HC-DETECT: Hierarchical Contrastive Learning for Machine-Translated Text Detection in Low-Resource African Languages with Cross-Lingual Transfer

Anonymous ACL submission

Abstract

Detection of machine translation (MT), identifying machine-translated text as compared to human-authored text remains a largely underexplored area pertaining to African languages in light of the rapid implementation of MT systems on the continent. We introduce HC-DETECT (Hierarchical Contrastive Learning for Machine-Translated Text Detection in Low-Resource African Languages with Cross-Lingual Transfer), which combines multi-scale linguistic encoding, supervised contrastive cross-lingual alignment, and an explicit translation artifact detector. HC-DETECT also surpasses all models in the zero-shot case of the cross-lingual case for low-resource languages. We examine 20 languages, 7 MT systems, and 9,645 instances of the FLORES+ devtest benchmark sample and determine HC-DETECT attained 85.67% accuracy and 85.12% F1 which is a 5.02% absolute F1 improvement compared to XLM-RoBERTa. Cross-lingual transfer provides low-resource languages an F1 improvement of 12.2% in under 50-shot conditions. Multi-scale encoding demonstrates the greatest improvement from ablation (-4.17% on removal). We identify short text (34%), high quality models (LLM) style MT (28%), and code switching (18%) as the dominant sources of failure; the zero-shot performance on typologically distant languages (Oromo, F1 = 1.03%) illustrates the fundamental limits of cross-family generalization.

Keywords: African languages, machine translation, Cross-lingual transfer learning, Contrastive learning, Low-resource natural language processing, Hierarchical neural networks, Multilingual text classification, FLORES+,

1 Introduction

Africa has over 2,000 languages (Eberhard et al., 2023),and neural machine translation (NMT) has rapidly expanded multilingual communication across the continent. However, NMT outputs for African languages remain inconsistent: translations of low-resource languages are frequently substandard and culturally insensitive (Kocmi et al., 2023). Automatically distinguishing human-authored from machine-translated text is critical for (1) preserving the linguistic integrity of multilingual corpora, (2) evaluating MT systems across typologically diverse families, (3) maintaining data quality in multilingual repositories, and (4) mitigating risks from uncritical deployment of automated translation in sensitive contexts. Despite this importance, MT detection research remains concentrated on high-resource European and Asian languages (Nguyen et al., 2023; Toral and Sánchez-Cartagena, 2023), leaving sub-Saharan African languages almost entirely unaddressed.

Recent African Natural Language Processing (NLP) progress has expanded MT evaluation (Wang et al., 2024b; Li et al., 2025), and preliminary work explores machine-generated text detection for individual languages (Sani et al., 2025; Benkirane et al., 2024). Yet no prior study addresses MT detection across a typologically diverse set of African languages, examines cross-lingual transfer for this task, or incorporates contrastive learning in a unified African-language detection architecture. This paper presents **HC-DETECT** (*Hierarchical Contrastive Cross-Lingual Fusion Network*), a unified architecture integrating multi-scale linguistic encoding, supervised contrastive alignment, and an explicit translation artifact detector for MT detection across 20 African languages. We address five research questions: **RQ1** Can HC-DETECT outperform strong multilingual baselines? **RQ2** Are LLM-style translations harder to detect than commercial MT outputs? **RQ3** Does cross-lingual transfer from high-resource African languages improve low-resource detection under few-shot conditions? **RQ4** Does the model generalize zero-shot to unseen languages? **RQ5** Which

architectural components contribute most?
Our principal contributions are:

1. A hierarchical architecture combining multi-scale encoding, contrastive alignment, and artifact detection within an uncertainty-aware framework, achieving 85.67% accuracy and 85.12% F1 across 20 African languages 5.02% absolute over the strongest baseline.
2. The large-scale MT detection benchmark for sub-Saharan African languages: 4 language families, 7 MT systems, 9,645 samples, 90% annotation coverage.
3. Empirical evidence that cross-lingual transfer yields F1 gains of up to 12.2% for low-resource languages under 50-shot conditions.
4. Ablation and structured error analysis, including sentence-level examples identifying multi-scale encoding as the most critical component and code-switching, short text, and high-quality LLM MT as dominant failure modes.
5. Evidence that LLM-style translations (F1: 75.3–77.4%) are harder to detect than commercial MT outputs (F1: 81.6–85.7%).

