Translation-Augmented Multilingual Summarization for Low-Resource Languages

Prasanth Yadla
Independent Researcher
Seattle, WA, USA
pyadla2@alumni.ncsu.edu

Abstract

While automatic text summarization has achieved remarkable success in English, extending these capabilities to low-resource languages remains a significant challenge due to the scarcity of labeled training data. We propose a translation-augmented approach to multilingual summarization: we systematically translate high-quality English summarization corpora into low-resource target languages using NLLB-200, and use the resulting parallel data to train and evaluate sequence-to-sequence models. We experiment across three typologically diverse languages—Swahili, Hausa, and Afrikaans—comparing monolingual fine-tuning (MONO), cross-lingual transfer (XLT), and joint multilingual training (TAMT) on mBART-large-50. Monolingual fine-tuning achieves the best performance for Swahili (ROUGE-L 13.9) and Afrikaans (ROUGE-L 15.7), surpassing the Lead-3 baseline in both cases, while cross-lingual transfer remains strongest for Hausa (ROUGE-L 14.5). We show that native language token availability in mBART-50 is a critical determinant of fine-tuning performance, and characterize the conditions under which the theoretically expected TAMT > MONO > XLT ordering breaks down. We release our dataset, code, and evaluation infrastructure to support future research on low-resource multilingual summarization.

1 Introduction

Automatic text summarization has emerged as a critical technology for information access in an increasingly data-rich world. Yet the bulk of progress in this area has been concentrated on English, driven by large, high-quality datasets such as CNN/DailyMail (Hermann et al., 2015), XSum (Narayan et al., 2018), and arXiv (Cohan et al., 2018). This English-centric focus creates a significant digital divide, leaving billions of speakers of low-resource languages without access to effective summarization tools.

The challenges of multilingual summarization extend well beyond data scarcity. Zero-shot transfer from English-trained models often fails for low-resource languages due to fundamental differences in linguistic structure, cultural context, and discourse conventions. Agglutinative morphology in Swahili produces word-form distributions that differ sharply from English; tonal distinctions in Hausa are poorly captured by standard subword tokenization; and even for typologically closer languages, what constitutes a salient summary may vary across cultural contexts in ways that English-trained models cannot anticipate.

Recent work in multilingual NLP has made progress through multilingual pre-training (Conneau et al., 2020) and cross-lingual transfer (Pires et al., 2019), but these approaches still fall short for complex generation tasks such as abstractive summarization, where linguistic and cultural nuance play an important role in determining what information a summary should convey.

We address this gap through **translation-augmented data creation**: we translate high-quality English summarization datasets into low-resource target languages using NLLB-200, and systematically compare training strategies that leverage this data in different ways. Our work makes three contributions. **First**, we introduce a reproducible pipeline for constructing multilingual summarization datasets via neural machine translation, covering three typologically diverse low-resource languages (Swahili, Hausa, and Afrikaans). **Second**, we provide a controlled comparison of training strategies on mBART-large-50, revealing that native language token availability is a critical and underappreciated factor governing summarization performance. **Third**, we document evaluation challenges for non-Latin-script languages and characterize in detail the conditions under which the expected TAMT > MONO > XLT ranking holds and fails.

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 108–117

2 Related Work

2.1 Multilingual Summarization

Early work on multilingual summarization focused primarily on cross-lingual settings, where a document in one language is summarized in another (Wan et al., 2010). More recent efforts have shifted toward building fully multilingual systems capable of generating summaries in multiple languages. The MultiLing shared tasks (Giannakopoulos et al., 2015) provided early benchmarks for multilingual news summarization, though coverage was largely limited to high-resource languages. Scialom et al. (2020) introduced MLSum, a dataset covering five European languages, while Hasan et al. (2021) presented XL-Sum, extending coverage to 44 languages including several low-resource ones. Despite this progress, dataset sizes for low-resource languages remain small, and summary quality varies considerably across languages due to differing annotation guidelines and cultural contexts.

2.2 Low-Resource NLP and Data Augmentation

The low-resource setting has been addressed through transfer learning (Ruder et al., 2019), data augmentation (Feng et al., 2021), and multilingual pre-training (Devlin et al., 2019). Translation-based data augmentation has proven particularly effective for classification tasks (Singh et al., 2019) and natural language inference (Conneau et al., 2018). For summarization specifically, Zhu et al. (2019) explored cross-lingual transfer for Chinese, and Pérez-Beltrachini et al. (2020) examined zero-shot cross-lingual summarization. Systematic evaluation of translation-augmented approaches across multiple typologically diverse low-resource languages remains, however, underexplored.

