

A Comparative Evaluation of English-Sango Machine Translation Systems: LoRA-Based Adaptation and Out-of-Domain Performance for an Extremely Low-Resource Language

Rakhun Kim*, Axel Deo Gratias Ndema Yetomane**

*Lecturer, Dept. of English Education, Hongik University, Seoul, Korea

**Student, Dept. of English and Cultural Studies, Konkuk University, Chungju, Korea

[Abstract]

Sango is an extremely low-resource language with limited machine translation resources. This study compares six English-to-Sango translation systems and examines the out-of-domain effects of LoRA-based adaptation. The systems were evaluated on the complete FLORES-200 devtest set using BLEU, chrF₂, and TER, supplemented by native-speaker judgments on a subset. Google Translate achieved the strongest overall performance, followed by zero-shot NLLB-200. All three LoRA-adapted conditions showed significantly lower BLEU scores than the zero-shot baseline in paired-bootstrap comparisons. A quantity- and step-matched comparison further showed that combining pivot-generated data, back-translation, and orthographic normalization outperformed an equal-sized condition using additional directly mined data, but did not recover zero-shot performance. Exploratory native-speaker evaluations also identified substantial semantic distortions, including unrelated religious-domain content, in some adapted outputs. Importantly, because the present experiments were conducted under a relatively small-scale adaptation setting, the observed pattern should not be assumed to generalize to substantially larger fine-tuning datasets. Overall, the results indicate that fine-tuning on automatically mined data may reduce out-of-domain performance in extremely low-resource MT, underscoring the importance of strong zero-shot baselines, independent evaluation benchmarks, and native-speaker validation.

► **Key words:** Sango, low-resource machine translation, LoRA-based fine-tuning, out-of-domain evaluation

-
- First Author: Rakhun Kim, Corresponding Author: Rakhun Kim
 - *Rakhun Kim (kimrhee02@hongik.ac.kr), Dept. of English Education, Hongik University
 - **Axel Deo Gratias Ndema Yetomane (ndemadeo@gmail.com), Dept. of English and Cultural Studies, Konkuk University
 - Received: 2026. 07. 24, Revised: 2026. 08. 19, Accepted: 2026. 08. 20.

[요 약]

산고어(Sango)는 기계번역 자원이 매우 제한적인 극저자원 언어이다. 본 연구는 6개의 영어-산고어 번역 시스템을 비교하고, LoRA 기반 적응이 도메인 외(out-of-domain) 성능에 미치는 영향을 분석하였다. 각 시스템은 FLORES-200 devtest 전체 세트를 대상으로 BLEU, chrF₂, 및 TER,를 사용하여 평가하였으며, 일부 문장에 대해서는 산고어 원어민 평가를 추가로 실시하였다. Google Translate가 전반적으로 가장 높은 성능을 보였고, zero-shot NLLB-200이 그 뒤를 이었다. 세 가지 LoRA 적응 조건은 모두 대응 부트스트랩 검정을 비교한 결과, zero-shot 기준선보다 유의하게 낮은 BLEU 점수를 보였다. 또한 데이터 양과 학습 단계를 통제된 비교에서는 피벗 생성 데이터, 역번역, 철자 정규화를 결합한 조건이 동일한 규모의 추가 직접 수집 데이터를 사용한 조건보다 높은 성능을 보였으나, zero-shot 모델의 성능에는 도달하지 못했다. 탐색적 원어민 평가에서는 일부 적응 모델의 출력에서 관련 없는 종교적 내용이 포함되는 등 상당한 의미 왜곡도 확인되었다. 중요하게도, 본 연구의 실험은 비교적 소규모의 적응 설정에서 수행되었으므로, 여기서 관찰된 결과가 훨씬 더 큰 규모의 파인튜닝 데이터셋에도 그대로 일반화된다고 보기는 어렵다. 종합하면, 자동 수집 데이터에 기반한 파인튜닝은 극저자원 기계번역에서 도메인 외 성능을 저하시킬 수 있으며, 강력한 zero-shot 기준선, 독립적인 평가 벤치마크, 그리고 원어민 검증의 중요성을 보여준다.

▶ **주제어:** 산고어, 저자원언어 기계번역, LoRA 기반 파인튜닝, 도메인 외 평가

I. Introduction

Machine translation (MT) for low-resource languages can promote linguistic inclusion, expand access to information, and strengthen the digital representation of underserved communities [1], [2]. These concerns are especially salient in Africa, where extensive linguistic diversity coexists with limited parallel corpora, computational resources, and standardized evaluation data [2]. Sango is the national language and principal lingua franca of the Central African Republic, whereas French remains prominent in administration, education, and written communication [3]. Despite its communicative importance, Sango remains under-resourced in language technology, limiting both MT development and reliable evaluation.

The range of Sango translation systems has recently expanded. NLLB-200 provides multilingual support for Sango [4], and Google Translate added the language in June 2024 during an expansion covering 110 languages [5], [6]. More targeted initiatives include SangoAI, which supplies curated Sango lexical information to a general-purpose large language model (LLM) through

inference-time prompting [7], and a French-Sango system adapted from NLLB-200 using automatically mined parallel data [8]. These systems represent distinct strategies: commercial translation, multilingual zero-shot MT, vocabulary-augmented prompting, and corpus-based adaptation.

Their relative English-Sango translation quality, however, remains largely unexplained. No peer-reviewed study has compared the available systems under a common and reproducible protocol. Because they differ in model access, training resources, and adaptation strategies, results produced under separate datasets and evaluation conditions are not directly comparable. Moreover, developer-reported results for the adapted French-Sango model show an in-distribution gain but lower performance than the pretrained NLLB baseline on a FLORES-200 sample [8]. Although preliminary, this pattern is consistent with NMT research showing that domain-specific fine-tuning can improve adaptation-domain performance while degrading general-domain performance [9]. Such degradation

has been examined as catastrophic forgetting [10] and linked to adaptation-data properties such as target-vocabulary coverage [11]. Whether this pattern occurs in English-Sango translation has not been established through direct comparison with zero-shot NLLB-200.

This study evaluates six English-Sango MT systems: Google Translate, SangoAI, zero-shot NLLB-200, a directly fine-tuned NLLB system (Direct FT), a two-stage English-French-Sango system (French Pivot), and an augmented fine-tuning system (Augmented FT) combining pivot-generated data, back-translation, and orthographic normalization. It addresses the following questions:

RQ1. How does translation performance differ across the six English-to-Sango MT systems, as measured by automatic metrics and native-speaker judgments?

RQ2. How do the out-of-domain performances of the LoRA fine-tuning conditions compare with that of zero-shot NLLB-200, and does the combined augmentation procedure outperform a quantity- and step-matched control trained on directly mined data?

