwa	7.27	**11.98**	0.47	5.52	**9.61**	0.34	8.64	**10.99**	6.58	**11.91**	**9.63**	7.57	9.82	9.25
tshiluba	4.77	**11.30**	0.46	**11.99**	9.48	0.33	7.81	**11.29**	9.58	**15.37**	8.04	10.74	9.55	**12.48**
somali	0.00	**9.92**	0.51	0.00	**8.89**	0.34	9.29	**11.57**	8.70	**15.00**	4.96	4.45	10.43	**11.85**
borana	0.00	**8.24**	0.38	0.00	**8.52**	0.30	**12.70**	11.00	7.99	**14.62**	4.12	4.26	11.85	**11.31**
burji	**10.68**	7.81	0.40	**10.68**	7.78	0.31	7.81	**11.10**	10.25	**14.70**	9.25	9.23	9.46	**12.48**
Avg	**6.52**	**10.87**	**0.49**	**8.42**	**9.93**	**0.35**	**10.19**	**11.30**	**9.08**	**14.10**	**8.69**	**9.18**	**10.75**	**11.59**

Table 27: Few-shot (3-Shot) **literal and figurative** results across Afri-Proverb languages using **Qwen-3-4B-It (3-shot)**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is bolded; Row Avg is the mean of (**BLEU+ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages, all in **bold**.

lang	Llama-3.2-3B (3-shot)										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	3.22	**8.98**	0.45	9.82	**13.42**	0.39	7.46	**10.98**	**17.83**	11.44	6.10	11.62	9.22	**14.64**
chiga	0.00	**10.83**	0.49	11.38	**13.01**	0.40	**12.55**	10.23	**18.39**	10.42	5.41	12.20	11.39	**14.41**
ganda	5.52	**7.43**	0.31	2.58	**9.27**	0.35	4.77	**10.40**	**14.37**	9.56	6.48	5.93	7.58	**11.97**
rufumbira	4.77	**9.57**	0.48	0.00	**12.66**	0.39	3.75	**10.10**	0.00	**9.07**	7.17	**6.33**	6.93	4.54
runyoro	5.52	**10.66**	0.49	10.52	**13.96**	0.40	0.00	**11.19**	0.00	**12.63**	8.09	**12.24**	5.60	6.31
soga	3.39	**11.17**	0.50	**15.51**	12.88	0.38	3.09	**11.61**	12.51	**13.49**	7.28	**14.19**	7.35	13.00
tooro	0.00	**10.99**	0.47	9.84	**13.31**	0.40	9.29	**11.19**	6.83	**12.33**	5.50	**11.58**	10.24	9.58
gweno	4.03	**13.23**	0.51	9.63	**13.62**	0.39	8.64	**10.11**	6.13	**7.16**	8.63	**11.63**	9.38	6.65
kihangaza	3.09	**12.12**	0.50	9.82	**10.08**	0.38	4.77	**10.89**	4.20	**9.81**	7.61	**9.95**	7.83	7.01
kihara	3.13	**13.34**	0.50	10.89	**13.97**	0.39	2.66	**10.48**	7.77	**10.74**	8.24	**12.43**	6.57	9.26
makonde	4.77	**17.93**	0.52	7.86	**15.03**	0.41	0.00	**11.98**	**18.92**	11.82	11.35	**11.45**	5.99	15.37
nyaturu	0.00	**14.11**	0.56	**15.44**	11.75	0.39	**14.54**	9.36	**10.80**	8.08	7.06	**13.60**	11.95	9.44
pare	6.57	**13.13**	0.54	**12.98**	12.87	0.40	**10.68**	10.38	9.03	**12.51**	9.85	**12.93**	10.53	10.77
sukuma	6.74	**12.39**	0.50	10.05	**12.01**	0.40	**15.62**	9.83	**11.88**	10.42	9.56	**11.03**	12.73	11.15
zigula	2.91	**20.33**	0.55	5.87	**10.33**	0.38	4.93	**12.03**	4.75	**10.26**	**11.62**	8.10	8.48	7.51
digo	4.07	**18.75**	0.56	**13.95**	13.37	0.41	8.26	**10.30**	2.84	**9.45**	**11.41**	13.66	9.28	6.15
ekegusii	6.57	**10.99**	0.47	**14.63**	12.66	0.39	3.67	**10.62**	6.05	**11.17**	8.78	**13.65**	7.15	8.61
gikuyu	3.09	**10.62**	0.49	8.16	**12.00**	0.40	3.09	**10.81**	8.91	**11.52**	6.85	**10.08**	6.95	10.22
kamba	5.60	**9.57**	0.49	**14.99**	11.89	0.40	4.07	**9.56**	**12.98**	10.15	7.59	**13.44**	6.81	11.57
luo	3.39	**9.80**	0.48	**34.10**	12.27	0.40	6.27	**10.22**	**11.18**	10.98	6.60	**23.19**	8.25	11.08
maasai	2.45	**8.63**	0.45	4.62	**11.16**	0.38	4.77	**10.67**	6.74	**11.22**	5.54	**7.89**	7.72	8.98
meru	3.75	**10.94**	0.47	0.00	**11.98**	0.40	3.75	**10.70**	0.00	**10.59**	7.35	**5.99**	7.23	5.29
nandi	7.81	**9.05**	0.43	9.67	**11.91**	0.39	7.81	**10.46**	**17.03**	10.25	8.43	**10.79**	9.14	13.64
nubian_2	2.63	**8.77**	0.46	**12.01**	11.28	0.37	4.46	**9.95**	2.63	**10.08**	5.70	**11.65**	7.21	6.36
nubian	6.57	**1.03**	0.39	6.57	**12.94**	0.42	0.00	**1.79**	6.57	**9.42**	3.80	**9.75**	0.90	8.00
nyala	6.57	**10.70**	0.48	4.03	**12.23**	0.39	3.75	**10.64**	**15.81**	9.26	8.64	**8.13**	7.19	12.54
olusamia	3.13	**9.74**	0.48	**13.04**	10.83	0.39	4.20	**9.94**	4.75	**7.78**	6.43	**11.94**	7.07	6.27
orma	3.38	**8.96**	0.41	1.96	**12.00**	0.37	3.67	**10.38**	**26.76**	10.15	6.17	6.98	7.03	**18.46**
rendille	0.00	**8.45**	0.41	8.98	**11.91**	0.37	8.64	**10.56**	**12.87**	10.28	4.23	**10.45**	9.60	11.58
samburu	0.00	**8.79**	0.41	7.92	**10.14**	0.34	4.37	**10.97**	3.98	**11.58**	4.40	9.03	7.67	**7.78**
teso	0.00	**10.11**	0.47	**15.66**	13.20	0.39	0.00	**10.37**	6.74	**11.04**	5.05	**14.43**	5.19	8.89
tugen	2.84	**9.40**	0.45	10.34	**11.96**	0.38	0.00	**9.73**	2.84	**10.12**	6.12	**11.15**	4.87	6.48
turkana	0.00	**8.77**	0.41	8.91	**11.66**	0.37	0.00	**10.51**	0.00	**10.78**	4.38	**10.29**	5.26	5.39
kwele	0.00	**7.33**	0.41	6.57	**12.21**	0.35	0.00	**7.79**	3.39	**8.68**	3.67	**9.39**	3.90	6.04
tetela	2.15	**9.79**	0.47	**13.04**	11.64	0.37	0.00	**9.36**	6.27	**9.65**	5.97	**12.34**	4.68	7.96
bangubangu	3.09	**11.69**	0.47	12.61	**13.47**	0.37	**14.54**	11.29	0.00	**12.29**	7.39	**13.04**	12.92	6.14
hema	0.00	**11.15**	0.49	9.82	**12.33**	0.40	0.00	**10.23**	7.44	**10.52**	5.58	**11.07**	5.12	8.98
hemba	6.57	**10.47**	0.45	9.82	**11.64**	0.36	8.64	**10.43**	2.83	**11.56**	8.52	**10.73**	9.54	7.20
holoholo	**10.68**	10.51	0.49	**16.52**	11.29	0.36	7.81	**10.24**	6.05	**11.08**	10.60	**13.91**	9.03	8.57
nande	5.67	**11.34**	0.49	5.58	**13.21**	0.39	**12.22**	10.74	9.48	**11.59**	8.51	**9.40**	11.48	10.54
taabwa	3.46	**11.63**	0.46	9.38	**11.97**	0.37	**21.93**	10.81	**11.67**	7.44	7.55	10.68	**16.37**	9.56
tshiluba	4.20	**10.76**	0.45	8.85	**12.58**	0.38	8.26	**11.24**	6.58	**11.89**	7.48	**10.72**	9.75	9.24
somali	0.00	**9.50**	0.48	**25.89**	11.75	0.39	3.09	**9.09**	2.84	**10.29**	4.75	**18.82**	6.09	6.57
borana	4.20	**8.81**	0.39	**13.33**	12.48	0.37	**14.92**	10.63	6.44	**12.53**	6.51	**12.91**	12.78	9.49
burji	2.45	**8.26**	0.41	10.34	**11.39**	0.38	7.27	**9.34**	3.93	**11.15**	5.36	**10.87**	8.31	7.54
Avg	**3.30**	**10.79**	**0.47**	**9.27**	**11.63**	**0.38**	**6.28**	**10.78**	**8.49**	**11.97**	**7.05**	**10.45**	**8.53**	**10.23**

Table 28: Few-shot (3-shot) **literal and figurative** results across Afri-Proverb using **Llama-3.2-3B (3-shot)**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is bolded; Row Avg is the mean of (**BLEU+ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages, all in **bold**.