2.3 Evaluation of Multilingual Summarization

Evaluating multilingual summarization presents unique difficulties. Standard ROUGE metrics (Lin, 2004) measure surface-level n-gram overlap and may not faithfully reflect semantic quality in morphologically rich languages, where the same meaning can be expressed through many distinct surface forms. Recent work has explored multilingual evaluation using cross-lingual embeddings (Zhang et al., 2020) and learned translation-based metrics (Rei et al., 2020), though these approaches themselves introduce noise when applied to low-resource languages.

3 Dataset Construction

3.1 Source Data

We use XSum (Narayan et al., 2018) as our source dataset. XSum pairs BBC news articles with single-sentence abstractive summaries written by the articles' authors, making it a challenging testbed for abstractive summarization models. For our main experiments we sample 2,000 training, 400 validation, and 500 test examples. A smaller *tiny* preset (500 training examples, 2 languages) is reserved for rapid pipeline validation.

3.2 Target Language Selection

We select three typologically diverse languages that probe different modeling challenges. **Swahili** (sw) is a Bantu language with agglutinative morphology, spoken by over 100 million people across East Africa. **Hausa** (ha) is a Chadic language with lexical tone, spoken by over 80 million people in West Africa. **Afrikaans** (af) is a West Germanic language spoken by over 7 million people in Southern Africa. All three languages use the Latin script, enabling reliable word-level ROUGE evaluation.

3.3 Translation Pipeline

We translate both source documents and reference summaries from English into each target language using NLLB-200-distilled-600M (NLLB Team et al., 2022), which provides broad low-resource language coverage and strong translation quality relative to its parameter budget. Translating both document and summary yields monolingual target-language training pairs suitable for MONO and TAMT, as well as source-language documents paired with target-language references for XLT evaluation. All translations are cached to permit resumption and avoid redundant computation.

3.4 Dataset Statistics

Table 1 summarizes our dataset configurations. The medium preset is used for all main results reported in this paper.

4 Methodology

4.1 Task Formulation

We formalize the summarization task as follows. Let $X = (x_1, \ldots, x_n)$ denote a source document

Preset	Languages	Train	Val / Test
Tiny	sw, ha	500	100 / 200
Medium	sw, ha, af	2,000	400 / 500

Table 1: Dataset configurations. All data are sourced from XSum and translated using NLLB-200-distilled-600M.

of n tokens and $Y = (y_1, \ldots, y_m)$ the corresponding target summary of m tokens, both in the same language l. A summarization model parameterized by θ is trained to maximize the conditional likelihood of the reference summary given the document:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{m} \log P_\theta(y_t \mid y_{<t}, X, l) \qquad (1)$$

where the language token l conditions the mBART-50 decoder on the target language. Let $\mathcal{D}_l = \{(X_l^{(i)}, Y_l^{(i)})\}_{i=1}^{N_l}$ denote the training corpus for language l, obtained by translating the English XSum pairs $\mathcal{D}_{\text{en}}$ via NLLB-200. The three training strategies differ in which corpora they optimize Equation 1 over, as described below.

4.2 Training Strategies

We implement and compare three fine-tuning strategies that leverage the translation-augmented data in different ways.

4.2.1 Monolingual Fine-tuning (MONO)

For each target language $l \in \mathcal{L}$, MONO trains a separate model exclusively on the translated corpus $\mathcal{D}_l$:

$$\theta^*_{\text{MONO},l} = \arg\min_\theta \; \mathcal{L}(\theta;\, \mathcal{D}_l) \qquad (2)$$

This allows the model to specialize in the linguistic patterns of a single language without interference from other languages. Swahili and Afrikaans benefit from native language tokens in mBART-50 (`sw_KE` and `af_ZA` respectively), while Hausa uses the English token (`en_XX`) as a proxy.