The remainder of this paper is organized as follows. Section II reviews low-resource MT for African languages, existing Sango translation systems, fine-tuning, LoRA, and adaptation-related out-of-domain degradation, and linguistic characteristics of Sango relevant to the present study. Section III presents the evaluation framework, method, results, and discussion. Section IV concludes the paper and outlines directions for future research.

II. Preliminaries

1. Low-Resource Machine Translation for African Languages

Low-resource MT increasingly combines three complementary strategies: transfer from pretrained

multilingual models, data-centric resource construction through mining and back-translation, and LLM-assisted translation or synthetic-data generation [4], [12]–[15]. These strategies are often integrated rather than used in isolation. NLLB-200 is central to this development because it supports 200 languages and provides a pretrained basis for language-specific adaptation [4]. Its associated bitext resources were automatically mined using the Stopes pipeline and LASER3 encoders [13], [16]; consequently, corpus scale should be distinguished from alignment accuracy, domain diversity, and suitability for a specific language pair.

African-language MT has also advanced through community-driven resource development. Masakhane established a participatory framework for building datasets and systems with relevant language communities [2], while JW300 supplied parallel data for more than 300 languages from Jehovah’s Witnesses publications [17]. These resources supported systems for Fon-French, Hausa-English, and several South African languages [18]–[20]. Their domain composition remains consequential. Religious corpora can provide otherwise unavailable supervision, but may poorly represent general-purpose use. MENYO-20k evaluations demonstrated that domain and diacritic variation materially affect Yorùbá-English MT performance, highlighting the risks of assessing general-purpose systems only on religious or otherwise narrow distributions [21].

Fine-tuning and augmentation have nevertheless produced substantial low-resource gains. Issaka et al. reported broad improvements across 31 African languages after combining curated data with model adaptation [22], while English-Kokborok NLLB adaptation using professionally translated, religious-domain, and back-translated data achieved strong held-out results [23]. LLMs have also served as translation baselines and synthetic-data generators [15], and multi-armed-bandit selection has been proposed when further fine-tuning is impractical [24]. These

studies establish the potential of adaptation and resource construction, but not the transfer of held-out gains to an independent out-of-domain benchmark.

2. Existing Sango Translation Systems and Evaluation Evidence

Current Sango systems represent several methodological families, but peer-reviewed comparative evidence is sparse. Their published descriptions also differ in evaluation sets and reporting practices, preventing reliable cross-system comparison. NLLB-200 provides the principal open multilingual baseline and the FLORES-200 benchmark used here [4]. Google Translate supports Sango, but its Sango-specific model, training data, and adaptation procedure are undisclosed; it is therefore treated as a commercial black box [5], [6].

SangoAI uses vocabulary-augmented prompting, supplying curated lexical information and language-specific instructions to a general-purpose LLM without task-specific fine-tuning or conventional parallel-corpus training [7]. This developer-authored documentation establishes the reported design, not independently verified translation quality, and no reproducible peer-reviewed evaluation of SangoAI has been identified.

By contrast, MEYNG/nllb-sango-finetuned-600m is described in its developer-authored model card as a full-parameter adaptation of NLLB-200-distilled-600M trained on more than 500,000 automatically mined French-Sango sentence pairs, rather than as a LoRA-based adapter [8]. According to the same documentation, French-to-Sango BLEU increased from 13.79 to 19.49 on an adaptation-distribution sample but decreased from 6.11 to 3.39 on a FLORES-200 sample [8]. Because these results are reported only in developer documentation and have not been independently reproduced in a controlled peer-reviewed study, they are treated here as preliminary artifact-level evidence.

Sango has also been included in broad FLORES-200 benchmarking of small language models [25], but that work neither compares the English-to-Sango translation systems examined here nor tests adaptation against the corresponding zero-shot baseline. Thus, existing evidence does not establish either the relative performance of current English→Sango systems or whether Sango adaptation improves out-of-domain performance.

3. Fine-Tuning, LoRA, and Out-of-Domain Degradation in Low Resource MT

Full-parameter fine-tuning adapts pretrained models to particular language pairs or distributions but requires updating and maintaining a complete model for each adaptation condition. Parameter-efficient fine-tuning (PEFT) reduces this burden by updating a small parameter subset or adding lightweight modules to a shared frozen backbone [12], [26]. Low-Rank Adaptation (LoRA) is particularly relevant for three reasons. First, it freezes pretrained weights and learns low-rank updates, reducing trainable parameters and adaptation memory [26]. Second, lightweight adapters can specialize a common backbone for multiple languages or domains. Third, low-resource MT studies show that PEFT can provide a competitive efficiency-quality trade-off [27], [28]. LoRA is therefore motivated primarily by computational feasibility and scalable specialization, not by an assumption of superior translation quality.

Empirical results are mixed. Across low-resource translation conditions, several PEFT configurations improved in- and out-of-domain performance, but standard LoRA was not consistently the strongest method [28]. On NLLB-200-distilled-600M across 13 Indigenous-to-Spanish pairs, LoRA updated 0.19% of parameters and offered a strong efficiency-quality trade-off, whereas full fine-tuning achieved the highest aggregate held-out score [27]. Specialized approaches have reported stronger gains: MLAS-LoRA separates language-general and language-specific modules to reduce multilingual

interference [29], and low-resource Indic systems have combined LoRA and DoRA (Weight-Decomposed Low-Rank Adaptation) with normalization, filtering, transliteration, and post-processing [30], [31]. Conversely, business-domain experiments found that LoRA largely retained general-domain performance but learned domain-specific patterns less effectively than full fine-tuning [32]. Computational efficiency, adaptation quality, and retention of previous capability are therefore distinct outcomes.

This distinction is essential because adaptation gains may not generalize beyond the training distribution. Domain-specific NMT fine-tuning can improve adaptation-domain quality while degrading performance elsewhere [9]. This loss has been analyzed as catastrophic forgetting [10], and its severity may depend on properties of the adaptation data [11]. Out-of-domain degradation denotes the observed decline, whereas catastrophic forgetting is an explanatory interpretation requiring evidence that the same model performs worse on the same previously supported condition after adaptation [10], [33]. A low post-adaptation score alone is insufficient because it may reflect weak initial capability, noisy data, inadequate learning, or optimization failure.