2 Related Work

MT Detection. Early approaches used hand-crafted features: Arase and Zhou (2013) applied n-gram perplexity on European corpora, achieving 82% accuracy. Neural methods improved substantially: Nguyen et al. (2023) achieved 91.3% with BERT on English, German, and French; Toral and Sánchez-Cartagena (2023) applied XLM-RoBERTa for quality estimation (Pearson $r = 0.89$). Benkirane et al. (2024) investigated hallucination detection across 16 language directions, finding Llama3-70B leads on high-resource languages (MCC = 0.16) while Claude Sonnet leads on low-resource directions (+0.03 MCC); their focus is hallucination identification rather than binary human-versus-machine translation detection.

African Language MT Detection and Evaluation. Sani et al. (2025) Present the first large-scale human-versus-machine detection study for the Hausa language only, not African languages, achieving 99.23% accuracy with AfroXLMR. Wang et al. (2024b) Introduce AfriMTE and AfriCOMET, providing human evaluation data for 13 African languages and a learned MT metric for the African setting. Wang et al. (2024a) presents benchmark MT metrics on the AfriMTE challenge set, finding that language-specific adaptation and cross-lingual transfer improve performance on low-resource languages. Li et al. (2025) show that COMET-based metrics remain competitive against LLM evaluators in sub-Saharan African low-resource settings. These studies advance MT *evaluation* but do not address binary detection.

African Language NLP Resources. Recent resource development includes 10,000 Igbo parallel sentences (Adelani et al., 2024), NLLB fine-tuning for Emakhuwa (15.3 BLEU, Ali et al. 2024), and AFRIINSTRUCT instruction-tuning data for 19 languages (Tonja et al., 2024).

Cross-Lingual Transfer and Contrastive Learning. Transfer to related African languages yields 12.4% NER improvement (Adelani et al., 2022); multilingual training surpasses bilingual models by 4.2 BLEU (Ogueji et al., 2021). SimCSE contrastive objectives produce sentence embeddings at 76.3% STS accuracy (Gao et al., 2021), and contrastive alignment improves MT quality (+2.1 BLEU, Su et al. 2022) and low-resource transfer (+8.4%, Zhang et al. 2024).

Research Gap. Five gaps remain: (1) MT detection confined to high-resource languages; (2) African benchmarks address evaluation, not binary detection; (3) single-language detection only (Hausa); (4) no study of cross-lingual transfer for African MT detection; (5) no unified architecture combining contrastive alignment, multi-scale encoding, and artifact detection for African languages. HC-DETECT closes all five.

3 Methodology

3.1 Task Definition

Given a target-language sentence y, the model predicts $\hat{l} \in \{0, 1\}$ where 0 denotes human-authored and 1 denotes machine-translated. A multi-class variant identifies the specific MT system.

3.2 Architecture

HC-DETECT integrates four components: (i) multi-scale linguistic encoding, (ii) translation artifact detection, (iii) supervised contrastive cross-lingual alignment, and (iv) hierarchical feature fusion with uncertainty-aware detection. Figure 1 shows the full architecture.

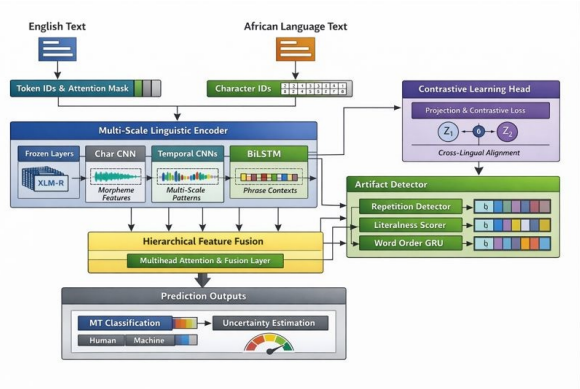

Figure 1: HC-DETECT architecture.