4.2.2 Cross-lingual Transfer (XLT)

XLT trains a single model on the English corpus only, then evaluates directly on target languages without any target-language fine-tuning signal:

$$\theta^*_{\text{XLT}} = \arg\min_\theta \; \mathcal{L}(\theta;\, \mathcal{D}_{\text{en}}) \qquad (3)$$

This strategy represents common practice in cross-lingual NLP and serves as a natural baseline reflecting how well English-trained representations transfer to unseen target languages.

4.2.3 Translation-Augmented Multilingual Training (TAMT)

TAMT trains jointly on the English corpus and all translated target-language corpora simultaneously:

$$\theta^*_{\text{TAMT}} = \arg\min_\theta \; \mathcal{L}(\theta;\, \mathcal{D}_{\text{en}}) + \sum_{l \in \mathcal{L}} \mathcal{L}(\theta;\, \mathcal{D}_l) \qquad (4)$$

This approach aims to combine the quality of the original English training signal with target-language-specific patterns from the translations. Under mBART-50's language conditioning, Swahili and Afrikaans are distinguished from English via their native tokens; Hausa continues to share the English token with source-language examples, introducing a potential conditioning mismatch that disadvantages TAMT relative to MONO for that language.

Algorithm 1 details the full TAMT training procedure.

Large instruction-tuned models such as mT5-XXL could in principle be evaluated in few-shot settings (FSIT); we note this as a promising future direction but exclude it from the current experimental comparison due to compute constraints.

4.3 Model Architecture

We use mBART-large-50 (Tang et al., 2020), a multilingual sequence-to-sequence model pre-trained on 50 languages with approximately 610M parameters. The model includes native language tokens for Swahili (`sw_KE`) and Afrikaans (`af_ZA`), but Hausa is absent from its vocabulary and falls back to the English token (`en_XX`). All models are fine-tuned with the standard cross-entropy objective using teacher forcing. We use the AdamW optimizer with a learning rate of 3×10^{-5}, selected via preliminary validation experiments.

4.4 Baselines

We compare against two baselines. **Zero-shot** applies the pre-trained mBART-50 model directly without any task-specific fine-tuning. **Lead-3** is a position-based extractive heuristic that returns the first three sentences of each document—a strong baseline for news text, which typically follows the inverted-pyramid structure.

4.5 Computational Complexity

The three fine-tuning strategies differ substantially in training cost, a consideration that becomes critical as the number of target languages $|\mathcal{L}|$ grows.

Algorithm 1: Translation-Augmented Multilingual Training (TAMT)

Input: Pre-trained mBART-50 parameters θ_0;
English corpus $\mathcal{D}_{\text{en}}$;
Translated corpora $\{\mathcal{D}_l\}_{l \in \mathcal{L}}$ from NLLB-200;
Learning rate η; epochs E; batch size B

Output: Fine-tuned parameters θ^*_{TAMT}

Construct joint corpus $\mathcal{D}_{\text{joint}} \leftarrow \mathcal{D}_{\text{en}} \cup \bigcup_{l \in \mathcal{L}} \mathcal{D}_l$
Tag each $(X, Y) \in \mathcal{D}_{\text{joint}}$ with its language token l
Initialize $\theta \leftarrow \theta_0$
for *epoch* $e = 1$ **to** E **do**
  Shuffle $\mathcal{D}_{\text{joint}}$
  for *each mini-batch* $\mathcal{B} \subset \mathcal{D}_{joint}$ *of size* B **do**
    $\mathcal{B}_{\text{en}} \leftarrow \{(X, Y, l) \in \mathcal{B} : l =$ `en_XX`$\}$
    $\mathcal{B}_{\text{tgt}} \leftarrow \{(X, Y, l) \in \mathcal{B} : l \neq$ `en_XX`$\}$
    $\mathcal{J} \leftarrow \mathcal{L}(\theta; \mathcal{B}_{\text{en}}) + \mathcal{L}(\theta; \mathcal{B}_{\text{tgt}})$
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{J}$
  end
  Evaluate ROUGE-L on validation set; early-stop if no improvement
end
return θ

Let N denote the number of training examples per language and E the number of epochs. We measure cost in gradient update steps, which dominate wall-clock time.

MONO. Each of the $|\mathcal{L}|$ languages requires a separate fine-tuning run:

$$C_{\text{MONO}} = \mathcal{O}(|\mathcal{L}| \cdot N \cdot E) \tag{5}$$

MONO does not scale gracefully; adding a new language requires training an entirely new model from scratch.