LoRA is not inherently exempt. Although pretrained weights remain frozen, inference combines them with learned low-rank updates, so parameter preservation does not guarantee preservation of the base model's behavior [26]. Controlled multilingual MT experiments found comparable forgetting under LoRA and full-parameter fine-tuning on translation directions excluded from adaptation: larger models were generally more resistant, while greater volumes of selective fine-tuning data intensified the loss [33]. Most low-resource PEFT studies, however, emphasize the final quality and efficiency of adapted systems rather than directly comparing the same model before and after adaptation on the same out-of-domain set [27], [28], [33]. Evidence

remains limited for same-direction out-of-domain generalization after LoRA adaptation with automatically mined data.

Accordingly, the unresolved issue is not whether LoRA can reduce training cost, but whether its adaptation gains survive a distributional shift in an extremely low-resource setting. The present study addresses this gap using two LoRA-adapted NLLB-200 systems. Direct FT uses automatically mined English-Sango pairs, whereas Augmented FT applies the same adaptation procedure to data additionally incorporating pivot-generated and back-translated pairs with orthographic normalization. Their FLORES-200 performance is compared with zero-shot NLLB-200 to determine whether adaptation preserves, improves, or degrades quality outside the adaptation distribution.

4. Linguistic Characteristics of Sango Relevant to the Present Study

Sango is a tonal language written in a Latin-based orthography. Descriptive accounts identify seven oral vowels and three level tones represented through diacritics [34]–[36]. Digital texts may omit or inconsistently represent these marks because of input constraints and competing practices [37], creating surface variation that affects corpus matching and reference-based evaluation.

French has also exerted substantial lexical influence on Sango. Established French loanwords may be phonologically and semantically integrated and should not automatically be treated as mistranslations [38], [39]. Extended minimally adapted French sequences, however, may indicate untranslated material, source-target interference, or noise in mined bitext. This distinction informs interpretation of French-pivot outputs and lexical intrusions in adapted systems.

III. The Proposed Evaluation Scheme

1. Evaluation Design and Experimental Conditions

To address the two research questions stated in Section I, this study employed two related but analytically distinct evaluation frameworks.

RQ1 concerns the comparative performance of six English-to-Sango MT systems: Google Translate, SangoAI, zero-shot NLLB-200, French Pivot, Direct FT, and Augmented FT. All six systems were evaluated on the complete 1,012-sentence FLORES-200 English-Sango devtest set using BLEU, chrF, and TER. Descriptive judgments from a native Sango speaker on an 18-sentence subset were additionally used as supplementary evidence for the central patterns observed in the automatic evaluation.

RQ2 examines the out-of-domain performance of the two primary LoRA-adapted systems, Direct FT and Augmented FT, relative to zero-shot NLLB-200. Direct FT was trained on 8,000 directly mined English-Sango sentence pairs, whereas Augmented FT combined the same 8,000 directly mined pairs with 500 French-pivot-generated pairs and 500 back-translated pairs. Direct-9K was additionally included as a quantity- and step-matched control. It was trained on the original 8,000 directly mined pairs plus 1,000 additional, non-overlapping directly mined pairs and matched Augmented FT in both total training-pair count and optimizer-step budget, but contained no pivot-generated or back-translated data. The comparison between Direct-9K and Augmented FT was therefore used to examine whether the performance difference was associated with the combined augmentation procedure rather than simply with a larger training pool or greater optimization exposure.

Table 1 summarizes the six primary translation systems included in RQ1 and the additional Direct-9K fine-tuning condition included as a quantity- and step-matched control for RQ2.

Table 1. Translation systems and experimental conditions used for RQ1 and RQ2

Condition	System	Type	Analysis
No Additional Training	Google Translate	Commercial	RQ1
	SangoAI	Prompting-based	RQ1
	Zero-shot NLLB	Pretrained baseline	RQ1, RQ2
	French Pivot	Two-stage EN→FR→SG pipeline system	RQ1
Fine-tuned Condition	Direct FT	LoRA-adapted	RQ1, RQ2
	Direct-9K	Quantity- and step-matched control	RQ2
	Augmented FT	LoRA-adapted	RQ1, RQ2

Note. Direct-9K is an additional control condition and is not included among the six primary systems compared for RQ1.

Figure 1 summarizes the output-generation pathways for the four systems evaluated without additional training and the three fine-tuning conditions, together with the automatic and native-speaker evaluation procedures.

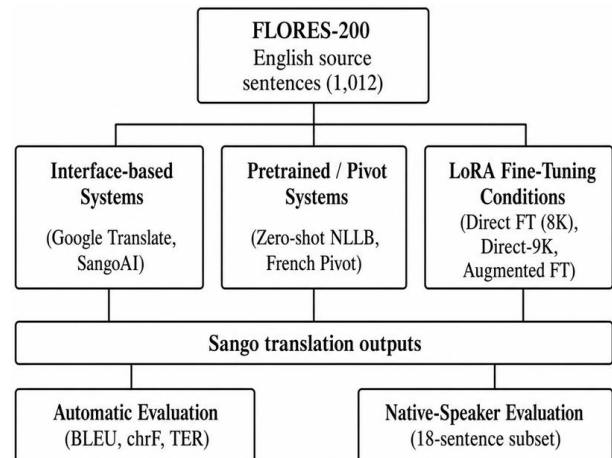


Fig. 1. Overall Workflow for System Output Generation and Evaluation

2. Data Construction, System Configuration, and Evaluation Procedures

2.1. Data Construction for Fine Tuning

The FLORES-200 English-Sango devtest set was used as the common out-of-domain evaluation benchmark. In this study, out-of-domain evaluation refers operationally to testing on FLORES-200,

which was neither drawn from the automatically mined corpus used for adaptation nor used for training or model selection. This definition does not imply that the topical distribution of the mined corpus was exhaustively characterized. The benchmark contains 1,012 English source sentences with professionally translated Sango references [4]. Table 2 summarizes the training-data composition and optimizer-step budgets for the three main fine-tuning conditions.

Table 2. Training-data composition and optimizer-step budget within FT Condition

Condition	Direct mined pairs	Pivot-generated pairs	Back-translated pairs	Optimizer Steps
Direct FT	8,000	0	0	3,000
Direct -9K	9,000	0	0	3,375
Augmented FT	8,000	500	500	3,375

Note. Direct-9K and Augmented FT were matched in total training-pair count and optimizer-step budget. Pivot-generated pairs refer to synthetic English-Sango pairs produced through the EN→FR→SG pipeline and should not be confused with the standalone French Pivot system. Rule-based orthographic normalization was applied to the Sango side of the 8,000 directly mined base pairs used in Augmented FT; the pivot-generated and back-translated components were not normalized.