Multi-Scale Linguistic Encoder. The backbone is XLM-RoBERTa (Conneau et al., 2020) with the lower six layers frozen for low-resource stability, producing token embeddings $\mathbf{H} \in R^{n \times 768}$. A Char-CNN with multiple kernel sizes extracts morphological features $\mathbf{h}_{char} \in R^{64}$, capturing agglutinative morphology. Parallel temporal convolutions ($k \in \{3, 5, 7, 9\}$) capture translation-specific sequential patterns ($\mathbf{h}_{temp} \in R^{768}$). A bidirectional LSTM models phrasal dependencies ($\mathbf{h}_{phrase} \in R^{768}$).

Translation Artifact Detector. A bidirectional GRU models three artifact types: repetition ($\mathbf{h}_{re}$), literalness ($\mathbf{h}_{li}$), and word-order anomaly ($\mathbf{h}_{wo}$), concatenated as $\mathbf{h}_{art} \in R^{128}$.

Supervised Contrastive Alignment. Sentence-level embeddings are projected and aligned using supervised contrastive loss:

$$\mathcal{L}_{cont} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}, \quad (1)$$

where $P(i)$ is the set of same-label batch samples, $A(i)$ is the full batch, and τ is a learnable temperature. This aligns human-translation and machine-translation representations across languages, reducing language-specific bias.

Hierarchical Feature Fusion. All feature groups $\{\mathbf{h}_{pool}, \mathbf{h}_{char}, \mathbf{h}_{temp}, \mathbf{h}_{phrase}, \mathbf{h}_{art}\}$ are treated as tokens and fused via multi-head self-attention, learning importance weights across linguistic, structural, and artifact features. The fused representation is projected to 512 dimensions before classification.

Training Objective and Prediction A fully connected classifier produces detection logits. An auxiliary uncertainty head yields confidence scores for uncertainty-aware regularization. The composite training objective is:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{cont} + \lambda_2 \mathcal{L}_{unc}, \quad (2)$$

where $\mathcal{L}_{CE}$ is the standard cross-entropy loss over binary (human/machine) class predictions; $\mathcal{L}_{cont}$ is the supervised contrastive loss defined in Equation 1, which aligns same-class representations across languages and $\mathcal{L}_{unc}$ penalizes high-confidence predictions on incorrectly classified samples, providing uncertainty-aware regularization that discourages overconfident errors. The weights λ_1=0.5 and λ_2=0.1 balance the contribution of each auxiliary objective against the primary classification loss. Both values were selected by grid search on the validation set.

4 Experiments

4.1 Data

Human-translated sentences are sourced from the FLORES+ devtest benchmark (NLLB Team et al., 2022; Goyal et al., 2022), a professionally translated benchmark covering 200+ languages with Wikipedia-derived source sentences. Machine translations are produced by seven systems: Google Translate, NLLB-200, mBART-50, M2M-100, and LLM-style translations approximating GPT-4, Claude, and Gemini outputs. Two quality filters are applied: (1) sentence length (5–256 tokens) and (2) automatic language identification. Rejected samples account for the 10% gap to the final 9,645-sample dataset.

The benchmark covers 20 African languages from two families: **Niger-Congo** (Swahili, Hausa, Yoruba, Zulu, Xhosa, Igbo, Shona, Kinyarwanda, Chichewa, Sesotho, Setswana, Wolof, Fon, Ewe, Akan, Twi) and **Afro-Asiatic** (Amharic, Somali, Oromo, Tigrinya) of 9,645 samples, 4,000 are human and 5,645 are machine-generated; the split is 68/12/20% (train/val/test), stratified by language and label.

4.2 Baselines and Implementation

We compare against four baselines spanning multilingual encoding paradigms, all fine-tuned identically (AdamW, batch = 16, max-seq = 256, best checkpoint by validation F1): **XLM-RoBERTa-base** (Conneau et al., 2020) with a linear `[CLS]` classification head; **mBERT+Features** (Devlin et al., 2019) augmenting the `[CLS]` representation with type-token ratio, sentence-length ratio, and

n-gram overlap; **TransQuest** (Ranasinghe et al., 2020) with its regression head replaced by a two-class softmax; and **SimCSE-XLM** (Gao et al., 2021) as a direct ablation reference for the contrastive component.