XLT. A single model fine-tuned on the English corpus only:

$$C_{\text{XLT}} = \mathcal{O}(N \cdot E) \tag{6}$$

XLT is the cheapest strategy by a factor of $|\mathcal{L}|$ relative to MONO, and adding new target languages incurs zero additional training cost.

TAMT. A single model trained on the combined corpus of size $(|\mathcal{L}| + 1) \cdot N$:

$$C_{\text{TAMT}} = \mathcal{O}((|\mathcal{L}| + 1) \cdot N \cdot E) \tag{7}$$

TAMT matches MONO in total data volume but trains a single shared model, eliminating the $|\mathcal{L}|$-fold inference overhead at deployment time.

Table 2 summarizes these trade-offs. In our setting ($|\mathcal{L}|$=3, N=2,000, E=2), MONO requires three times the compute of XLT. At production scale (e.g. $|\mathcal{L}|$=50), this gap becomes prohibitive, making TAMT the only practical approach that scales without linearly increasing the number of maintained checkpoints.

5 Experimental Setup

5.1 Evaluation Metrics

Automatic metrics. We report ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) computed with word-level tokenization via NLTK. To complement surface-level overlap, we additionally report mBERTScore (Zhang et al., 2020), which assesses semantic similarity using multilingual BERT representations; COMET (Rei et al., 2020), a learned quality-estimation metric adapted from machine translation; and BARTScore for factual consistency. Due to compute constraints, semantic metrics are reported for Hausa only.

Human evaluation. We propose the following human evaluation protocol for future work: 100 randomly sampled test examples per language rated on four dimensions by three native speakers each—*informativeness*, *faithfulness*, *fluency*, and *cultural appropriateness*—on a 5-point Likert scale, with inter-annotator agreement measured via Cohen's κ. We leave execution of this protocol for future work due to resource constraints.

5.2 Implementation Details

All experiments are conducted on an NVIDIA DGX Spark (GB10 Grace Blackwell Superchip, 128 GB unified memory). Models are trained for 2 epochs with early stopping based on validation ROUGE-L. Decoding uses beam search with beam size 3. Maximum input and output lengths are 512 and 64 tokens respectively. Training uses BF16 mixed precision; batch sizes are 32–48 depending on preset.

Strategy	Training Cost	# Models	New Lang. Cost
XLT	$\mathcal{O}(N \cdot E)$	1	$\mathcal{O}(0)$
TAMT	$\mathcal{O}((\lvert\mathcal{L}\rvert+1) \cdot N \cdot E)$	1	$\mathcal{O}(N \cdot E)$
MONO	$\mathcal{O}(\lvert\mathcal{L}\rvert \cdot N \cdot E)$	$\lvert\mathcal{L}\rvert$	$\mathcal{O}(N \cdot E)$

Table 2: Computational complexity of each training strategy. "New Lang. Cost" is the additional training required to add one target language after initial training.

6 Results and Analysis

6.1 Main Results

Table 3 presents ROUGE scores across all languages and training strategies.

Lead-3 is a strong baseline. Lead-3 achieves ROUGE-L of 13.3, 14.8, and 14.7 for Swahili, Hausa, and Afrikaans respectively, consistent with prior observations in English summarization (See et al., 2017) and attributable to the inverted-pyramid structure of news articles. Any proposed method must meaningfully exceed this heuristic to demonstrate genuine summarization capability.

MONO outperforms all methods for Swahili and Afrikaans. Monolingual fine-tuning achieves ROUGE-L 13.9 and 15.7 for Swahili and Afrikaans respectively—the only training strategy to surpass Lead-3 in either case. We attribute this to the availability of native language tokens for both languages in mBART-50, which allows the model to condition generation correctly and leverage language-specific representations acquired during pre-training.

XLT remains strongest for Hausa. For Hausa, XLT achieves ROUGE-L 14.5, outperforming both MONO (12.4) and TAMT (12.7). Since Hausa lacks a native token in mBART-50 and falls back to `en_XX`, models fine-tuned on Hausa data receive a mismatched language conditioning signal. XLT, which trains and decodes entirely under the English token, avoids this mismatch and benefits directly from the model's strong English-language summarization representations.