Training data were drawn from the English-Sango and French-Sango portions of the allenai/nllb mined corpus [13], [16]. Sentence pairs with LASER3 similarity scores of at least 1.0 were retained, yielding 3,447,841 English-Sango pairs and 971,276 French-Sango pairs. Before further cleaning, 200 English-Sango pairs were reserved as a held-out in-distribution set and excluded from all model training. This set was used only for diagnostic monitoring of validation loss and was not used for checkpoint selection or the reported translation-quality evaluation. The remaining English-Sango data were screened on the Sango side for script validity, degenerate textual patterns, and French-language contamination. Target strings were retained only when at least 90% of their characters belonged to the permitted Latin-based character set. Strings containing URLs,

HTML-related patterns, or sequences of five or more consecutive digits were removed. French contamination was calculated as the proportion of whitespace-tokenized items matching the 581-entry french field of the MEYNG/sango-vocabulary dataset, and pairs with a ratio above 0.30 were excluded. These procedures removed 48,946 pairs and yielded a cleansed pool of 3,398,695 English-Sango pairs.

Direct FT used the first 8,000 pairs of the ordered cleansed pool as a fixed training budget under the available single-GPU setting. Direct-9K used the same 8,000-pair base set and added 1,000 non-overlapping directly mined pairs randomly sampled from the remaining pool with a fixed seed of 43. Augmented FT used the same base set together with 500 pivot-generated pairs and 500 back-translated pairs. The pivot-generated component was created from 500 unique English source sentences obtained after deduplicating the English side of the base set. Each sentence was translated into French using zero-shot NLLB-200 and subsequently into Sango using MEYNG/nllb-sango-finetuned-600m. The back-translated component used the first 500 Sango-side sentences from the LASER-filtered French-Sango corpus, which were translated into English using zero-shot NLLB-200 in the Sango-to-English direction, producing synthetic-English-original-Sango pairs. Both synthetic-data generation procedures used greedy decoding, a maximum output length of 128 tokens, and the appropriate forced target-language token. Before Augmented FT was assembled, rule-based normalization informed by established descriptions of Sango orthography [34], [37] and phonology [35], [36] was applied only to the Sango side of the 8,000 directly mined base pairs. The procedure standardized whitespace, apostrophes, selected vowel symbols, Unicode representations of diacritics, and nasal notation before “b” and “p”; the pivot-generated and back-translated components were not normalized. Direct-9K and

Augmented FT were matched in total training-pair count and optimizer-step budget, allowing their comparison to assess differences associated with data composition rather than data quantity or optimization exposure alone.

Figure 2 summarizes the data-construction pipeline from the mined corpus to the fixed 8K base set and illustrates how Direct FT, Direct-9K, and Augmented FT were constructed from that base.

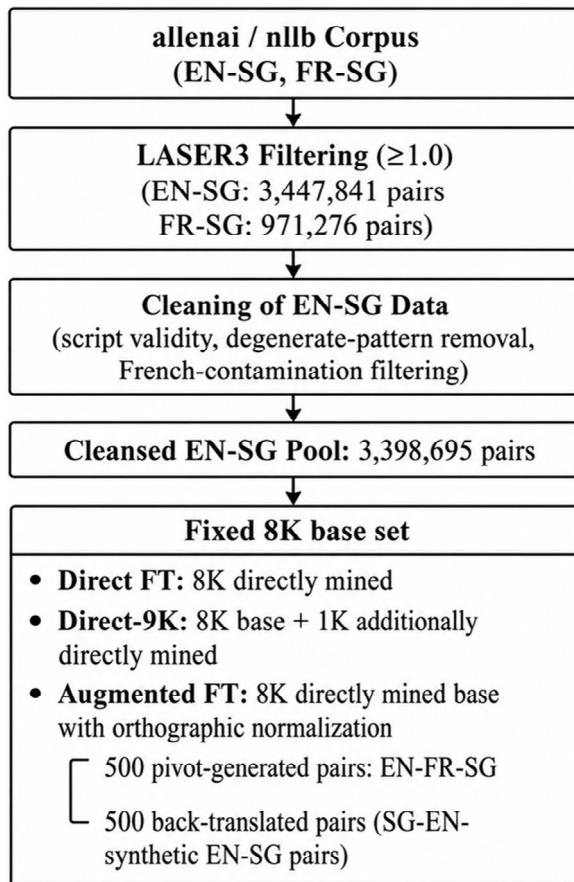


Fig. 2. Data-construction Pipeline for the LoRA Fine Tuning Conditions.

For the component-specific follow-up analysis, three additional training conditions were constructed. The normalization-only condition used the same 8,000 directly mined base pairs as Direct FT, with the rule-based orthographic normalization described above applied to the Sango targets. The pivot-only condition combined the 8,000-pair direct-data base with 1,000 pivot-generated English-Sango pairs, while the back-translation-only condition combined the same 8,000-pair base with

1,000 back-translated pairs. The synthetic pairs were generated using the same translation pipelines and decoding settings described above for the corresponding components of Augmented FT. These conditions were evaluated as single-run follow-up experiments and were intended to provide descriptive component-specific evidence rather than strict one-factor-removal ablations, because the synthetic-data conditions used 1,000 pairs of each respective component rather than the 500-pair quantity used for each component within Augmented FT.

To assess potential overlap with FLORES-200, the final training sets used for Direct FT (8,000 pairs), Direct-9K (9,000 pairs), and Augmented FT (9,000 pairs, including its 500 pivot-generated and 500 back-translated components) were directly compared against the complete FLORES-200 devtest set (1,012 sentences). After Unicode NFKC normalization, lowercasing, and punctuation removal, exhaustive exact-match comparisons and character 8-gram Jaccard near-duplicate searches (similarity threshold ≥ 0.7) found no exact or near-duplicate matches in any of the three final training sets on either the English or Sango side. The 1,000 synthetic pairs included in Augmented FT were also examined separately and likewise showed no overlap.

As a supplementary corpus-level check, a reconstructed candidate pool of 3,403,161 English-Sango pairs corresponding to the direct-mined corpus was examined using the same normalization procedure. No exact matches were found in the full reconstructed pool, and no near-duplicates were detected in a random sample of 200,000 pairs.

2.2. System Configuration and Output Collection

Zero-shot NLLB, Direct FT, Direct-9K, and Augmented FT were based on facebook/nllb-200-distilled-600M. The three fine-tuned conditions used LoRA adapters applied to the query, key, value, and output projection matrices of the transformer attention layers [26]. The LoRA configuration used rank $r = 32$, scaling

factor $\alpha = 64$, and dropout of 0.05, resulting in approximately 1.51% of the model parameters being trainable.