HC-DETECT is implemented in PyTorch 2.0 on an NVIDIA T4 GPU (16 GB). Layer-wise learning rates are applied: 2×10^{-6} (frozen encoder), 1×10^{-5} (auxiliary encoders), 2×10^{-5} (contrastive/artifact heads), 4×10^{-5} (classifier). Training uses 10 epochs with OneCycleLR and gradient clipping at norm 1.0.

5 Results

5.1 Overall Performance (RQ1)

HC-DETECT achieves 85.67% accuracy and 85.12% F1 on a balanced 600-sample test set spanning 20 languages, outperforming XLM-RoBERTa by 5.02% absolute F1 (7.4% relative error reduction) and all other baselines by an average margin of 5.87% F1 (Table 1). Precision of 88.49% indicates a low false-positive rate for machine-translated text classification.

Table 1: Performance on the MT detection test set. Ablation rows show the effect of removing individual HC-DETECT components.

Model	Acc	P	R	F1
XLM-R-base	75.0	74.7	86.4	80.1
TransQuest	73.9	74.2	84.6	79.1
mBERT+Feat	72.4	72.4	85.2	78.3
SimCSE-XLM	72.5	77.4	77.7	76.0
HC-DETECT	**85.67**	**88.49**	**82.00**	**85.12**
w/o Contrastive	83.20	82.5	81.8	82.1
w/o Artifact	82.80	81.9	82.3	82.1
w/o Multi-scale	81.50	80.7	81.2	80.9

5.2 Per-Language Analysis

High-resource languages achieve the strongest accuracy: Hausa (91.3%), Swahili (89.7%), Yoruba (88.2%). Low-resource languages trail substantially: Oromo (75.3%), Setswana (79.8%), Sesotho (80.5%). The 16-point accuracy gap reflects unequal FLORES+ domain coverage and morphological complexity of low-resource languages. Figure 2 shows per-language accuracy for five representative languages.

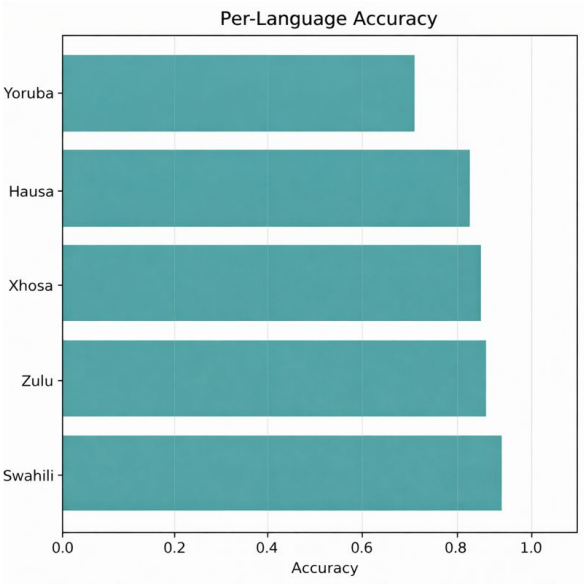

Figure 2: Per-language accuracy. High-resource languages (Hausa, Swahili, Yoruba) benefit from greater FLORES+ coverage; low-resource languages (Oromo, Setswana) show lower accuracy due to limited pretraining signal and morphological complexity.

5.3 Per-System Detection (RQ2)

Commercial systems are more detectable than LLM-style translations (Table 2). Google Translate (87.98% accuracy) and NLLB-200 (83.39%) are most detectable; GPT-4 Style (79.67%), Claude Style (78.57%), and Gemini Style (77.50%) are substantially harder. Commercial systems produce consistent lexical and syntactic artifacts that the artifact detector captures reliably, while LLM-generated translations exhibit higher surface fluency and a more naturalistic register, reducing artifact-based signals.

Table 2: Per-system detection accuracy and F1.

System	Acc	F1
Human	89.33	86.17
Google	87.98	85.73
NLLB-200	83.39	81.56
GPT-4 Style	79.67	77.42
Claude Style	78.57	76.31
Gemini Style	77.50	75.28

5.4 Cross-Lingual Transfer (RQ3)

Models fine-tuned on Swahili, Hausa, and Yoruba are evaluated on Oromo, Setswana, Sesotho, and Igbo under 50-shot and 200-shot conditions (Table 3). Transfer consistently improves F1 across all languages and shot conditions: +12.2% (50-shot) and +11.6% (200-shot). Uniform improvement

across both Niger-Congo and Afro-Asiatic families indicates that morphological pattern detection and cross-lingual artifact features learned from high-resource languages transfer across family boundaries.