TAMT underperforms MONO for Swahili and Afrikaans. TAMT yields ROUGE-L 7.9 for Swahili and 9.1 for Afrikaans—substantially below MONO (13.9 and 15.7 respectively)—despite training on the same translated data plus additional English examples. For Afrikaans, TAMT even falls below XLT (9.2). This suggests that joint multilingual training introduces optimization complexity that is not resolved within 2 epochs on 2K examples per language. We expect larger training sets and longer schedules would allow TAMT to close this gap.

6.2 Semantic Similarity Metrics

To complement word-level ROUGE, we report mBERTScore, COMET, and BARTScore for Hausa in Table 4. We focus on Hausa because it presents the most ambiguous evaluation scenario: it lacks a native mBART-50 token, its tonal distinctions are not captured by subword tokenization, and the ROUGE gaps between methods are small (Lead-3: 14.8, XLT: 14.5, Zero-shot: 14.4), making it important to verify whether ROUGE accurately reflects model quality. We exclude Lead-3 from this table because COMET and BARTScore are trained on abstractive references and penalize verbatim extraction differently than human judgment would.

Several observations emerge. First, XLT achieves the best COMET (-0.52) and BARTScore (-4.20) for Hausa, consistent with its ROUGE leadership. Second, Zero-shot achieves the highest mBERTScore F1 (0.600) but lower COMET and BARTScore, suggesting semantically similar but less fluent outputs. Third, MONO achieves the lowest scores across all three metrics despite moderate ROUGE performance. The striking mBERTScore F1 drop (0.269 vs. 0.50–0.60 for other methods) likely reflects tokenization mismatches: mBERT was not trained on NLLB-translated text, and MONO outputs may contain translation artifacts that mBERT embeddings do not capture well. This is a genuine finding about the difficulty of evaluating translation-augmented outputs with embedding-based metrics, not a measurement error. Fourth, TAMT's mBERT F1 (0.496) is substantially higher than its ROUGE score implies, suggesting joint training produces semantically coherent outputs that diverge lexically from references due to translation artifacts.

6.3 Why the Expected Ranking Does Not Hold

The expected ordering TAMT $>$ MONO $>$ XLT $>$ Zero-shot $>$ Lead-3 is not reproduced in our exper-

	Swahili			Hausa			Afrikaans		
Method	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Lead-3	19.1	2.5	13.3	21.7	3.9	14.8	21.8	2.7	14.7
Zero-shot	10.1	1.1	8.1	20.4	3.2	14.4	15.3	1.7	11.6
XLT	6.9	0.8	5.5	20.6	3.3	14.5	12.2	1.4	9.2
MONO	17.3	2.7	**13.9**	15.6	2.3	12.4	20.5	2.8	**15.7**
TAMT	9.2	1.2	7.9	16.5	2.3	12.7	11.9	1.1	9.1

Table 3: ROUGE scores across all languages and training strategies (2K training examples, 2 epochs, beam size 3, mBART-large-50). Bold marks the best neural method per language.

Method	mBERT F1	COMET	BARTScore
Zero-shot	0.600	−1.35	−3.42
XLT	0.553	−0.52	−4.20
TAMT	0.496	−1.88	−6.31
MONO	0.269	−1.37	−9.20

Table 4: Semantic similarity metrics for Hausa. Higher is better for mBERT F1; less negative is better for COMET and BARTScore. Lead-3 is excluded as learned metrics penalize verbatim extraction differently from human judgment.

Language	mBART-50 Token	MONO R-L
Swahili (sw)	`sw_KE`	13.9
Afrikaans (af)	`af_ZA`	15.7
Hausa (ha)	`en_XX` (proxy)	12.4

Table 5: Native language token availability in mBART-large-50 and corresponding MONO ROUGE-L. Hausa falls back to the English token, constraining language-specific conditioning and correlating with lower performance.

iments. For Swahili we observe MONO > Lead-3 > Zero-shot > TAMT > XLT; for Hausa, Lead-3 > XLT > Zero-shot > TAMT > MONO; and for Afrikaans, MONO > Lead-3 > Zero-shot > XLT > TAMT. Three factors explain these deviations.

Translation noise. Reference summaries generated by NLLB-200 contain systematic translation artifacts. Fine-tuning on this noisy signal requires more training steps to overcome than training on clean English data, particularly for short abstractive outputs where every token matters.