The final LoRA configuration was informed by preliminary pilot runs rather than a systematic hyperparameter sweep. A lower-rank configuration restricted to the query and value projections showed target-language instability, including occasional French output, which motivated the use of $r = 32$ with LoRA applied to the query, key, value, and output projections in the formal experiments. The scaling factor was set to $\alpha = 64$ ($\alpha/r = 2$) as part of this fixed configuration. Because α and α/r were not independently varied, the selected setting should not be interpreted as optimal.

Training was conducted for three epochs with the Adafactor optimizer, a learning rate of 1×10^{-4} , and an effective batch size of 8 [44]. All fine-tuning and local-model inference were conducted on a single NVIDIA T4 GPU. Training and validation loss trajectories for the three main fine-tuning conditions are provided in Appendix. Validation loss was evaluated every 100 optimizer steps on the 200-pair held-out in-distribution set described in Section 2.1 and was used only for diagnostic monitoring, not for checkpoint selection or early stopping.

For inference, zero-shot NLLB and the three fine-tuning conditions used greedy decoding, a maximum output length of 128 tokens, and the `sag_Latn` target-language token. The French Pivot system was implemented as a two-stage inference pipeline and involved no additional training in this study. Each FLORES-200 English source sentence was first translated into French using zero-shot NLLB-200. The resulting French output was then translated into Sango using the externally released MEYNG/nllb-sango-finetuned-600m French-to-Sango model, which was developed through full-parameter fine-tuning [8]. Only the final Sango output was retained for evaluation.

Preliminary full-parameter fine-tuning attempts produced degenerate outputs under the tested configurations and were excluded from the formal

evaluation. Because alternative technical explanations were not systematically ruled out, these attempts served only as the procedural motivation for adopting LoRA and were not interpreted as evidence that full-parameter fine-tuning is generally unsuitable for English-Sango translation.

Google Translate outputs were collected through the Google Cloud Translation API using English as the source language and Sango as the target language. Outputs for all 1,012 sentences were stored and evaluated using the same metric pipeline as the other systems.

SangoAI outputs were collected through its public developer API [7]. The system is publicly described as using vocabulary-augmented prompting without task-specific fine-tuning [7]. Following interruptions caused by rate limiting, a resumable collection protocol was implemented using fixed inter-request delays, chunk-level saving, exponential backoff, and per-sentence logging. A length-stratified sample of 500 sentences and the remaining 512 sentences were collected in two non-overlapping passes and subsequently merged according to their original FLORES-200 indices. This procedure yielded complete SangoAI output coverage for all 1,012 source sentences.

To consolidate the experimental settings relevant to reproducibility, Table 3 summarizes the backbone checkpoint, LoRA configuration, trainable parameter count, data-selection and training seeds, software-library versions, decoding settings, automatic-evaluation signatures, and API/output collection dates. The table distinguishes deterministic construction of the base 8K training set from seed-based supplemental sampling and model training and separately reports the decoding configurations used for evaluation, pivot generation, and back-translation.

Table 3. Experimental Configuration and Reproducibility Details

Category	Setting
Backbone	facebook/nllb-200-distilled-600M
LoRA configuration	$r = 32$, $\alpha = 64$, dropout = 0.05
LoRA target modules	q_proj, k_proj, v_proj, out_proj
Trainable parameters	9,437,184 / 624,510,976 (1.51%)
Epochs	3
Learning rate / scheduler	1×10^{-4} ; linear scheduler; warmup ratio = 0.1
Optimizer	Adafactor
Data-selection seeds	Base 8K: deterministic first 8,000 pairs; Direct-9K supplemental sampling; seed 43
Training seeds	Original run: seed 42; multi-seed analysis: seeds 42, 43, 44
Evaluation decoding	Greedy (do_sample=False, num_beams=1); max_length=128; target = sag_Latn
Pivot decoding	EN→FR→SG; greedy; max_length=128; forced target-language token at each stage
Back-translation decoding	SG→EN; greedy; max_length=128; target = eng_Latn
Library versions	transformers 5.13.1; datasets 3.6.0; peft 0.19.1; torch 2.11.0+cu128; sacreBLEU 2.6.0
BLEU signature	nrefs:1 case:mixed eff:no tok:13a smooth:exp version:2.6.0
chrF signature	nrefs:1 case:mixed eff:yes nc:6 nw:0 space:no version:2.6.0
TER signature	nrefs:1 case:lc tok:tercom norm:no punct:yes asian:no version:2.6.0
FLORES-200 revision (post-hoc verification)	71abf77d8b7beb5cfef59898d6b24d92ab7654fc; original run not revision-pinned; verified counts: dev = 997, devtest = 1,012
API/output	collection date July 8, 2026 (Google Translate, SangoAI)

The FLORES-200 dataset used in the original experiments was loaded without pinning a specific repository revision. A post-hoc verification at the revision reported in Table 3 confirmed the same dataset sizes originally observed (dev = 997; devtest = 1,012). The reported revision is therefore provided as a post-hoc reproducibility reference rather than as the revision explicitly pinned during the original experiments.

2.3. Evaluation Procedures

Automatic evaluation was conducted on the complete 1,012-sentence FLORES-200 devtest set for the six primary systems and the Direct-9K

control. Translation quality was assessed at the corpus level using BLEU [41], chrF [42], and TER [43]. All metrics were computed with Sacre BLEU [40].

BLEU measures modified word-level n -gram precision while applying a brevity penalty to outputs that are shorter than the reference. It is calculated as follows [41]:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right)$$

where p_n is the modified precision for n -grams of order n , w_n is the corresponding weight, and BP is the brevity penalty. The BP is defined as 1 when $c > r$ and $\text{BP} = \exp\left(1 - \frac{r}{c}\right)$ when $c \leq r$, where c and r denote the total candidate and effective reference lengths, respectively. Higher BLEU scores indicate greater word-level correspondence with the reference [41].

chrF evaluates translation quality by combining character n -gram precision and recall. It is calculated as follows [42]:

$$\text{chrF}_\beta = \frac{(1 + \beta^2) \cdot \text{chrP} \cdot \text{chrR}}{\beta^2 \cdot \text{chrP} + \text{chrR}}$$

where chrP and chrR denote character n -gram precision and recall, respectively, and β determines the relative weight assigned to recall. In this study, β was set to 2; thus, the reported scores correspond to chrF_2 . Because chrF operates at the character level, it is generally more tolerant than word-based metrics of tokenization differences and minor orthographic variation, although it remains sensitive to differences in surface form [42]. This characteristic is particularly relevant to written Sango, in which diacritics and orthographic representations may vary [34], [37]. Higher chrF scores indicate greater character-level correspondence with the reference.