lang	SmolLM3-3B (3-shot)										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	**15.97**	9.08	0.46	**19.00**	11.67	0.38	6.57	**10.60**	2.71	**12.40**	12.53	**15.34**	8.59	7.56
chiga	8.12	**10.66**	0.46	7.50	**12.01**	0.36	8.64	**9.93**	**21.59**	11.22	9.39	9.75	9.29	**16.41**
ganda	5.52	**7.48**	0.29	5.52	**9.93**	0.38	6.57	**10.96**	8.64	**11.24**	6.50	7.73	8.77	**9.94**
rufumbira	3.67	**9.55**	0.48	12.23	**13.75**	0.39	7.81	**10.78**	9.20	**11.83**	6.61	**12.99**	9.30	10.52
runyoro	4.02	**10.84**	0.47	5.52	**12.73**	0.38	6.57	**10.59**	4.62	**13.13**	7.43	9.13	8.58	**8.88**
soga	**15.97**	11.35	0.51	4.07	**9.92**	0.34	6.27	**11.25**	8.07	**12.71**	**13.66**	6.99	8.76	10.39
tooro	0.00	**11.27**	0.47	8.45	**14.08**	0.39	7.81	**11.32**	10.02	**13.45**	5.63	**11.27**	9.57	11.74
gweno	7.81	**14.13**	0.52	6.57	**12.50**	0.39	**14.06**	9.56	**8.97**	7.00	**10.97**	9.54	11.81	7.99
kihangaza	10.68	**12.79**	0.52	**14.26**	9.60	0.35	7.81	**10.36**	**5.87**	11.85	11.74	**11.93**	9.09	8.86
kihara	8.64	**13.58**	0.52	10.10	**10.24**	0.35	7.27	**10.98**	5.20	**12.09**	**11.11**	10.17	9.12	8.64
makonde	0.00	**19.39**	0.56	6.57	**12.92**	0.39	8.12	**11.38**	11.18	**11.99**	**9.70**	9.75	9.75	11.59
nyaturu	0.00	**14.84**	0.58	0.00	**11.87**	0.38	10.68	**9.88**	9.58	**10.43**	**7.42**	5.94	10.28	10.01
pare	10.68	**14.08**	0.56	6.44	**11.93**	0.40	6.74	**8.96**	**9.84**	9.77	**12.38**	9.19	7.85	9.81
sukuma	9.65	**12.37**	0.54	8.12	**12.28**	0.40	3.13	**8.01**	**16.04**	10.74	11.01	10.20	5.57	**13.39**
zigula	19.00	**22.92**	0.60	4.37	**11.62**	0.39	6.74	**12.03**	**14.54**	12.42	**20.96**	8.00	9.38	13.48
digo	3.74	**18.79**	0.57	5.52	**12.71**	0.41	4.93	**10.74**	**17.29**	12.65	**11.27**	9.12	7.84	14.97
ekegusii	8.12	**12.16**	0.47	8.12	**12.13**	0.41	**12.14**	11.30	7.14	**12.24**	10.14	10.12	11.72	**9.69**
gikuyu	0.00	**11.08**	0.48	**13.17**	12.01	0.39	8.64	**10.68**	12.04	**12.33**	5.54	**12.59**	9.66	12.19
kamba	7.27	**9.60**	0.50	**13.72**	12.45	0.40	5.52	**10.19**	10.25	**12.11**	8.44	**13.08**	7.86	11.18
luo	8.12	**9.67**	0.47	6.84	**12.05**	0.39	4.77	**10.51**	7.06	**12.20**	8.90	**9.45**	7.64	9.63
maasai	0.00	**8.46**	0.45	6.47	**11.48**	0.37	6.74	**10.95**	**18.36**	12.28	4.23	**8.98**	8.84	15.32
meru	3.74	**11.00**	0.47	**12.46**	11.92	0.39	6.27	**10.59**	9.57	**12.37**	7.37	**12.19**	8.43	10.97
nandi	0.00	**8.14**	0.42	6.57	**10.90**	0.35	**19.00**	10.07	**11.70**	9.94	4.07	8.74	**14.54**	10.82
nubian_2	**10.68**	9.72	0.45	7.50	**11.34**	0.36	5.67	**10.43**	**16.52**	12.04	10.20	9.42	8.05	**14.28**
nubian	4.62	**6.56**	0.44	8.03	**12.45**	0.43	7.81	**8.32**	**17.16**	11.88	5.59	10.24	8.07	**14.52**
nyala	10.68	**10.73**	0.48	7.50	**11.79**	0.36	0.00	**10.96**	**17.40**	10.15	10.70	9.65	5.48	**13.78**
olusamia	**15.97**	9.82	0.48	7.35	**11.73**	0.40	7.81	**10.35**	**9.40**	8.60	**12.90**	9.54	9.08	9.00
orma	**9.65**	8.94	0.40	9.65	**10.69**	0.33	5.67	**10.92**	**12.61**	11.62	9.30	10.17	8.30	**12.12**
rendille	4.62	**8.38**	0.40	10.68	**11.13**	0.36	**12.70**	11.37	10.28	**10.67**	6.50	10.91	**12.03**	10.48
samburu	0.00	**6.60**	0.40	6.29	**9.56**	0.34	7.81	**9.80**	**10.55**	10.39	3.30	7.93	8.81	**10.47**
teso	3.74	**10.68**	0.48	0.00	**11.95**	0.36	0.00	**10.89**	3.00	**12.32**	7.21	5.98	5.45	**7.66**
tugen	5.52	**8.82**	0.43	6.29	**11.00**	0.34	7.81	**10.38**	**13.67**	11.36	7.17	8.65	9.10	**12.52**
turkana	0.00	**7.96**	0.43	0.00	**12.03**	0.37	7.81	**9.94**	0.00	**10.83**	3.98	**6.02**	8.88	5.42
kwele	3.83	**7.80**	0.41	6.57	**10.61**	0.33	6.57	**10.38**	5.69	**11.74**	5.82	8.59	8.48	**8.72**
tetela	3.39	**9.77**	0.47	0.00	**10.99**	0.37	6.27	**9.97**	6.57	**9.81**	6.58	**5.50**	8.12	8.19
bangubangu	6.57	**11.21**	0.47	11.50	**12.71**	0.37	6.57	**11.56**	12.19	**12.53**	8.89	**12.11**	9.07	12.36
hema	0.00	**10.71**	0.48	7.35	**12.31**	0.37	5.52	**10.26**	7.50	**12.50**	5.36	**9.83**	7.89	10.00
hemba	**15.97**	10.25	0.46	4.79	**10.70**	0.36	9.65	**9.88**	7.76	**11.32**	**13.11**	7.75	9.77	9.54
holoholo	0.00	**10.72**	0.47	6.07	**11.65**	0.37	**10.68**	9.77	**17.29**	10.11	5.36	8.86	10.23	**13.70**
nande	5.67	**11.65**	0.49	5.67	**12.86**	0.38	5.30	**11.72**	8.69	**12.93**	8.66	**9.27**	8.51	10.81
taabwa	7.27	**11.58**	0.47	3.74	**11.69**	0.38	**19.00**	10.24	**9.31**	8.02	9.43	7.72	**14.62**	8.67
tshiluba	5.06	**11.08**	0.45	5.67	**10.94**	0.36	**11.04**	10.88	**13.65**	12.51	8.07	8.31	10.96	**13.08**
somali	0.00	**9.78**	0.50	0.00	**11.64**	0.39	**10.68**	10.38	6.11	**11.54**	4.89	**5.82**	10.53	8.83
borana	0.00	**8.27**	0.38	6.44	**10.74**	0.34	5.67	**10.36**	10.28	**11.96**	4.14	8.59	8.01	**11.12**
burji	4.03	**8.46**	0.40	8.45	**10.28**	0.35	6.57	**10.23**	9.86	**11.49**	6.25	9.37	8.40	**10.68**
Avg	**6.09**	**10.96**	**0.48**	**8.21**	**11.47**	**0.38**	**8.13**	**10.64**	**10.21**	**11.63**	**8.52**	**9.84**	**9.38**	**10.92**

Table 29: Few-shot (3-shot) **literal and figurative** results across Afri-Proverb using **SmolLM3-3B**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is bolded; Row Avg is the mean of (**BLEU+ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages, all in **bold**.