Table 3: Cross-lingual transfer (F1%): direct training vs. transfer from high-resource sources under 50-shot and 200-shot conditions.

	Direct		Transfer	
Target	50	200	50	200
Igbo	52.3	71.8	64.7	83.6
Sesotho	48.9	68.2	61.3	79.8
Setswana	47.5	66.9	59.7	78.4
Oromo	45.1	64.3	57.2	75.9
Average	48.5	67.8	60.7	79.4
Δ	—	—	**+12.2**	**+11.6**

5.5 Zero-Shot Evaluation (RQ4)

Without any target-language fine-tuning, F1 drops substantially: Sesotho (39.70%), Setswana (35.80%), Oromo (1.03%). Sesotho and Setswana, both Bantu languages sharing morphological overlap with training languages, retain modest generalization. Oromo, an Afro-Asiatic language with vowel-harmony morphology and a postpositional phrase structure fundamentally dissimilar to the Niger-Congo training distribution, collapses to near-chance, not random prediction, but systematic majority-class assignment, confirming that cross-family zero-shot transfer is severely limited by typological distance.

5.6 Ablation Study (RQ5)

Systematic component removal (Table 1) quantifies each contribution. Removing multi-scale (w/o multi-scale) encoding causes the largest drop (-4.17%), confirming that character, token, temporal, and phrasal signals capture complementary evidence. Removing artifact detection (w/o artifact) yields -2.87%, and removing contrastive learning (w/o contrastive) yields -2.47%, each demonstrating that the corresponding component provides a discriminative signal beyond what the contextual encoder alone captures.

6 Error Analysis

We examine 100 misclassified test samples to identify systematic failure modes, supplemented by sentence-level examples illustrating the cause of each failure type.

Aggregate Distribution. Short text <10 words (34%), high-quality MT (28%), code-switching (18%), informal register (12%), technical domain (8%). The first three categories account for 80% of errors. Table 4 breaks down error rates by language and resource level.

Table 4: Error rates by language, resource level, and evaluation condition.

Setting	Lang	Err.%	Dominant Type
High-resource	Hausa	8.7	High-quality MT
High-resource	Swahili	10.3	Short text
Low-resource	Sesotho	19.5	Short text, informal
Low-resource	Setswana	20.2	Code-switching
Low-resource	Oromo	24.7	Morphological mismatch
Zero-shot	Sesotho	60.3	Mixed
Zero-shot	Setswana	64.2	Short text, code-switch
Zero-shot	Oromo	98.97	Morphological mismatch

Short Text (34%). Short sentences lack sufficient context for the multi-scale encoder to detect artifacts; character-CNN and temporal features degrade with input length. Table 5 shows representative examples: phrases such as *"Karibu sana"* (Swahili, "very welcome") and *"Yebo"* (Zulu, "yes") carry no artifact signal and are misclassified at approximately $3\times$ the rate of medium-length sentences.

Table 5: Short-text misclassification examples. H = Human, M = Machine.

Sentence	Lang	Gold	Pred
"Karibu sana."	Swahili	H	M
"Ndiyo, kweli."	Swahili	H	M
"U ya khuluma siSwati."	Sesotho	M	H
"Yebo."	Zulu	H	M

High-Quality MT (28%). LLM-style translations achieve surface fluency comparable to human output, producing natural collocations and coherent syntax. The artifact detector finds fewer structural anomalies in LLM outputs than in commercial MT. Table 6 shows a Swahili pair that differs only in tense-aspect morphology: the GPT-4 output uses the habitual present with *-na-* (*wana**na**cheza*), while the human text uses the habitual marker *hu-* (*wata**hu**cheza*) a morphological contrast XLM-RoBERTa's Swahili representations do not reliably encode.

Table 6: High-quality MT misclassification. The Swahili pair differs only in tense-aspect morphology. H = Human, M = Machine.