TER measures the minimum number of edits required to transform a system output into the reference translation [43]. It is calculated as

follows:

$$\text{TER} = \frac{\text{number of edits}}{\text{average number of reference words}}$$

An edit consists of the substitution of a word, deletion of an unnecessary word, insertion of a missing word, or movement of a contiguous word sequence to another position. At the corpus level, the numerator represents the minimum total number of such edits required to transform the system outputs into the reference translations, and the denominator corresponds to the total number of words in the reference translations. Because one reference translation was used for each source sentence, this denominator is equivalent to the aggregate reference length across the evaluation set.

SacreBLEU reports TER on a 0-100 scale [40]. Unlike BLEU and chrF₂, lower TER scores indicate better translation performance because they show that fewer edits are required to reproduce the reference translation.

Differences in corpus BLEU were assessed using sentence-level paired bootstrap resampling with replacement [45]. All reported comparisons were calculated using a uniform protocol of 1,000 bootstrap resamples. The analysis included adjacent comparisons in the observed RQ1 BLEU ordering and the prespecified contrasts required for RQ2: zero-shot NLLB versus each fine-tuning condition and Augmented FT versus the quantity- and step-matched Direct-9K control. Each test reported the observed BLEU difference, a percentile-based 95% confidence interval, and a two-sided empirical p-value. When no resample favored the lower-scoring system, the result was reported as $p < .001$.

Automatic evaluation was supplemented by judgments from one native Sango speaker (Evaluator 1) on an 18-sentence subset shared by the six primary systems. The 18 source sentences were randomly sampled without replacement from

the first 100 sentences of the FLORES-200 devtest set using a fixed random seed of 7. The first 100 sentences constituted the sampling frame because they corresponded to the portion for which SangoAI outputs were available when the original native-speaker evaluation instrument was prepared. The corresponding 0-based positions in the FLORES-200 devtest sequence were 4, 6, 7, 8, 9, 11, 12, 19, 27, 41, 46, 50, 53, 55, 64, 68, 74, and 83. Because the sampling frame was restricted to the first 100 sentences, this 18-sentence subset should not be regarded as fully representative of the complete 1,012-sentence devtest set; accordingly, the native-speaker evaluation is treated as exploratory. Each system output was rated using 0-5 scales for semantic accuracy, naturalness, and orthographic appropriateness. Evaluator 1 also completed two forced-choice tasks: one comparing zero-shot NLLB-200 with the corresponding fine-tuned output presented for each item, and another comparing Google Translate with Augmented FT. In this initial evaluation, system identities were not masked and the outputs were not randomized. Because the Direct FT and Augmented FT outputs in this initial evaluation had been generated from earlier checkpoints rather than the checkpoints used in the final automatic evaluation, these judgments were treated as exploratory and were not considered direct validation of the final automatic-evaluation results.

To address the checkpoint mismatch, a second native-speaker evaluation was conducted using the exact Direct FT and Augmented FT checkpoints (seed = 42) whose scores are reported in Table 4. The same 18 source sentences were used. The Direct FT and Augmented FT outputs were regenerated using these checkpoints, while the outputs of the other four systems were retained. For each sentence, system identities were masked and the six outputs were presented in randomized order.

Because Evaluator 1 was unavailable during the revision period, the checkpoint-matched

re-evaluation was conducted by a second native Sango speaker (Evaluator 2), who was provided with the same evaluation rubric and instructions used in the initial evaluation. Evaluator 2 used the same 0–5 scales for semantic accuracy, naturalness, and orthographic appropriateness as Evaluator 1. Because the two evaluations involved different evaluators and, for Direct FT and Augmented FT, different model checkpoints, their results were analyzed descriptively and reported separately rather than pooled. The checkpoint-matched evaluation was likewise treated as exploratory because it involved a single evaluator and only 18 sentences.

Following completion of the initial evaluation sheet, Evaluator 1 was contacted to clarify references to “Bible content” in the free-text comments and the interpretation of the error-annotation columns. The evaluator confirmed that these comments referred to unrelated religious or preaching-like content appearing in multiple system outputs within the evaluated subset, particularly in the Direct FT and Augmented FT outputs. The evaluator also provided English back-translations for three representative outputs. In the error-annotation columns, numerical entries represented approximate counts of problematic words or segments, whereas “A” indicated that an output was too inaccurate, unrelated, or incoherent for a meaningful assessment of the specified error type. These follow-up explanations and back-translations were used only for qualitative interpretation and were not used in the quantitative analyses.

3. Results

3.1. Comparative Performance of the Six Systems: RQ1

3.1.1. Quantitative Analysis

Table 4 presents the automatic evaluation results for the six primary English-to-Sango MT systems on the complete 1,012-sentence FLORES-200 devtest set. Google Translate achieved the highest

performance across all three metrics, obtaining BLEU 9.953, chrF₂ 40.179, and TER 82.922. Zero-shot NLLB-200 ranked second across all three metrics, with BLEU 7.832, chrF₂ 34.583, and TER 91.326.

Table 4. Automatic evaluation results for the six primary systems on FLORES-200

System	BLEU	chrF ₂	TER
Google Translate	9.953	40.179	82.922
Zero-shot NLLB	7.832	34.583	91.326
French Pivot	6.137	31.631	97.012
Augmented FT	5.628	29.355	95.820
Direct FT	4.886	28.049	95.216
SangoAI	3.454	25.395	93.732

For BLEU and chrF₂, the remaining systems were ordered as French Pivot, Augmented FT, Direct FT, and SangoAI. French Pivot achieved BLEU 6.137 and chrF₂ 31.631, followed by Augmented FT with BLEU 5.628 and chrF₂ 29.355. Direct FT obtained BLEU 4.886 and chrF₂ 28.049, while SangoAI produced the lowest BLEU and chrF₂ scores, at 3.454 and 25.395, respectively. The TER results did not reproduce the same ordering. SangoAI achieved a lower, and therefore better, TER score than French Pivot, Augmented FT, and Direct FT. Among the three systems, French Pivot produced the highest TER value.

Paired bootstrap comparisons confirmed that the adjacent differences in the observed BLEU ordering were statistically significant. Google Translate exceeded zero-shot NLLB-200 by 2.12 BLEU points, 95% CI [1.68, 2.56], $p < .001$. Zero-shot NLLB-200 outperformed French Pivot by 1.69 points, 95% CI [1.22, 2.16], $p < .001$. French Pivot exceeded Augmented FT by 0.51 points, 95% CI [0.07, 0.95], $p = .022$, and Augmented FT exceeded Direct FT by 0.74 points, 95% CI [0.28, 1.16], $p = .002$. Direct FT also outperformed SangoAI by 1.43 BLEU points, 95% CI [1.00, 1.93],

$p < .001$. These tests support the observed BLEU ordering for the comparisons directly examined, although the magnitude of the differences was considerably smaller between French Pivot, Augmented FT, and Direct FT than between the two highest-ranked systems.