lang	Gemma-3-4B-It (3-shot)										Row Avg (BLEU + ChrF++)			
	swa_lit			eng_lit			swa_fig		eng_fig		swa_lit	eng_lit	swa_fig	eng_fig
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++				
alur	**19.00**	9.20	0.48	**20.41**	12.46	0.38	**19.00**	11.29	11.04	**11.90**	14.10	**16.44**	15.15	11.47
chiga	0.00	**11.77**	0.50	10.80	**16.26**	0.41	3.39	**10.42**	5.20	**12.11**	5.88	**13.53**	6.91	8.66
ganda	5.52	**7.72**	0.29	7.35	**8.97**	0.33	3.09	**10.07**	6.27	**10.83**	6.62	8.16	6.58	**8.55**
rufumbira	6.57	**10.01**	0.51	9.58	**15.64**	0.42	2.15	**13.12**	8.85	**11.52**	8.29	**12.61**	7.63	10.18
runyoro	5.52	**11.23**	0.50	**24.34**	15.55	0.41	6.27	**13.43**	9.86	**12.68**	8.38	**19.95**	9.85	11.27
soga	**15.97**	12.35	0.53	7.46	**12.65**	0.38	2.63	**9.88**	5.52	**12.64**	**14.16**	10.05	6.26	9.08
tooro	0.00	**11.59**	0.47	7.99	**18.22**	0.40	6.74	**13.00**	11.84	**14.14**	5.79	**13.11**	9.87	12.99
gweno	7.81	**14.54**	0.53	6.57	**14.66**	0.40	2.29	**10.29**	4.17	**6.80**	11.18	**10.61**	6.29	5.49
kihangaza	10.68	**13.12**	0.51	9.29	**9.97**	0.38	8.64	**8.99**	**19.08**	12.96	11.90	9.63	8.82	**16.02**
kihara	8.64	**13.90**	0.51	12.98	**15.60**	0.39	4.99	**11.44**	7.55	**11.67**	11.27	**14.29**	8.22	9.61
makonde	6.57	**19.52**	0.55	11.39	**16.35**	0.40	6.10	**11.44**	7.81	**13.02**	**13.04**	13.87	8.77	10.42
nyaturu	0.00	**14.91**	0.63	7.82	**12.28**	0.39	5.52	**8.32**	5.93	**9.67**	**7.46**	10.05	6.92	7.80
pare	**15.97**	14.56	0.61	8.26	**12.44**	0.39	5.52	**8.12**	**15.62**	9.83	15.27	10.35	6.82	**12.72**
sukuma	9.65	**13.08**	0.56	9.24	**13.36**	0.40	**14.06**	9.55	**10.92**	10.82	11.37	**11.30**	11.81	10.87
zigula	19.00	**21.13**	0.60	**19.30**	14.29	0.40	4.77	**9.75**	**15.62**	12.48	**20.06**	16.79	7.26	14.05
digo	7.27	**19.96**	0.63	5.52	**15.69**	0.42	2.72	**10.99**	10.55	**13.97**	**13.62**	10.61	6.85	12.26
ekegusii	8.12	**12.36**	0.46	6.83	**15.01**	0.40	**32.47**	12.88	**22.59**	13.29	10.24	10.92	**22.68**	17.94
gikuyu	5.52	**9.45**	0.52	**14.17**	13.41	0.39	**34.57**	12.41	10.89	**13.35**	7.49	13.79	**23.49**	12.12
kamba	7.27	**10.29**	0.51	11.04	**14.16**	0.40	**25.41**	12.85	**15.62**	13.61	8.78	12.60	**19.13**	14.62
luo	**14.06**	9.42	0.51	**26.76**	12.19	0.40	**24.27**	11.57	**37.99**	13.64	11.74	19.48	17.92	**25.82**
maasai	**15.97**	8.71	0.47	**17.29**	13.03	0.39	9.65	**12.38**	**17.97**	13.06	12.34	**15.16**	11.02	15.52
meru	4.77	**10.54**	0.49	**25.73**	12.66	0.39	0.00	**11.49**	**17.97**	13.30	7.66	**19.19**	5.75	15.64
nandi	6.57	**8.28**	0.43	**13.38**	10.69	0.35	**15.97**	9.77	2.63	**8.76**	7.42	**12.04**	12.87	5.70
nubian_2	0.00	**10.24**	0.46	**16.04**	12.86	0.39	**19.30**	10.85	**30.21**	10.95	5.12	14.45	15.08	**20.58**
nubian	0.00	**2.69**	0.38	**23.36**	11.96	0.41	**20.41**	10.58	**37.99**	9.98	1.34	17.66	15.50	**23.99**
nyala	0.00	**10.74**	0.50	**16.32**	13.02	0.38	4.77	**9.45**	4.88	**9.49**	5.37	**14.67**	7.11	7.19
olusamia	**15.97**	9.56	0.50	**16.78**	12.98	0.40	**24.88**	12.22	1.12	**8.11**	12.77	14.88	**18.55**	4.61
orma	**9.65**	9.12	0.39	7.13	**12.13**	0.37	8.12	**9.85**	6.74	**11.98**	9.39	**9.63**	8.99	9.36
rendille	**10.68**	8.08	0.41	**10.68**	10.62	0.36	8.12	**10.66**	**10.05**	9.65	9.38	**10.65**	9.39	9.85
samburu	0.00	**7.44**	0.42	6.27	**9.05**	0.34	**23.64**	9.89	7.81	**11.50**	3.72	7.66	**16.76**	9.66
teso	4.77	**10.91**	0.48	9.11	**13.52**	0.36	0.00	**10.72**	3.23	**14.05**	7.84	11.32	5.36	**8.64**
tugen	5.52	**9.04**	0.44	9.86	**10.90**	0.35	0.00	**10.51**	**18.36**	9.74	7.28	10.38	5.25	**14.05**
turkana	0.00	**8.69**	0.45	0.00	**11.28**	0.37	**28.12**	9.86	**21.11**	11.74	4.35	5.64	**18.99**	16.42
kwele	0.00	**7.63**	0.41	6.92	**11.73**	0.34	**14.54**	10.31	**12.55**	10.64	3.81	9.33	**12.42**	11.59
tetela	8.12	**10.00**	0.49	**14.26**	12.34	0.37	6.27	**9.11**	4.99	**10.84**	9.06	**13.30**	7.69	7.92
bangubangu	9.65	**11.61**	0.48	10.68	**12.16**	0.36	2.45	**10.49**	**19.64**	14.46	10.63	11.42	6.47	**17.05**
hema	0.00	**11.15**	0.51	12.19	**15.23**	0.42	0.00	**8.29**	8.64	**12.60**	5.58	**13.71**	4.15	10.62
hemba	**15.97**	10.61	0.47	**36.13**	12.20	0.35	4.20	**9.95**	**23.36**	12.25	13.29	**24.17**	7.08	17.81
holoholo	0.00	**9.48**	0.49	**11.99**	11.77	0.37	4.77	**8.53**	**11.60**	9.82	4.74	**11.88**	6.65	10.71
nande	5.67	**12.05**	0.49	7.86	**13.33**	0.38	7.81	**11.42**	**19.00**	12.91	8.86	10.59	9.62	**15.95**
taabwa	7.27	**12.16**	0.47	5.52	**13.84**	0.38	0.00	**9.87**	7.86	**8.50**	9.71	**9.68**	4.94	8.18
tshiluba	4.99	**11.61**	0.45	13.04	**14.10**	0.38	2.45	**10.94**	7.13	**12.93**	8.30	**13.57**	6.70	10.03
somalia	6.57	**10.40**	0.51	12.55	**16.08**	0.41	11.87	**12.19**	**31.95**	12.73	8.49	14.32	12.03	**22.34**
borana	**19.00**	8.83	0.39	**16.37**	12.73	0.36	**23.64**	11.44	10.57	**12.47**	13.91	**14.55**	17.54	11.52
burji	**10.68**	8.03	0.40	7.77	**10.96**	0.37	4.46	**9.60**	9.29	**11.32**	9.36	**9.37**	7.03	10.31
Avg	**7.44**	**11.06**	**0.48**	**12.54**	**13.12**	**0.38**	**10.22**	**10.67**	**13.10**	**11.66**	**9.25**	**12.83**	**10.45**	**12.38**

Table 30: Few-shot (3-shot) **literal and figurative** generation results across Afri-Proverb using **Gemma-3-4B-It**. We report **BLEU**, **ChrF++**, and **COMET** for **swa_lit** and **eng_lit**; and **BLEU** and **ChrF++** for **swa_fig** and **eng_fig**. For each language and direction, the better of **BLEU** vs. **ChrF++** is bolded; Row Avg is the mean of (**BLEU+ChrF++**)/2 per direction (COMET excluded), with the best Row Avg per language bolded. The bottom row shows column-wise averages, all in **bold**.

ISO code	Name	Family	Region	# Speakers	Vitality	Digital Support
kam	Kamba	Niger-Congo	Kenya & Tanzania	4.6M	Institutional	Ascending
dig	Digo	Niger-Congo	Kenya & Tanzania	580K	Institutional	Ascending
mas	Maasai	Nilo-Saharan	Kenya & Tanzania	1.5M	Stable	Ascending
kcn	Nubi	Creole	Kenya & Uganda	50K	Stable	Emerging
luo	Dholuo	Nilo-Saharan	Kenya & Tanzania	4.2M	Stable	Ascending
kik	Gikuyu	Niger-Congo	Kenya	6.6M	Stable	Ascending
guz	Ekegusii	Niger-Congo	Kenya & Tanzania	2.7M	Stable	Ascending
gax	Borana	Afro-Asiatic	Kenya, Ethiopia & Somalia	9.6M	Stable	Ascending
som	Somali	Afro-Asiatic	Kenya & Somalia	24M	Institutional	Vital
gwe	Gweno	Niger-Congo	Tanzania	2.5K	Endangered	Emerging
han	Hangaza	Niger-Congo	Tanzania	155K	Stable	Emerging
reg	Kara	Niger-Congo	Tanzania	100K	Endangered	Emerging
rim	Nyaturu	Niger-Congo	Tanzania	1.3M	Stable	Ascending
asa	Pare	Niger-Congo	Tanzania	500K	Stable	Ascending
ziw	Zigula	Niger-Congo	Tanzania	480K	Endangered	Ascending
alz	Alur	Nilo-Saharan	DRC & Tanzania	1.7M	Institutional	Ascending
cgg	Chiga	Niger-Congo	Uganda	1.6M	Institutional	Ascending
bnx	Bangubangu	Niger-Congo	DRC	170K	Stable	Emerging
hem	Hemba	Niger-Congo	DRC	180K	Stable	Emerging
suk	Sukuma	Niger-Congo	Tanzania	1.8M	Stable	Ascending
lua	Tshiluba	Niger-Congo	DRC	15M	Institutional	Ascending
nnb	Nande	Niger-Congo	DRC	1.1M	Stable	Ascending
bkw	Kwele	Niger-Congo	DRC	120k	Stable	Emerging
nix	Hema	Niger-Congo	DRC	130k	Stable	Emerging
tll	Telela	Niger-Congo	DRC	760k	Stable	Ascending
bji	Burji	Afro-Asiatic	Ethiopia	83k	Endangered	Emerging
hoo	Holoholo	Niger-Congo	DRC	16k	Stable	Emerging
tap	Taabwa	Niger-Congo	DRC & Zambia	282k	Stable	Emerging
lsm	Olusaamya	Niger-Congo	Kenya & Uganda	480k	Institutional	Emerging
orc	Orma	Afro-Asiatic	Kenya	92k	Stable	Emerging
teo	Teso	Nilo-Saharan	Kenya & Uganda	2.78M	Institutional	Ascending
tuy	Tugen	Nilo-Saharan	Kenya	140k	Stable	Still
ttj	Tooro	Niger-Congo	Uganda	490k	Institutional	Ascending
xog	Soga	Niger-Congo	Uganda	3.1M	Institutional	Ascending
nyo	Runyoro	Niger-Congo	Uganda	670k	Institutional	Ascending
tuv	Turkana	Nilo-Saharan	Kenya	1.02M	Stable	Ascending
saq	Samburu	Nilo-Saharan	Kenya	240k	Stable	Ascending
rel	Rendille	Afro-Asiatic	Kenya	96k	Stable	Ascending
kin	Kinyarwanda	Niger-Congo	Uganda & Rwanda	15M	Institutional	Vital
nle	Nyala	Niger-Congo	Kenya	47.7K	Stable	Emberging
niq	Nandi	Nilo-Saharan	Kenya	950K	Stable	Ascending
mer	Meru	Niger-Congo	Kenya	2M	Institutional	Ascending
kde	Makonde	Niger-Congo	Tanzania & Mozambique & kenya	2.1M	Stable	Ascending
lug	Ganda	Niger-Congo	Uganda	11M	Institutional	Vital

Table 31: **Overview of the languages represented in our dataset.** For each language, we report its ISO code, language family, geographic region of use, estimated speaker population, vitality status, and level of digital support, compiled as compiled *Ethnologue*[5] and *Glottolog*[6] and related resources.