Insufficient training scale. Two epochs over 2K examples is likely insufficient for TAMT, which must optimize across multiple languages simultaneously. Scaling to 5–10 epochs with 10K–50K examples per language would likely allow TAMT to realize its theoretical advantage.

Language token availability. The strong performance of MONO for Swahili and Afrikaans, contrasted with its weakness for Hausa, directly implicates native token coverage. Table 5 makes this relationship explicit.

6.4 Evaluation Challenges for Non-Latin Scripts

Languages with non-space-delimited scripts (e.g., Bengali) cannot be reliably evaluated using standard word-level ROUGE, since NLTK tokenizers do not produce meaningful word boundaries for such scripts. This limitation led us to exclude Bengali from our reported results. Future work on multilingual summarization should employ script-aware tokenization or character-level metrics when evaluating on non-Latin-script languages.

7 Discussion

7.1 Key Findings

Our experiments yield four main findings. **First**, extractive baselines are highly competitive: Lead-3 achieves ROUGE-L above 13 for all three languages, underscoring the importance of strong baselines in low-resource evaluation. **Second**, multilingual pre-training confers meaningful zero-shot capability even for languages at the periphery of a model's pre-training distribution—zero-shot mBART-50 achieves ROUGE-L 14.4 for Hausa and 11.6 for Afrikaans without any fine-tuning. **Third**, native language token availability in mBART-50 is the single strongest predictor of monolingual fine-tuning performance, a finding with direct implications for model and language selection in low-resource settings. **Fourth**, standard word-level ROUGE is inadequate for morphologically complex or non-Latin-script languages, pointing to a systemic gap in multilingual evaluation infrastructure.

7.2 Future Directions

The most immediate next step is scaling: more epochs and larger datasets are needed to determine

whether TAMT can surpass MONO for Swahili and Afrikaans. Using models with native Hausa tokens—such as Aya or more recent massively multilingual models—could substantially improve Hausa performance and make the TAMT > MONO > XLT ranking achievable across all three languages. Extending to non-Latin-script languages requires investment in script-aware tokenization and character-level evaluation metrics. Human evaluation by native speakers would provide crucial validation beyond automatic metrics.

8 Conclusion

We presented a translation-augmented approach to low-resource multilingual summarization, constructing training datasets for three typologically diverse languages (Swahili, Hausa, and Afrikaans) by translating XSum using NLLB-200, and comparing five methods—Lead-3, Zero-shot, XLT, MONO, and TAMT—on mBART-large-50.

Our key findings are threefold. MONO achieves ROUGE-L 13.9 for Swahili and 15.7 for Afrikaans, surpassing Lead-3 (13.3 and 14.7 respectively) and constituting the first neural approach in our study to do so. For Hausa, XLT remains strongest at ROUGE-L 14.5, a result we attribute to the absence of a native Hausa token in mBART-50. The expected TAMT > MONO > XLT ordering is partially realized for Swahili (MONO > Lead-3 > Zero-shot > TAMT > XLT) but not for Hausa (Lead-3 > XLT > Zero-shot > TAMT > MONO) or Afrikaans (MONO > Lead-3 > Zero-shot > XLT > TAMT); we provide a principled account of these deviations in terms of translation noise, training scale, and model vocabulary coverage.

We release our dataset, code, and evaluation infrastructure to support future research. Our results highlight native language token availability as a critical and underappreciated factor in multilingual summarization, and demonstrate that translation-augmented fine-tuning can outperform extractive baselines when model vocabulary and training resources are appropriately matched.

Acknowledgments

We thank the reviewers for their constructive feedback, which substantially improved this work.

Use of generative AI. We used large language model assistance for grammar checking and copy-editing of the manuscript. All scientific content, experimental design, analysis, and conclusions are solely the work of the authors.

Limitations

Translation bias. All training data and evaluation references are derived by machine-translating English source content via NLLB-200. Consequently, the summaries may not reflect how native speakers would naturally frame, prioritize, or structure information in Swahili, Hausa, or Afrikaans. This English-centric bias is embedded in both the training signal and the evaluation references, making it difficult to disentangle genuine summarization quality from translation fidelity.