The automatic evaluation was supplemented by descriptive judgments from Evaluator 1, a native Sango speaker, on the 18-sentence subset. Table 5 reports the mean ratings from this initial evaluation for semantic accuracy, naturalness, and orthographic appropriateness. Google Translate received the highest mean rating on all three dimensions, followed by zero-shot NLLB-200. Direct FT received the lowest ratings, while Augmented FT consistently scored higher than Direct FT.

Table 5. Mean Native-Speaker Ratings in the Initial Evaluation

System	Semantic accuracy	Naturalness	Orthography
Google Translate	3.94	3.50	3.00
Zero-shot NLLB	2.50	2.22	2.83
SangoAI	1.61	1.44	2.17
French Pivot	1.56	1.61	2.17
Augmented FT	1.17	1.39	1.94
Direct FT	0.78	1.06	1.78

Note. Ratings were provided by Evaluator 1 on 0-5 scales and are reported descriptively. The Direct FT and Augmented FT outputs were generated using earlier checkpoints rather than the exact checkpoints used in the final automatic evaluation.

The initial native-speaker ratings broadly reflected the principal automatic-evaluation pattern. Google Translate and zero-shot NLLB-200 received higher ratings than the two LoRA-adapted systems, and Augmented FT received higher ratings than Direct FT across all three dimensions. However, the human ratings did not reproduce the complete automatic-metric ordering. SangoAI received a slightly higher semantic-accuracy rating than French Pivot, whereas French Pivot received a higher naturalness rating, and the two systems obtained the same orthographic rating.

Evaluator 1's scalar ratings were supplemented by two pairwise preference comparisons. When zero-shot NLLB-200 was compared with the corresponding fine-tuned output for each of the 18 items, the evaluator preferred the zero-shot output in 11 cases (61%), preferred the fine-tuned output in two cases (11%), and judged both outputs to be poor in five cases (28%). In the separate comparison between Google Translate and Augmented FT, Google Translate was preferred in all 18 cases.

To directly assess the final checkpoints used in the automatic evaluation, a checkpoint-matched native-speaker re-evaluation was conducted by Evaluator 2 using the same 18 source sentences. Table 6 presents the mean ratings from this re-evaluation.

Table 6. Mean Native-Speaker Ratings in the Checkpoint-Matched Re-Evaluation

System	Semantic accuracy	Naturalness	Orthography
Google Translate	4.00	3.53	3.00
Zero-shot NLLB	2.53	2.24	2.82
SangoAI	1.62	1.50	2.12
French Pivot	1.65	1.59	2.18
Augmented FT	1.81	1.81	2.44
Direct FT	1.35	1.47	2.35

Note. Ratings were provided by Evaluator 2 on the same 0-5 scales used in the initial evaluation and are reported descriptively. The Direct FT and Augmented FT outputs were generated using the exact seed-42 checkpoints used in the final automatic evaluation. System identities were masked, and the six outputs for each sentence were presented in randomized order.

In the checkpoint-matched re-evaluation, Augmented FT received higher mean ratings than Direct FT on all three dimensions: semantic accuracy (1.81 vs. 1.35), naturalness (1.81 vs. 1.47), and orthographic appropriateness (2.44 vs. 2.35). This relative ordering was consistent with the initial evaluation, in which Augmented FT also received higher ratings than Direct FT across all three dimensions. Google Translate again received the highest mean ratings across all three dimensions,

followed by zero-shot NLLB-200. Because the checkpoint-matched evaluation involved a single evaluator and only 18 sentences, these findings are interpreted descriptively as exploratory rather than as definitive validation of the automatic results.

Taken together, the two native-speaker evaluations provided converging descriptive evidence for the higher ratings of Google Translate and zero-shot NLLB-200 relative to the two LoRA-adapted systems and, more specifically, for the relative advantage of Augmented FT over Direct FT. However, because the evaluations were conducted by different evaluators and each involved only 18 sentences, their scores were not pooled and should not be interpreted as definitive human validation of the system-level rankings. Evaluator 1's free-text comments further clarified the types of semantic failure underlying some of the low ratings, as discussed in the following subsection.

3.1.2. Qualitative Analysis of Native Speaker Comments

Evaluator 1's free-text comments and follow-up clarification from the initial evaluation revealed domain-incongruent religious content in multiple system outputs within the 18-sentence evaluation subset. In several cases, the translations introduced religious expressions, preaching-like formulations, or biblical references that were absent from the English source. Although the evaluator did not identify these expressions as quotations from specific Bible verses, their lexical content and discourse style were clearly religious and unrelated to the source topics, which included technology, politics, and medical research.

Table 7 presents an illustrative case involving a report that criticized the Executive's policy toward Iraq. The Augmented FT output did not preserve the political content of the source sentence. Instead, it introduced an explicit reference to the Bible and replaced the report's call for a policy change with a statement concerning religious influence and the changing of times.

Table 7. Illustrative example of domain-incongruent religious content in an Augmented FT output

Type	Content
English Source	The Report is highly critical of almost every aspect of the present policy of the Executive towards Iraq and it urges an immediate change of direction.
Augmented FT output	<i>A yeke fa so a yeke na ngangu ti sara ngangu na ndo ti aye kue so ayeke na ya ti Bible, na a yeke mu maboko na e ti changé anbeni ngoi.</i>
Sango Native-speaker back-translation into English	<i>"It shows that it has the power to influence everything in the Bible, and it helps us to change times."</i>
Main divergence	The criticism of the Executive's Iraq policy is replaced by an unrelated reference to the Bible and a statement about changing times.

The native-speaker back-translation indicates that the central proposition of the source was not retained. The error therefore extended beyond a local lexical choice or an orthographic difference: the political topic and the communicative purpose of the sentence were replaced by semantically unrelated religious content.

Table 8. Contrasting religious-domain intrusions in Direct FT and Augmented FT outputs for the same English source, "The Iraq Study Group presented its report at 12.00 GMT today."

Type	Content
English Source	The Iraq Study Group presented its report at 12.00 GMT today.
1) Direct FT	<i>Na nda ni, a-Wafango ye ti Royaume ti Israël amû lani tënë ti ala na ngoi ti 12 h 00.</i>
Sango Native-speaker back-translation into English	<i>"Finally, the Teachers of the Kingdom of Israel gave their talk at 12:00 pm."</i>
Main divergence	The Iraq Study Group is replaced by "the Teachers of the Kingdom of Israel," while presented its report is rendered as "gave their talk."
2) Augmented FT output	<i>A yeke na ngoi so a yeke fa tënë ti Royaume ni na lango 12 ti nze ti ndapre.</i>
Sango Native-speaker back-translation into English	<i>"It is during the Kingdom-preaching work on the 12th of the month."</i>
Main divergence	The report-presentation event is replaced by "Kingdom-preaching work," and 12.00 GMT today is reinterpreted as a calendar date.