Category	Description
Language Vitality	
Institutional	The language is developed and sustained by institutions beyond the home and community.
Stable	The language is not institutionally supported but remains widely used in the home and community.
Endangered	It is no longer the norm that children learn and use the language.
Extinct	The language is no longer used and has no remaining community identity.
Digital Language Support	
Still	The language shows no signs of digital support.
Emerging	The language has some digital content and/or encoding tools.
Ascending	The language has tools such as spell checkers, localization, or machine translation.
Vital	The language is supported by multiple tools, including speech processing.
Thriving	The language has extensive support, including virtual assistants and advanced technologies.

Table 32: **Language vitality and digital support levels.** Vitality categories are derived from *Ethnologue*[7] and reflect intergenerational transmission, while digital support levels capture the extent of technological resources available for each language.

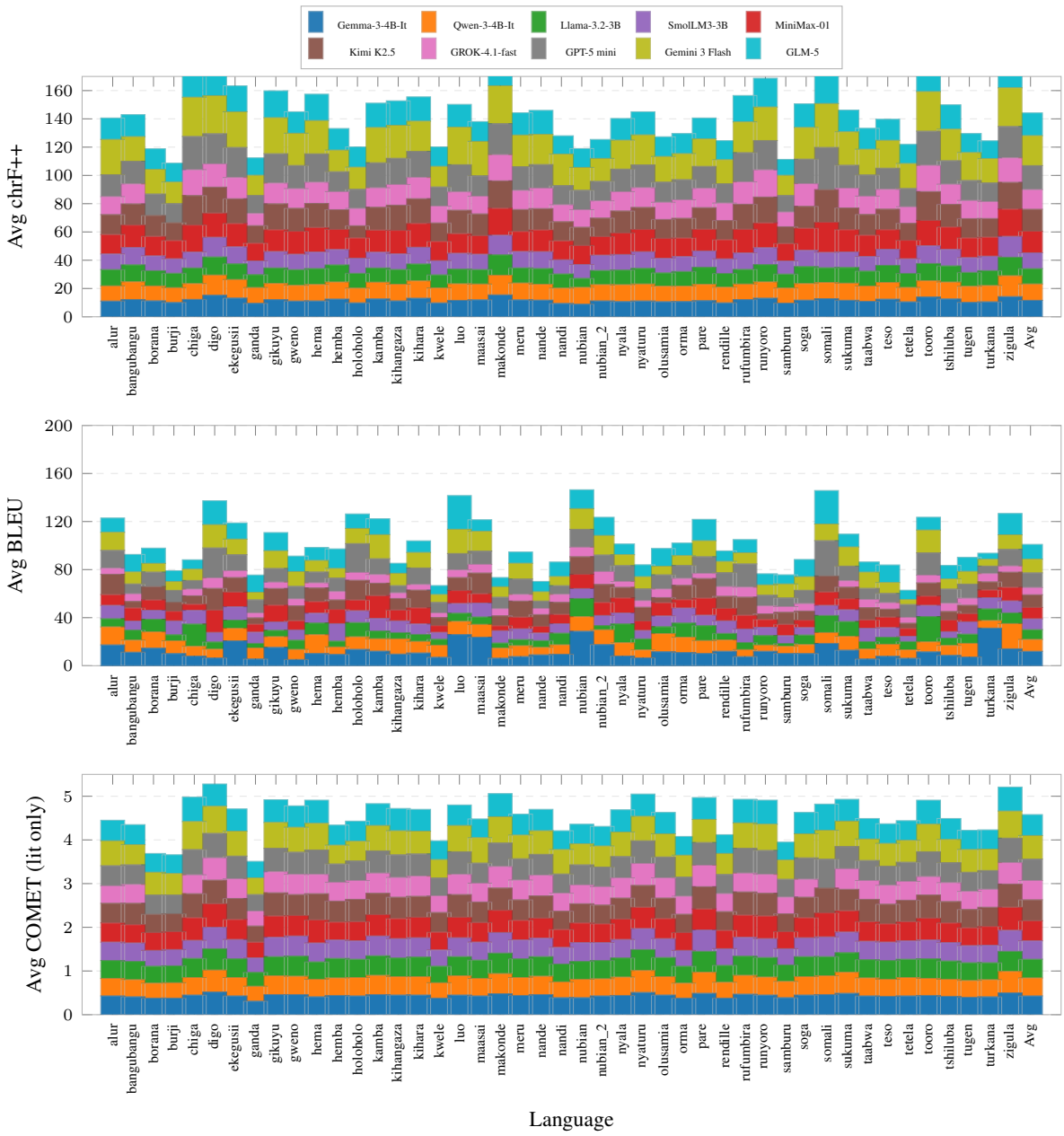

Figure 9: Stacked bars show per-model contributions. chrF++ and BLEU averaged across 4 tasks. COMET averaged across swa_lit and eng_lit only. Position 46 = average across all languages. GROK-4.1-fast has few languages missing.

Language	swa literal	eng literal	swa figurative	eng figurative
Kikuyu	1.17	1.60	1.80	2.72
Ekegusii	3.59	3.16	3.33	3.65
Luo	1.49	1.92	2.34	2.70
Maasai	2.48	1.96	1.11	1.41
Meru	2.70	2.92	2.64	2.79
Olusamia	1.30	1.97	1.93	2.13
Somali	1.48	1.52	2.27	1.96
Sukuma	1.13	1.69	1.71	2.63

Table 33: Human evaluation task averages by language across literal and figurative translation settings.

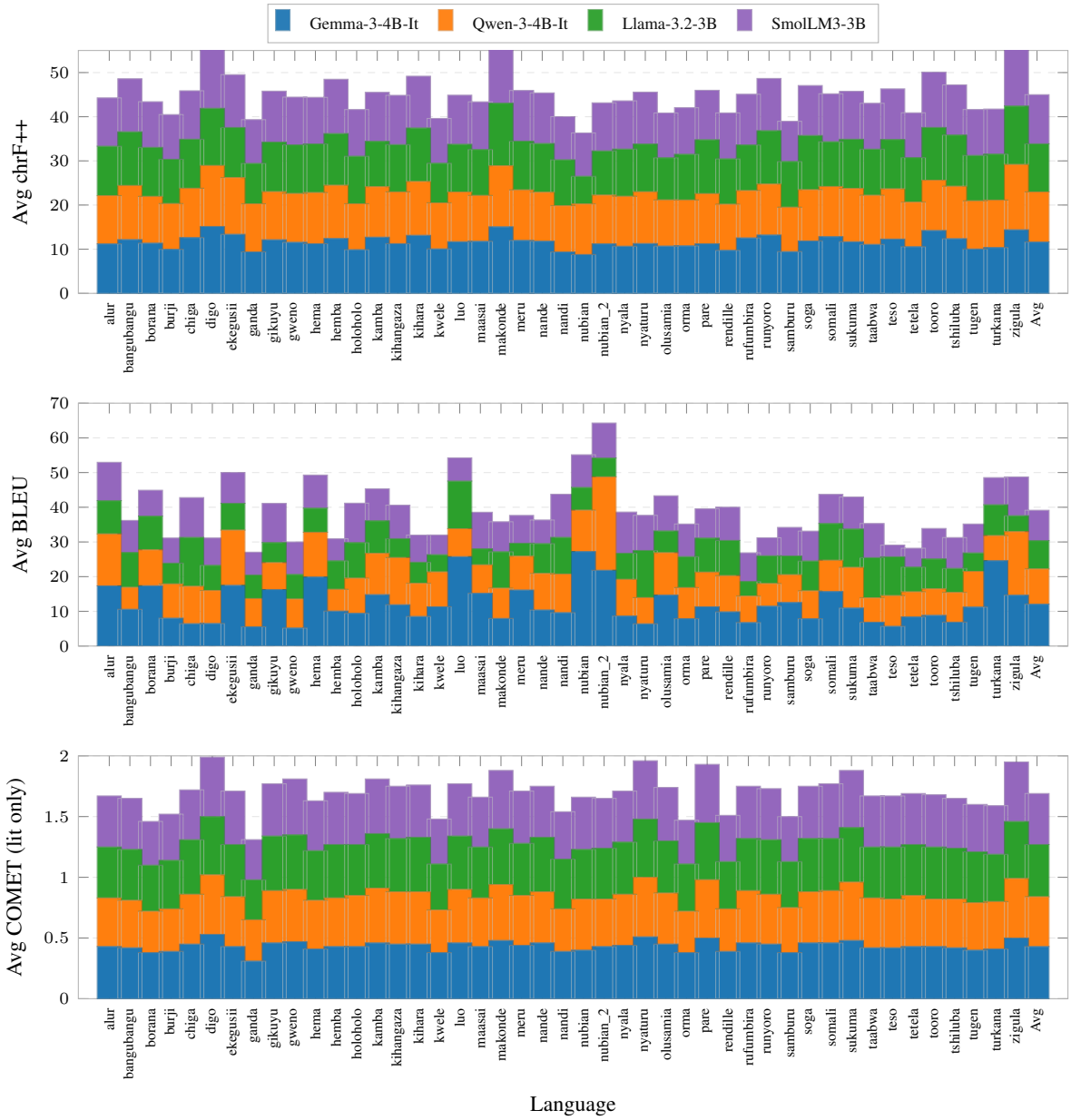

Figure 10: Few-shot (3-shot). Stacked bars show per-model contributions. chrF++ and BLEU averaged across 4 tasks. COMET averaged across swa_lit and eng_lit only. Position 46 = average across all languages.