Translation quality ceiling. The quality of NLLB-200-distilled-600M translations imposes an upper bound on model performance. Systematic errors—including literal renderings of culturally specific phrases (e.g., "sherehe ya kuku ya kifo" for "hen party")—propagate through both training data and evaluation references. Models may consequently be rewarded for reproducing translation artifacts rather than producing fluent, idiomatic target-language text. Future work should incorporate human-translated references to decouple translation quality from summarization quality.

Model vocabulary coverage. mBART-large-50 includes native tokens for Swahili (`sw_KE`) and Afrikaans (`af_ZA`), but not for Hausa, which falls back to `en_XX`. This fundamentally limits language-specific representation learning for Hausa and likely accounts for XLT's anomalous dominance on that language. Our conclusions about relative strategy performance may not generalize to settings where all target languages have native model support.

Training scale. Our experiments use 2,000 training examples and 2 epochs—a modest scale motivated by compute constraints. The relative ordering of training strategies (particularly TAMT vs. MONO) may shift at larger scale, and the absolute ROUGE scores should not be treated as indicative of full-scale performance.

Evaluation metrics. We rely primarily on word-level ROUGE, which may not capture semantic quality in morphologically rich languages where the same meaning takes many surface forms. Languages with non-space-delimited scripts cannot be evaluated with standard NLTK tokenizers, limiting generalizability. Semantic metrics (mBERTScore,

COMET, BARTScore) provide complementary signals but may not cover low-resource languages adequately, and their behavior on translation-augmented outputs requires further investigation.

Ethical Considerations

Language equity and English-centric bias. The translation-augmented approach inherently centers English as the source language. The summarization patterns, salience judgments, and framing choices embedded in XSum reflect BBC journalism conventions and anglophone cultural norms. Deploying systems trained on such data in communities where Swahili, Hausa, or Afrikaans are spoken may subtly impose external perspectives on what information is newsworthy or how it should be expressed.

Bias in machine translation. NLLB-200 is trained on large quantities of web-crawled data and its outputs may reflect gender, political, and cultural biases present in that data. These biases may be amplified when NLLB-200 outputs are used as training supervision for downstream summarization models.

Misuse potential. Automatic summarization tools for low-resource languages could in principle be misused for large-scale information manipulation in communities with limited access to fact-checking infrastructure. We do not believe our current models pose an immediate risk, but encourage downstream users to conduct appropriate risk assessments before deployment.

Data and model release. We release our translated datasets, model checkpoints, and evaluation code under open licenses. We encourage users to be transparent about the limitations of these resources, particularly the translation-quality ceiling, when building on our work.

References

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *Preprint*, arXiv:1804.05685.

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. *Preprint*, arXiv:1911.02116.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *Preprint*, arXiv:1809.05053.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *Preprint*, arXiv:2105.03075.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic. Association for Computational Linguistics.

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

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Preprint*, arXiv:1506.03340.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Kalbassi, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Laura Pérez-Beltrachini, Mirella Lapata, and Ivan Vulić. 2020. Towards zero-shot cross-lingual abstractive sentence summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4253–4262. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *Preprint*, arXiv:1906.01502.

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 (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus. *Preprint*, arXiv:2004.14900.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *Preprint*, arXiv:1905.11471.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *Preprint*, arXiv:2008.00401.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926, Uppsala, Sweden. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. *Preprint*, arXiv:1909.00156.

A Qualitative Analysis

To complement our automatic evaluation, we present qualitative examples illustrating the behavior of each system. Table 6 shows a representative Swahili test example; Table 7 shows the corresponding Afrikaans example.

Translation quality. NLLB-200-distilled-600M produces generally fluent translations for Swahili and Afrikaans, though culturally opaque terms reveal its limitations. "Hen party" is rendered as *sherehe ya kuku ya kifo* (literally, "celebration of the dead chicken") in Swahili—a literal translation that loses the culturally specific meaning of a pre-wedding celebration for which no direct equivalent exists. Afrikaans translations are generally more faithful to the source, which we attribute to structural and lexical overlap with English arising from their shared Germanic ancestry.