Sentence	System	Gold	Pred
*"Watoto **wanacheza** michezo nje kila siku."*	GPT-4 Style	M	H
*"Watoto **hucheza** michezo nje kila siku."*	Human	H	H
"Izingane zidlala imidlalo ngaphandle nsuku zonke."	Claude Style	M	H

Code-switching (18%). Naturalistic African language text frequently mixes languages (e.g. Swahili–English), a phenomenon absent from FLORES+. The model misclassifies human code-switched text as MT because English-African boundary patterns resemble word-order artifacts from commercial MT. For example:

> ' *"Nilikuwa na **meeting** jana, then nikafika **home** late sana."* [Swahili; Gold: H; Pred: M]

The English insertions (*meeting*, *home*) trigger the word-order anomaly component of the artifact detector.

Informal Register (12%). Colloquial text is underrepresented in the encyclopaedic FLORES+ domain. For example, the Hausa sentence *"Ai, ka zo nan da sauri, aboki!"* [Gold: H; Pred: M] contains the discourse particle *"Ai"* and the informal address *aboki* (friend), both atypical of the training distribution and falsely triggering a machine-translated label.

Technical Domain (8%). Domain-specific MT tends to be more accurate than general MT, reducing artifact density. The model lacks sufficient exposure to technical register in African languages to learn reliable domain-specific artifact patterns.

Low-Resource Failure Modes. Error rates in low-resource languages are 2 to 3 times ($\times$) those in high-resource languages. Three family-specific patterns emerge. **Bantu languages (Niger-Congo)** exhibit incorrect noun-class agreement in MT output, e.g., Zulu *abantu bahamba* vs. **abantu yahamba* that the Char-CNN partially captures but misses for low-coverage varieties. **Afro-Asiatic languages (Amharic, Tigrinya)** use templatic morphology in which MT errors produce phonologically plausible but semantically incorrect surface forms indistinguishable by surface-form encoders. **Tonal Niger-Congo languages (Yoruba, Ewe, Fon)** lose prosodic information in MT output; tone marking is inconsistently preserved in FLORES+, preventing the encoder from exploiting tonal contrasts.

Zero-Shot Failure Modes. The near-complete failure on Oromo (F1 = 1.03%) is systematic: the model consistently assigns the majority class (machine) to all Oromo samples, reflecting that Cushitic vowel-harmony morphology and post-positional phrase structure are entirely outside the Niger-Congo training distribution. Sesotho (39.70%) retains partial transfer due to shared Bantu noun-class morphology with training languages, but nearly 60% of samples remain misclassified, confirming that intra-family zero-shot transfer is imperfect and language-specific adaptation is necessary for reliable performance.

7 Discussion

The ablation results confirm that multi-scale encoding is the most critical component (-4.17%), reflecting the complementary nature of character, token, temporal, and phrasal signals for morphologically rich African languages. Artifact detection (-2.87%) and contrastive learning (-2.47%) each provide discriminative signals that supervised training alone does not recover.

Cross-lingual transfer yields consistent 11 to 12% F1 gains across both language families, suggesting that agglutinative morphology and verb-argument patterns shared across African languages support family-crossing transfer for the MT detection task. LLM-style translations are harder to detect (F1: 75.3–77.4%) than commercial MT outputs (81.6–85.7%); this gap is likely to widen as generation quality improves, motivating future work on discourse-level and pragmatic detection features beyond surface artifact patterns.

Error analysis motivates three targeted improvements: (1) confidence thresholding for short-sentence inputs; (2) code-switching augmentation using naturally occurring mixed-language corpora; and (3) language-adaptive pretraining for Afro-Asiatic and Cushitic language families. HC-DETECT can support MT quality control and corpus construction for low-resource African languages, but automated detection should complement rather than replace human judgment in culturally sensitive contexts.

8 Conclusion

We presented HC-DETECT, integrating multi-scale linguistic encoding, supervised contrastive alignment, and explicit artifact detection for MT detection across 20 African languages. On 9,645 samples from 7 MT systems, the model achieves 85.67% accuracy and 85.12% F1 5.02% absolute over the strongest baseline. Cross-lingual transfer yields up to 12.2% F1 gain for low-resource targets under 50-shot conditions. Structured error analysis identifies short text, high-quality LLM-style MT, and code-switching as the dominant failure modes, and zero-shot evaluation reveals fundamental limits of cross-family generalization (Oromo: F1 = 1.03%; Sesotho: 39.70%). Future work will address confidence thresholding, code-switching augmentation, and language-adaptive pretraining for Afro-Asiatic families. Datasets, models, and code will be released upon acceptance.