Lang	Minimax	Gemma	Llama	Qwen	SmolLM3	Tie
Kikuyu	**21.3**	13.3	8.5	5.9	5.3	45.7
Ekegusii	22.5	**25.0**	2.5	7.5	12.5	30.0
Luo	**27.5**	22.5	2.5	5.0	10.0	32.5
Maasai	2.0	5.5	5.0	**11.1**	9.5	66.8
Meru	20.1	18.1	15.5	**22.0**	18.8	31.7
Olusamia	**18.9**	14.7	6.3	12.6	7.4	48.4
Somali	**30.4**	5.2	0.4	1.6	1.2	61.2
Sukuma	**32.5**	17.5	6.3	11.3	8.8	23.8

Table 34: Win rates (%) by language. Best model per language is bolded.

Lang	MiniMax	Gemma	Llama	Qwen	SmolLM
Kikuyu	**2.49**	2.04	1.53	1.57	1.49
Ekegusii	3.21	3.42	3.31	**3.64**	3.58
Luo	**2.48**	2.40	2.11	1.62	1.64
Maasai	1.35	1.66	1.38	**2.27**	2.05
Meru	**2.88**	2.75	2.69	2.77	2.78
Olusamia	**2.15**	1.88	1.42	1.68	1.72
Somali	**2.95**	2.14	1.28	1.52	1.61
Sukuma	**2.43**	1.92	1.57	1.51	1.51

Table 35: Average human evaluation scores by model and language. Best model per language is bolded.

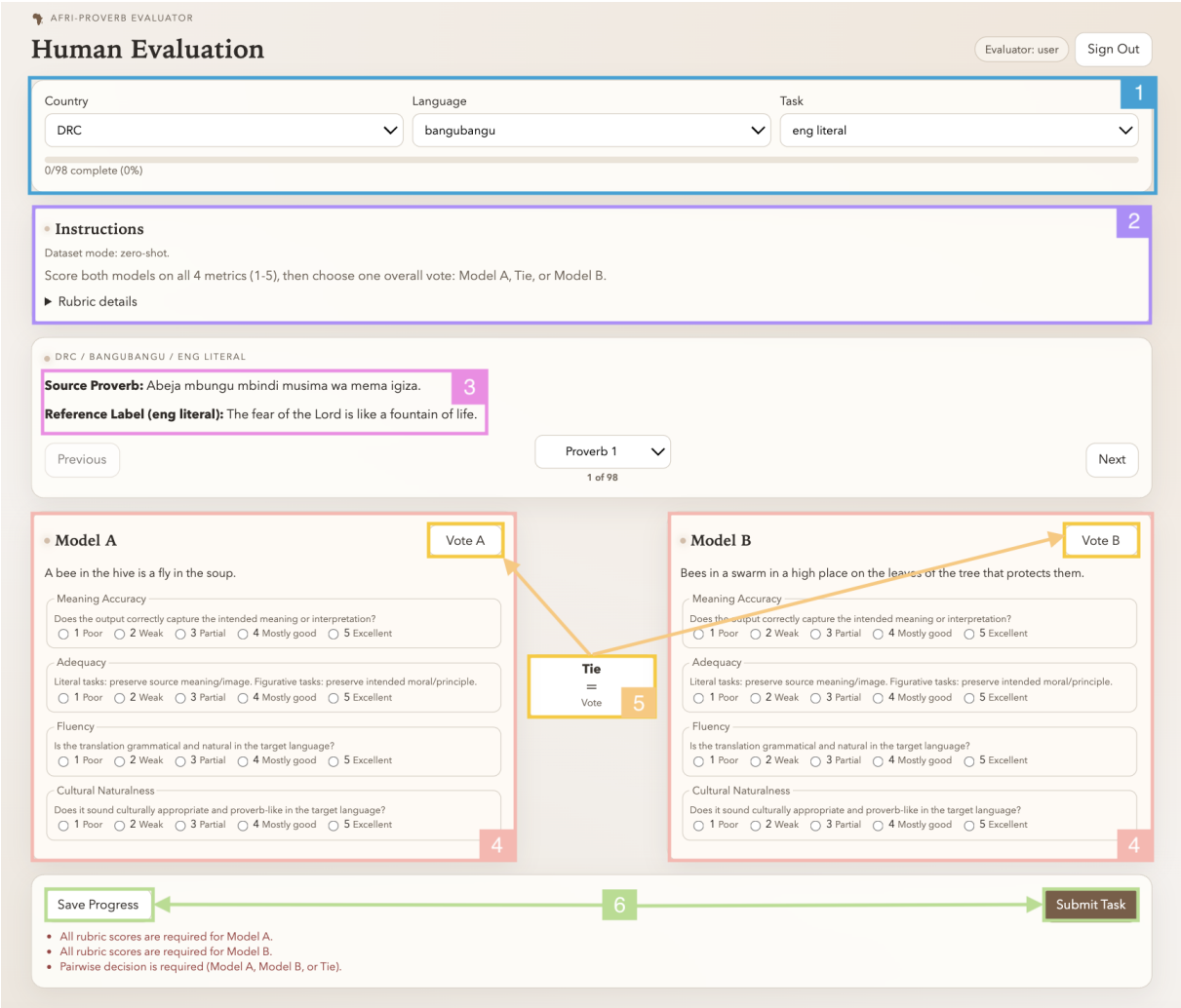

Figure 11: Annotation interface used in the human evaluation.
(1) Select the country, language, and task. (2) Read instructions and rubric. (3) Read source proverb and reference label. (4) Score each blind output on four rubric dimensions. (5) Vote for a final pairwise preference. (6) Completed annotations can be saved or submitted.

Annotator ID	Language	Task	Rows	Completion
A1	olusamia	eng_figurative	44/100	44%
A1	olusamia	swa_figurative	47/100	47%
A1	olusamia	swa_literal	53/100	53%
A1	olusamia	eng_literal	50/100	50%
A2	nandi	eng_figurative	100/100	100%
A3	luo	eng_figurative	52/519	10%
A3	luo	eng_literal	149/519	29%
A3	luo	swa_figurative	51/519	10%
A3	luo	swa_literal	52/519	10%
A4	luo	eng_literal	50/519	10%
A5	samburu	swa_literal	100/100	100%
A6	meru	eng_figurative	51/520	10%
A6	meru	eng_literal	67/520	13%
A6	meru	swa_figurative	51/520	10%
A6	meru	swa_literal	51/520	10%
A7	ekegusii	eng_figurative	50/514	10%
A7	ekegusii	eng_literal	318/514	62%
A7	ekegusii	swa_figurative	52/514	10%
A7	ekegusii	swa_literal	514/514	100%
A8	digo	eng_literal	28/389	7%
A9	maasai	eng_figurative	50/205	24%
A9	maasai	eng_literal	50/205	24%
A9	maasai	swa_figurative	50/205	24%
A9	maasai	swa_literal	50/205	24%
A10	gikuyu	eng_figurative	46/554	8%
A10	gikuyu	eng_literal	46/554	8%
A10	gikuyu	swa_figurative	50/554	9%
A10	gikuyu	swa_literal	46/554	8%
A11	somali	eng_figurative	50/500	10%
A11	somali	swa_figurative	50/500	10%
A11	somali	swa_literal	50/500	10%
A11	somali	eng_literal	100/500	20%
A12	sukuma	eng_figurative	100/500	20%
A12	sukuma	eng_literal	100/500	20%
A12	sukuma	swa_figurative	100/500	20%
A12	sukuma	swa_literal	100/500	20%

Table 36: Anonymized evaluator task coverage across languages and translation settings. Rows indicate completed annotations over total assigned instances.

Outcome	Wins	Win Rate (%)
minmax	40	21.3
gemma3-4b-it	25	13.3
llama-3.2-3b	16	8.5
qwen3-4b-it	11	5.9
smollm3-3b	10	5.3
Tie	86	45.7

Table 37: Pairwise human preference win rates for the Gikuyu evaluation set (N=188).

Model	Meaning	Adequacy	Fluency	Cultural	Avg
MinMax	2.36	2.36	2.36	2.36	**2.36**
Gemma-3-4B	2.22	2.22	2.22	2.22	2.22
Qwen-3-4B	2.15	2.15	2.15	2.15	2.15
SmolLM-3B	1.95	1.95	1.95	1.95	1.95
Llama-3.2-3B	1.88	1.88	1.88	1.88	1.88

Table 38: Average human evaluation scores for Olusamia, aggregated across tasks and evaluators.

Model	Wins	Losses	Ties	Win Rate (%)
MinMax	22	6	34	**78.6**
Gemma-3-4B	18	8	36	69.2
Qwen-3-4B	16	10	36	61.5
SmolLM-3B	10	14	38	41.7
Llama-3.2-3B	8	16	38	33.3

Table 39: Pairwise win rates for Olusamia evaluation (ties excluded from win-rate denominator).