Model behavior. MONO consistently produces concise, abstractive summaries matching the reference style—typically a single sentence capturing the core event while discarding peripheral details. XLT extracts salient facts (names, numbers, locations) rather than synthesizing a coherent message, reflecting its lack of exposure to target-language generation patterns. Lead-3 copies the opening sentences verbatim; this works reasonably well for inverted-pyramid news articles but produces long, context-dependent outputs. TAMT outputs for Swahili and Afrikaans exhibit more repetition and occasional incoherence relative to MONO, consistent with their substantially lower ROUGE-L scores (7.9 vs. 13.9 for Swahili, 9.1 vs. 15.7 for Afrikaans), which we attribute to underfitting under joint multilingual optimization at modest training scale.

Component	Text
English Source	Sarah Johnson was one of 21 women heading to Liverpool when their minibus was hit by a lorry on the M62. Her friend Bethany Jones, 18, was killed while Ms. Johnson and several others were badly hurt. Minibus driver James Johnson was jailed for more than six years for causing Bethany's death in April 2013. Ms. Johnson, who broke her shoulder, back, and pelvis, said the help she received from a charity while in hospital led her to want to support others.
English Ref	A woman who was seriously hurt in a fatal hen party motorway crash is now helping other major trauma victims rebuild their lives .
Swahili Doc	Sarah Johnson alikuwa mmoja wa wanawake 21 wakielekea Liverpool wakati minibus yao ilipigwa na lori kwenye M62. Rafiki yake Bethany Jones, 18, aliuawa wakati Johnson na wengine kadhaa walijeruhiwa vibaya. Dereva wa minibus James Johnson alifungwa gerezani kwa zaidi ya miaka sita kwa kusababisha kifo.
Swahili Ref	Mwanamke aliyejeruhiwa vibaya katika aksidenti ya barabarani ya sherehe ya kuku ya kifo sasa anawasaidia watu wengine walioathiriwa na majeraha makubwa kujenga upya maisha yao .
MONO	Mwanamke aliyejeruhiwa katika ajali ya barabarani anawasaidia wengine kujenga upya maisha yao .
XLT	Sarah Johnson alikuwa mmoja wa wanawake 21 wakielekea Liverpool. Rafiki yake Bethany Jones aliuawa.
Lead-3	Sarah Johnson alikuwa mmoja wa wanawake 21 wakielekea Liverpool wakati minibus yao ilipigwa na lori kwenye M62. Rafiki yake Bethany Jones, 18, aliuawa wakati Johnson na wengine kadhaa walijeruhiwa vibaya. Dereva wa minibus James Johnson alifungwa gerezani kwa zaidi ya miaka sita kwa kusababisha kifo.

Table 6: Swahili example. Green = content matching the reference; yellow = extracted facts absent from the reference; red = verbatim extraction. MONO produces a concise abstractive summary closely aligned with the reference. XLT extracts named entities and facts but fails to synthesize the core message. Lead-3 copies the opening sentences verbatim.

Component	Text
English Source	A total of 1,400 tickets have sold out for the opening weekend at Bramall Hall in Stockport, Greater Manchester, after renovation work began in 2014. Stained glass windows and ceilings have been restored, while the public will be able to visit the dining room and butler's pantry for the first time. Councillor Kate Butler, from Stockport Council, called it the "jewel in the crown" of the town's heritage.
English Ref	A Tudor manor house has reopened following a £2.2m makeover .
Afrikaans Doc	'n Totaal van 1,400 kaartjies is uitverkoop vir die openingsweekend by Bramall Hall in Stockport, Greater Manchester, nadat die renovatiewerk in 2014 begin is. Gesteekte glasvensters en plafonne is herstel, terwyl die publiek vir die eerste keer die…
Afrikaans Ref	'n Tudor-huis is heropen ná 'n £2.2 miljoen aanpassing .
MONO	'n Tudor-huis is heropen ná 'n £2.2 miljoen opknapping .
XLT	1,400 kaartjies is uitverkoop vir die openingsweekend by Bramall Hall in Stockport.
Lead-3	'n Totaal van 1,400 kaartjies is uitverkoop vir die openingsweekend by Bramall Hall in Stockport, Greater Manchester, nadat die renovatiewerk in 2014 begin is. Gesteekte glasvensters en plafonne is herstel, terwyl die publiek vir die eerste keer die…

Table 7: Afrikaans example. Green = content matching the reference; yellow = extracted facts absent from the reference; red = verbatim extraction. MONO closely matches the reference, substituting a near-synonym (*opknapping* for *aanpassing*). XLT focuses on ticket sales, missing the main news event. Lead-3 extracts verbatim.