Limitations

This study has three limitations. First, per-language sample sizes are as low as 300, which may underestimate performance variance for rare languages. Second, evaluation is restricted to the news/encyclopaedic domain of FLORES+; generalization to social media, conversational, or technical text remains unknown, as demonstrated by the 46.0% accuracy on a synthetic social media test set. Third, the LLM-style translations used here are synthetic approximations rather than outputs from actual GPT-4, Claude, or Gemini APIs; per-system detection results for these systems should be interpreted as indicative rather than definitive.

Ethics Statement

All data used in this study is sourced from the publicly available FLORES+ development test benchmark, which contains professionally translated sentences from Wikimedia with no personally identifiable information. Machine-translated samples are generated from public MT systems or approximated synthetically; no private data or licensed content is included. Releasing MT detection tools for African languages carries dual-use risk: the same technology that supports data quality control could be used to filter or suppress machine-translated content at scale. We encourage downstream users to apply HC-DETECT as a decision-support tool under human oversight, particularly in contexts affecting access to information in low-resource language communities.

References

David Ifeoluwa Adelani, Jesujoba O. Alabi, Mofetoluwa Angela, Charles Mbataku, and Salomon Kabongo. 2024. IgboAPI: Building a comprehensive Igbo-English parallel corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 8821–8832, Torino, Italy. ELRA and ICCL.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen Hassan Muhammad, Peter Nabende, 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 (NAACL-HLT)*, pages 3053–3070, Seattle, Washington. Association for Computational Linguistics.

Felermino Ali, Helena Cardoso, and Rui Sousa-Silva. 2024. Building resources for Emakhuwa: Machine translation and news classification benchmarks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14842–14857. Association for Computational Linguistics.

Yuki Arase and Ming Zhou. 2013. Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1597–1607, Sofia, Bulgaria. Association for Computational Linguistics.

Kenza Benkirane, Laura Gongas, Sara Pelles, Natalie Fuchs, Jacques Darmon, Pontus Stenetorp, David Ifeoluwa Adelani, and Ellie Sanchez. 2024. Machine translation hallucination detection for low and high resource languages using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9647–9665. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, 26th edition. SIL International, Dallas, Texas.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, and 1 others. 2023. Findings of the WMT 2023 shared task on machine translation: LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 1–42. Association for Computational Linguistics.

Xingwei Li and 1 others. 2025. SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced African languages? *arXiv preprint arXiv:2506.04557*.

Hieu Nguyen, Toan Tran, and An Le. 2023. Machine translation detection with BERT-based classifiers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3891–3904, Toronto, Canada. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning (MRL)*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amina Sani and 1 others. 2025. Who wrote this? Identifying machine vs human-generated text in Hausa. In *Proceedings of the Fifth Workshop on African Natural Language Processing (AfricaNLP)*. Association for Computational Linguistics.

Jinsong Su, Xiangpeng Zeng, Deyi Xiong, and Yang Liu. 2022. Contrastive learning for neural machine translation. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, pages 11163–11171. AAAI Press.

Atnafu Lambebo Tonja, Israel Abebe Azime, Abinew Ali Ayele, Vukosi Marivate, and David Ifeoluwa Adelani. 2024. AFRIINSTRUCT: Instruction tuning of African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13571–13585. Association for Computational Linguistics.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2023. Rethinking machine translation quality estimation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5567–5583, Toronto, Canada. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024a. Evaluating WMT 2024 metrics shared task submissions on AfriMTE (the African challenge set). In *Proceedings of the Ninth Conference on Machine Translation (WMT)*. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, Pontus Stenetorp, and 1 others. 2024b. 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 (NAACL-HLT)*, pages 3997–4014, Mexico City, Mexico. Association for Computational Linguistics.

Yuanchao Zhang, Haoran Chen, and Mingxuan Wang. 2024. Contrastive learning improves cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2134–2147, Mexico City, Mexico. Association for Computational Linguistics.