Error Type	Description	Example Output	Explanation
Source Copying	Model reproduces the input without translating it.	*"A kibba si mbaya."*	The output remains in the source language instead of English, indicating failure to perform translation due to limited lexical knowledge.
Hallucinated Proverb	Model generates a fluent but unrelated proverb.	*"A wolf eats its own tail."*	The output is grammatically correct and proverb-like but semantically unrelated to the source, reflecting reliance on memorized patterns rather than grounded translation.
Prompt Artifact	Output contains task instructions or formatting text.	*"The literal translation of the proverb in English is..."*	The model fails to separate instructions from generation, leading to leakage of prompt text into the output.
Incomplete Output	Output is truncated or unfinished.	*"The figurative meaning of the proverb is:"*	The model begins generation but does not complete the translation, indicating decoding instability or length issues.
Partial Semantic Alignment	Output captures part of the meaning but is incomplete or distorted.	*"The hunter is the hunted."*	The model approximates the intended meaning (reversal or consequence) but does not preserve the original imagery or literal structure.
Overgeneralization	Output reduces the proverb to a generic moral statement.	*"Persistence leads to success."*	The model ignores specific cultural imagery and replaces it with a high-frequency abstract statement.
Wrong Imagery Substitution	Model introduces unrelated entities or metaphors.	*"The elephant eats the biggest part."*	New imagery (elephant) replaces the original meaning, leading to semantic distortion.
Language Mixing	Output contains a mix of source and target languages.	*"Akiba si mbaya is good."*	The model partially translates but leaves source tokens, indicating weak alignment between languages.
Literal Misinterpretation	Model translates words but fails to capture intended meaning.	*"The child does not know the cow does not graze."*	Word-level translation is attempted, but the resulting sentence is semantically incoherent.
Fluent but Incorrect	Output is grammatically correct but meaning is wrong.	*"Actions speak louder than words."*	The model produces a well-formed sentence unrelated to the source, showing fluency without understanding.

Table 40: Qualitative error taxonomy for proverb translation. We identify common failure modes across models, including source copying, hallucination, and semantic misalignment.

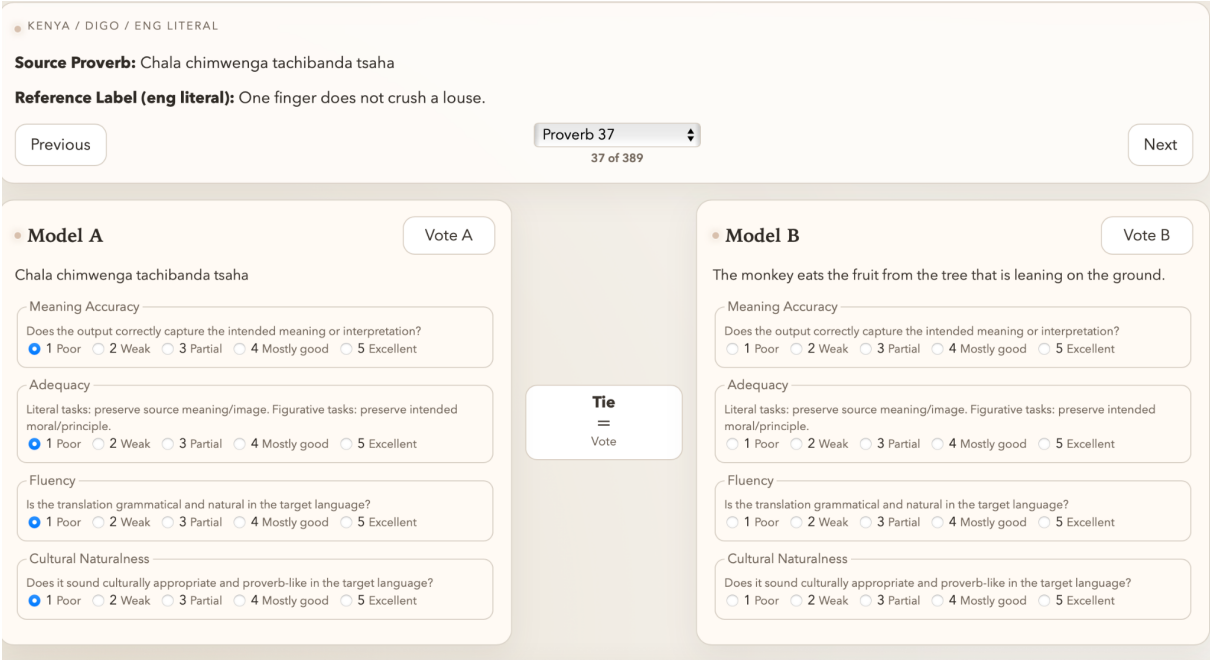

Figure 12: Example of the pairwise human evaluation interface used for proverb translation assessment. Annotators compare outputs from two models for the same proverb and rate each output on four dimensions: meaning accuracy, adequacy, fluency, and cultural naturalness. They then select the preferred output or mark the comparison as a tie. While both models fail to capture the intended meaning, Model A exhibits a degenerate failure mode by copying the source without translation, whereas Model B produces a fluent but semantically incorrect output. Non-translation is penalized more heavily than mistranslation, leading us to prefer Model B.

Language.	Model	n	Accuracy.	Adequacy.	Fluency.	Cultural.	Avg
Borana	minmax	82	**1.18**	**1.22**	**1.21**	**1.22**	**1.21**
	llama-3.2-3b	87	1.08	1.13	1.18	1.11	1.13
	gemma3-4b-it	79	1.04	1.04	1.08	1.05	1.05
	qwen3-4b-it	82	1.04	1.04	1.04	1.04	1.04
	smollm3-3b	82	1.00	1.01	1.02	1.01	1.01
Digo	minmax	81	**1.72**	**1.73**	**3.21**	**2.00**	**2.16**
	gemma3-4b-it	84	1.17	1.18	2.57	1.69	1.65
	qwen3-4b-it	81	1.12	1.12	2.01	1.37	1.41
	smollm3-3b	80	1.09	1.09	1.89	1.39	1.36
	llama-3.2-3b	82	1.00	1.00	1.87	1.35	1.30
Ekegusii	qwen3-4b-it	372	**3.55**	**3.57**	**3.69**	**3.74**	**3.64**
	smollm3-3b	375	3.47	3.53	**3.69**	3.66	3.58
	gemma3-4b-it	377	3.31	3.33	3.56	3.49	3.42
	llama-3.2-3b	373	3.21	3.21	3.43	3.39	3.31
	minmax	375	3.03	3.08	3.42	3.30	3.21
Gikuyu	minmax	75	**1.84**	**1.84**	**3.49**	**2.77**	**2.49**
	gemma3-4b-it	75	1.23	1.23	3.31	2.41	2.04
	qwen3-4b-it	75	1.09	1.11	2.20	1.87	1.57
	llama-3.2-3b	76	1.11	1.11	2.29	1.63	1.53
	smollm3-3b	75	1.08	1.08	2.07	1.73	1.49
Kamba	minmax	91	**2.11**	**2.11**	**3.63**	**2.87**	**2.68**
	gemma3-4b-it	96	1.40	1.42	3.18	2.42	2.10
	qwen3-4b-it	91	1.38	1.35	2.52	2.02	1.82
	llama-3.2-3b	93	1.27	1.28	2.62	2.08	1.81
	smollm3-3b	91	1.30	1.27	2.46	2.01	1.76
Kihangaza	minmax	83	**2.05**	**2.05**	**2.64**	**2.27**	**2.25**
	gemma3-4b-it	80	1.76	1.76	2.38	2.04	1.98
	llama-3.2-3b	84	1.55	1.55	2.26	1.98	1.83
	qwen3-4b-it	82	1.57	1.54	1.82	1.67	1.65
	smollm3-3b	81	1.48	1.48	1.90	1.70	1.64
Luo	minmax	141	**2.49**	**2.48**	**2.50**	**2.48**	**2.49**
	gemma3-4b-it	146	2.38	2.34	2.40	2.34	2.36
	llama-3.2-3b	142	2.12	2.10	2.15	2.13	2.13
	smollm3-3b	140	1.57	1.59	1.56	1.54	1.56
	qwen3-4b-it	141	1.57	1.57	1.52	1.55	1.55
Maasai	qwen3-4b-it	80	**2.30**	**2.29**	**2.29**	**2.30**	**2.29**
	smollm3-3b	82	2.05	2.05	2.06	2.09	2.06
	gemma3-4b-it	81	1.64	1.64	1.62	1.70	1.65
	llama-3.2-3b	83	1.39	1.35	1.36	1.39	1.37
	minmax	82	1.34	1.34	1.33	1.37	1.34
Meru	minmax	93	**3.03**	**2.86**	2.77	**2.61**	**2.82**
	qwen3-4b-it	92	2.93	2.67	**2.89**	2.50	2.75
	gemma3-4b-it	94	2.88	2.65	2.79	2.56	2.72
	smollm3-3b	91	2.88	2.66	2.86	2.46	2.71
	llama-3.2-3b	92	2.74	2.67	2.68	2.55	2.66
Nandi	qwen3-4b-it	122	**2.64**	**2.61**	**2.75**	**2.62**	**2.66**
	gemma3-4b-it	120	2.09	2.08	2.17	2.10	2.11
	smollm3-3b	122	2.06	2.06	2.17	2.06	2.09
	minmax	122	1.87	1.87	2.36	1.87	1.99
	llama-3.2-3b	120	1.69	1.69	1.77	1.69	1.71
Nyala	minmax	84	**1.61**	**1.63**	**1.88**	**1.62**	**1.68**
	gemma3-4b-it	84	1.31	1.32	1.49	1.35	1.37
	qwen3-4b-it	84	1.30	1.27	1.46	1.29	1.33
	smollm3-3b	85	1.24	1.24	1.41	1.25	1.28
	llama-3.2-3b	83	1.19	1.20	1.43	1.25	1.27
Olusamia	minmax	76	**2.22**	**2.24**	**2.24**	**2.24**	**2.23**
	gemma3-4b-it	80	1.93	1.93	1.93	1.93	1.93
	smollm3-3b	77	1.70	1.70	1.73	1.70	1.71
	llama-3.2-3b	78	1.64	1.64	1.64	1.69	1.65
	qwen3-4b-it	77	1.53	1.53	1.53	1.53	1.53
Pare	minmax	83	**1.36**	**1.36**	**3.08**	**1.87**	**1.92**
	llama-3.2-3b	82	1.07	1.06	2.83	1.71	1.67
	gemma3-4b-it	83	1.14	1.14	2.16	1.53	1.49
	smollm3-3b	84	1.08	1.08	2.01	1.39	1.39
	qwen3-4b-it	82	1.09	1.09	1.82	1.35	1.34
Samburu	qwen3-4b-it	40	**2.90**	**2.85**	**3.10**	**3.33**	**3.04**
	smollm3-3b	40	2.33	2.38	2.48	2.75	2.48
	llama-3.2-3b	40	1.73	1.77	1.98	2.20	1.92
	gemma3-4b-it	40	1.63	1.70	1.82	1.85	1.75
	minmax	40	1.18	1.20	1.32	1.50	1.30
Somali	minmax	101	**2.72**	**2.88**	**3.60**	**2.91**	**3.03**
	gemma3-4b-it	99	1.14	1.27	2.31	2.43	1.79
	smollm3-3b	99	1.03	1.05	1.68	1.73	1.37
	qwen3-4b-it	101	1.02	1.06	1.79	1.44	1.33
	llama-3.2-3b	100	1.04	1.04	1.46	1.36	1.23
Sukuma	minmax	160	**1.82**	**1.77**	**3.91**	**2.23**	**2.43**
	gemma3-4b-it	160	1.25	1.22	3.31	1.91	1.92
	llama-3.2-3b	160	1.07	1.06	2.53	1.61	1.57
	smollm3-3b	160	1.16	1.14	2.14	1.60	1.51
	qwen3-4b-it	160	1.11	1.11	2.38	1.42	1.51

Table 41: **Human evaluation by language and model.** Scores are 1–5 averages for meaning accuracy, adequacy, fluency, cultural naturalness, and overall mean. Languages with very small sample sizes are excluded. The languages were selected based on human annotators

Language	Gemma	LLaMA	MinMax	Qwen	SmolLM3	Best
borana	1.12	1.09	**1.23**	1.06	1.15	minmax
digo	1.65	1.30	**2.16**	1.41	1.36	minmax
ekegusii	3.42	3.31	3.21	**3.64**	3.58	qwen3-4b-it
gikuyu	2.04	1.53	**2.49**	1.57	1.49	minmax
kamba	2.10	1.80	**2.68**	1.82	1.76	minmax
kihangaza	2.24	2.05	2.04	**2.34**	2.15	qwen3-4b-it
luo	2.36	2.13	**2.49**	1.55	1.56	minmax
maasai	1.65	1.37	1.34	**2.29**	2.06	qwen3-4b-it
meru	2.72	2.66	**2.82**	2.75	2.71	minmax
nandi	2.63	2.04	2.30	**3.48**	2.61	qwen3-4b-it
olusamia	1.93	1.65	**2.23**	1.53	1.71	minmax
samburu	1.75	1.92	1.30	**3.04**	2.48	qwen3-4b-it
somali	1.79	1.23	**3.03**	1.33	1.37	minmax
sukuma	1.92	1.57	**2.43**	1.51	1.51	minmax

Table 42: **Human evaluation by language and model.** Each cell reports the overall mean score. Best model per language is shown in bold. Languages with very small sample sizes or unstable estimates are excluded for more reliable comparison.

Model	swa_lit			eng_lit			swa_fig		eng_fig	
	BLEU	ChrF++	COMET	BLEU	ChrF++	COMET	BLEU	ChrF++	BLEU	ChrF++
Gemma-3-4B-It	7.49	11.07	0.460	11.00	13.15	0.380	10.53	10.65	**14.88**	11.68
Qwen-3-4B-It	6.03	10.88	**0.488**	8.03	9.74	0.333	**10.95**	**11.26**	9.82	**14.29**
Llama-3.2-3B	3.95	**11.09**	0.468	**11.86**	**12.33**	**0.387**	6.02	10.11	8.14	11.73
SmolLM3-3B	5.46	10.94	0.470	8.10	11.74	0.370	8.07	10.49	10.49	11.46

Table 43: Average 0-shot performance of open-weight models across literal and figurative tasks on AFRI-PROVERB.

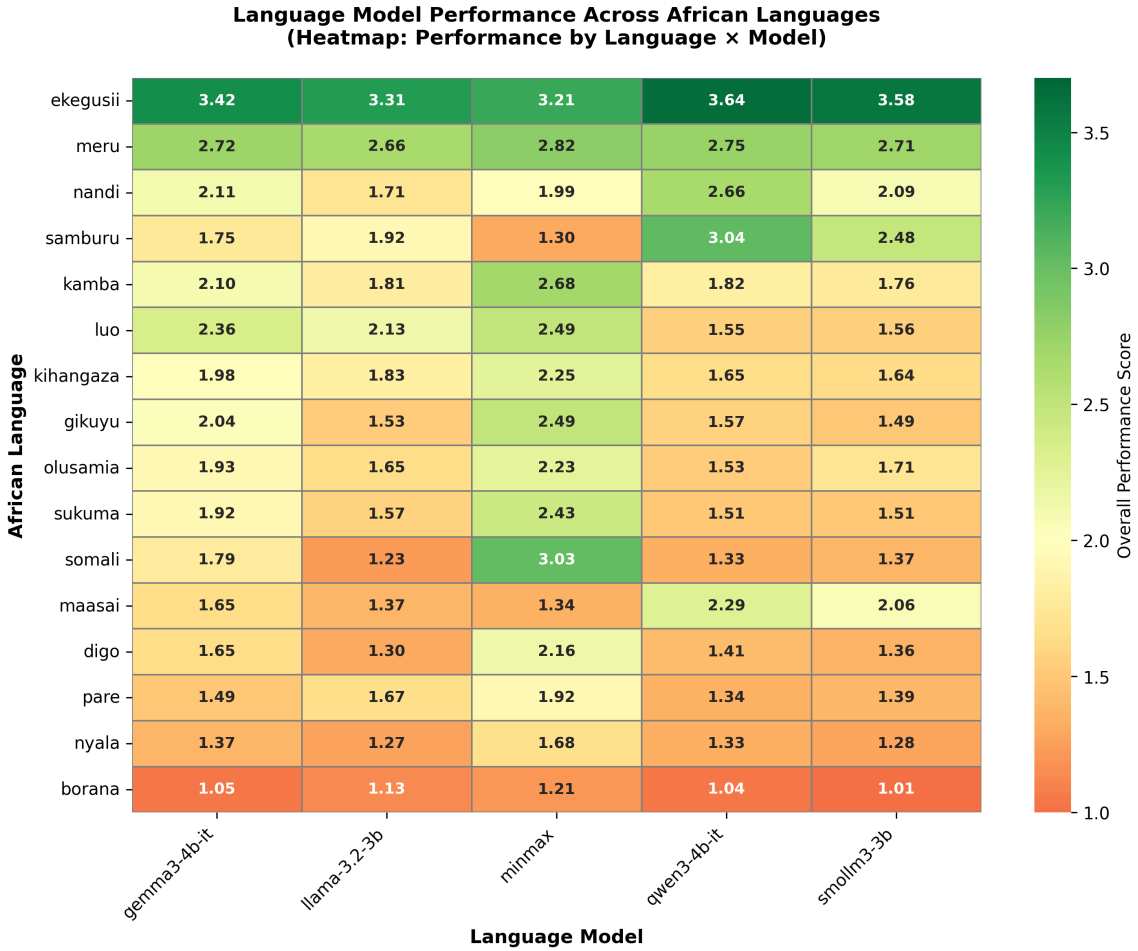

Figure 13: Human evaluation language model performance across languages. Each cell shows the average performance score for a language–model pair, with darker green shows higher performance and orange lower.

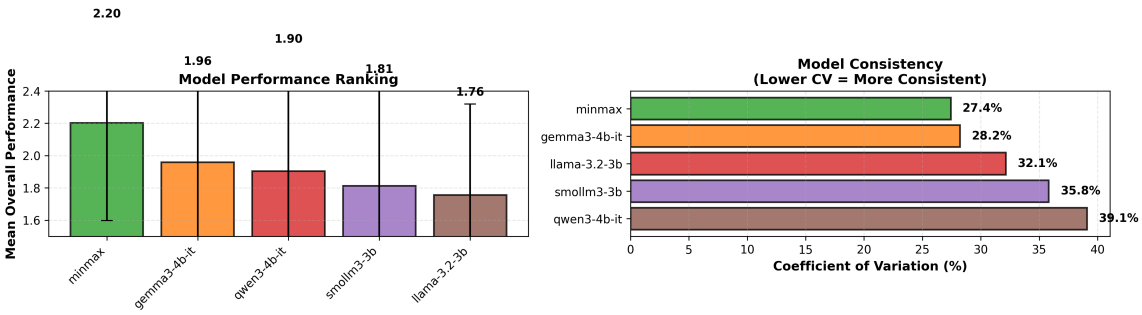

Figure 14: **Model-level human evaluation performance and consistency.** Left: mean overall performance across models with error bars showing score variability. Right: coefficient of variation (CV), where lower values indicate more consistent performance across languages and tasks. MinMax achieves the highest average performance and the lowest CV, indicating both strong and stable performance.

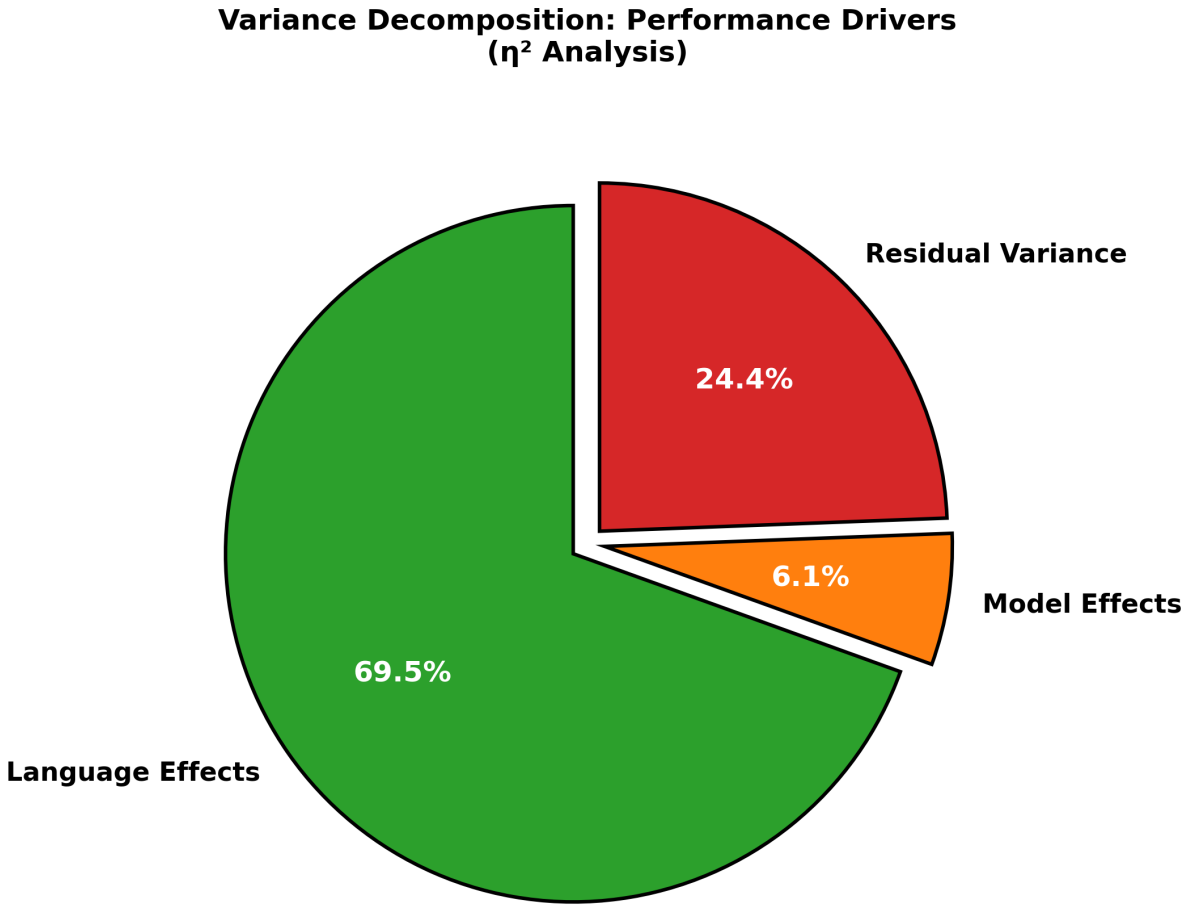

Figure 15: Variance decomposition analysis (η^2) showing the relative contribution of different factors to overall performance variation. Language effects account for the largest proportion of variance (69.5%), substantially exceeding model effects (6.1%), while the remaining unexplained variance accounts for 24.4%.

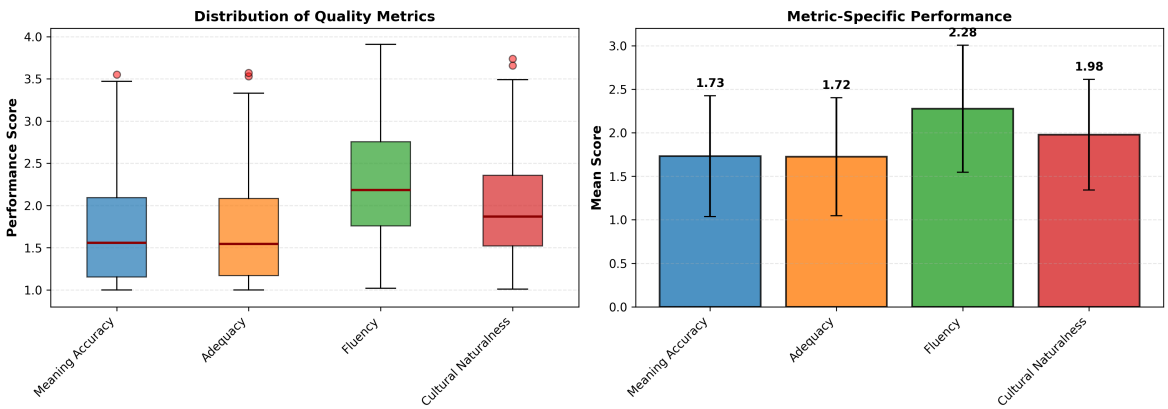

Figure 16: Analysis of human evaluation metrics across model outputs. Left: Distribution of scores for Meaning Accuracy, Adequacy, Fluency, and Cultural Naturalness using box plots. Right: Mean scores with standard deviation bars for each evaluation metric. Results reveal a clear fluency–meaning mismatch: models achieve comparatively higher fluency (2.28) than semantic accuracy (1.73) and adequacy (1.72), indicating that fluent outputs do not necessarily preserve intended meaning.

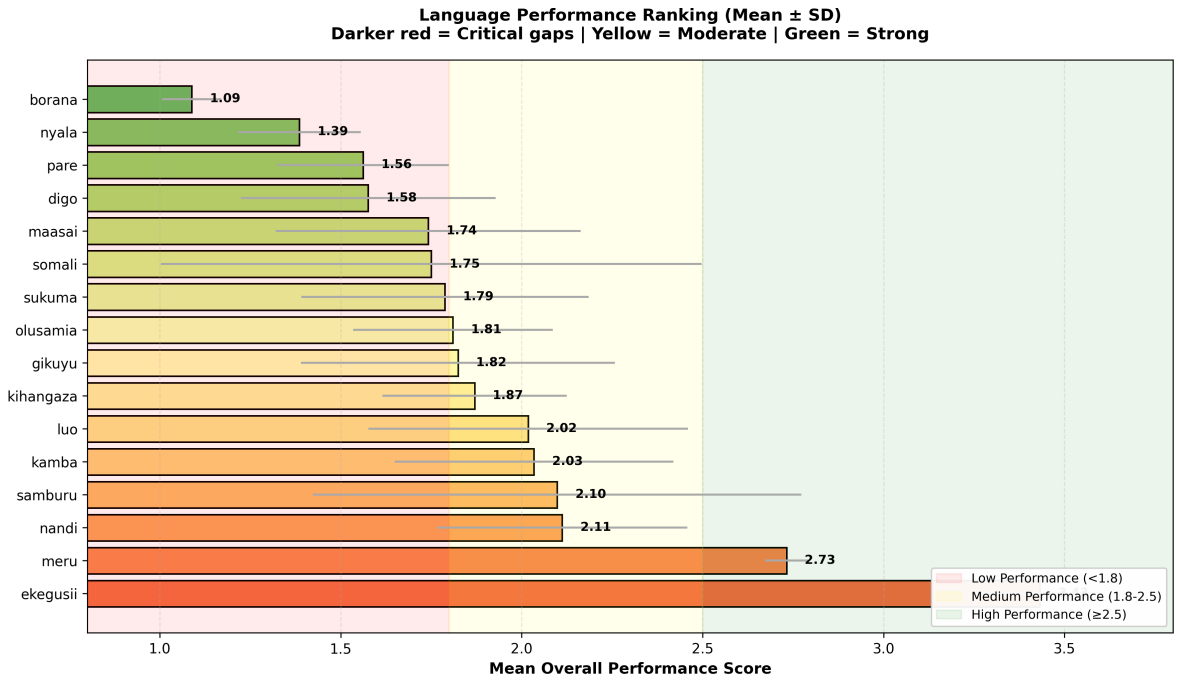

Figure 17: **Language-level performance distribution across the human evaluation study.** Languages are ranked by mean overall score, with error bars indicating standard deviation across evaluated instances. Background shading highlights performance strata: low performance (< 1.8), medium performance (1.8–2.5), and high performance (≥ 2.5). Performance varies substantially across languages, suggesting that language-specific characteristics contribute more strongly to outcome variation than model architecture alone.

Error Type	Count	%
Hallucination	1,270	45.8
Source copying	879	31.7
Wrong target language	842	30.4
Truncated/incomplete	382	13.8
Semantic distortion/drift	256	9.2
Meta/instruction leakage	87	3.1
Incoherent/nonsensical	69	2.5
Proverb substitution	46	1.7
Code-switching	9	0.3
Good output (none)	13	0.5

Table 44: Error distribution from 2,770 human-evaluated outputs. Multiple error categories may co-occur for a single output; percentages therefore do not sum to 100%.

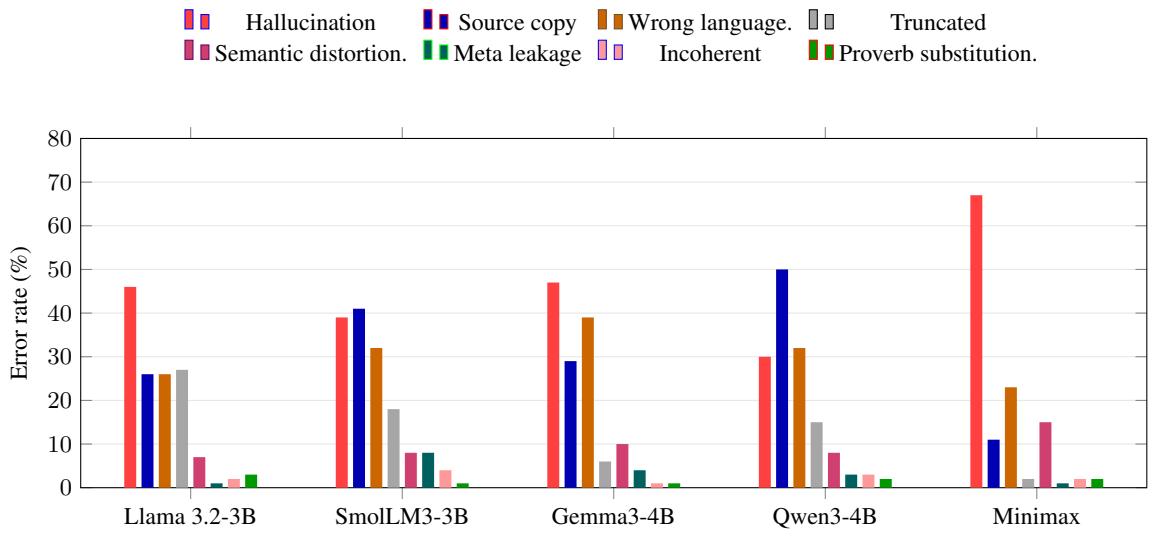

Figure 18: Error type distribution by model. Values indicate the percentage of evaluated outputs assigned each error category.