Table 8 compares Direct FT and Augmented FT outputs generated from the same English source sentence. The source reports the presentation of the Iraq Study Group’s report at a specified time. Both fine-tuned systems departed substantially from this political-reporting context, but they did so in different ways.

The native-speaker back-translations show that the same input elicited two different forms of religious-domain intrusion. Direct FT retained part of the temporal information but transformed the institutional subject into “the Teachers of the Kingdom of Israel” and the reporting event into the delivery of a talk. Augmented FT instead reconstructed the event as “Kingdom-preaching work” and misinterpreted the time expression as a date. Thus, although both outputs departed from the source proposition, the specific form of the semantic distortion differed across the two fine-tuned systems.

Taken together, the examples indicate that some of the low semantic-accuracy ratings reflected substantial departures from the source proposition rather than isolated lexical, syntactic, or orthographic errors. In the selected cases, central political content was replaced by religious-domain expressions that were neither licensed by the English source nor required by the Sango context. The findings also show that Direct FT and Augmented FT did not necessarily reproduce identical errors; the same source sentence could generate distinct forms of domain-incongruent content.

A post-hoc review of the original evaluation annotations was conducted to determine how often religious-domain content had been explicitly noted by the evaluators. In Evaluator 1’s initial evaluation sheet, explicit religion-related comments were recorded for 1 of the 18 source sentences (5.6%), affecting 4 of the 108 system outputs (3.7%). This annotation-based count should not be interpreted as the frequency of religious-domain intrusion in the initial evaluation, because Evaluator 1 did not

systematically label every such case during the original rating procedure. Indeed, the follow-up clarification and back-translations discussed above identified additional examples for which no religion-related comment had been entered in the original notes.

In the checkpoint-matched re-evaluation conducted by Evaluator 2, the free-text notes explicitly identified religious-domain intrusion (“hallucinates a Bible context”) in 2 of the 18 source sentences (11.1%), corresponding to 3 of the 108 system outputs (2.8%). These cases comprised a Direct FT output for a sentence describing the PALM clinical trial and both the Direct FT and Augmented FT outputs for a separate sentence concerning Iraq policy. Because neither evaluation employed a predefined religious-intrusion coding scheme, these annotation counts are reported only as descriptive information and should not be interpreted as estimates of prevalence across either the 18-sentence subset or the complete FLORES-200 devtest set. The two evaluations also involved different evaluators and, for Direct FT and Augmented FT, different checkpoints and presentation procedures; their annotation counts therefore should not be compared as evidence of an increase or decrease in the occurrence of religious-domain content.

3.2. Out-of-Domain Performance and Matched-Control Analysis: RQ2

Table 9 presents the automatic evaluation results for zero-shot NLLB-200 and the three fine-tuning conditions examined under RQ2. Zero-shot NLLB-200 outperformed Direct FT, Direct-9K, and Augmented FT across BLEU and chrF₂ and also achieved a lower TER score than all three conditions. Among the fine-tuning conditions, Augmented FT obtained the highest BLEU and chrF₂ scores. Direct FT, however, achieved a slightly lower TER score than Augmented FT, indicating that the relative advantage of Augmented FT was not uniform across all three metrics.

Table 9. Automatic evaluation results for the RQ2 conditions

Condition	Training Composition	BLEU	chrF ₂	TER
Zero-shot NLLB	No additional training	7.832	34.583	91.326
Direct FT (8K)	8K direct-mined	4.886	28.049	95.216
Direct-9K	9K direct-mined	4.743	27.711	96.189
Augmented FT	8K direct + 500 pivot-generated + 500 back-translated	5.628	29.355	95.820

Zero-shot NLLB-200 significantly outperformed all three fine-tuning conditions on the out-of-domain FLORES-200 benchmark. It exceeded Direct FT by 2.95 BLEU points, 95% CI [2.46, 3.43], $p < .001$; Direct-9K by 3.09 points, 95% CI [2.56, 3.58], $p < .001$; and Augmented FT by 2.20 points, 95% CI [1.71, 2.71], $p < .001$. Thus, none of the tested LoRA fine-tuning conditions improved FLORES-200 performance relative to the zero-shot pretrained baseline.

The comparison between Direct FT and Direct-9K showed that adding 1,000 directly mined sentence pairs, together with the corresponding increase in optimization exposure, did not improve out-of-domain BLEU. Direct-9K scored 0.14 points below Direct FT, 95% CI [-0.53, 0.23], and the difference was not statistically significant, $p = .460$.

By contrast, Augmented FT outperformed the quantity- and step-matched Direct-9K condition by 0.88 BLEU points, 95% CI [0.46, 1.28], $p < .001$. Because Direct-9K and Augmented FT contained the same total number of training pairs and used the same optimizer-step budget, the observed difference was associated with their differing data composition rather than simply with total training-pool size or optimization exposure. Under the tested configuration, replacing 1,000 additional directly mined pairs with 500 pivot-generated and 500 back-translated pairs was therefore associated with better out-of-domain BLEU performance.

Nevertheless, Augmented FT remained significantly below zero-shot NLLB-200. The combined

augmentation procedure therefore mitigated part of the performance loss observed among the fine-tuning conditions but did not eliminate the broader adaptation-related out-of-domain degradation. Because the matched-control comparison above evaluates the combined augmentation procedure as a whole, additional component-specific follow-up analyses are reported below.

To assess the stability of the single-run findings across training variability, the three fine-tuning conditions were evaluated across three training seeds (42, 43, and 44), while holding the LoRA configuration, learning rate, and all other hyperparameters constant. Table 10 reports the mean and standard deviation of BLEU, chrF, and TER across the three independently trained runs, together with the originally reported single-run values for comparison.

Table 10. Original Single-Run and Multi-Seed Results for the Three Fine-Tuning Conditions

Condition	Metric	Original (single-run)	Multi-Seed Mean $\pm$ SD (n = 3)	95% CI
Direct FT (8K)	BLEU	4.89	4.92 $\pm$ 0.13	[4.61, 5.24]
	chrF ₂	28.05	28.74 $\pm$ 0.07	[28.57, 28.91]
	TER	95.22	98.55 $\pm$ 1.52	[94.77, 102.32]
Direct-9K	BLEU	4.74	4.72 $\pm$ 0.16	[4.32, 5.12]
	chrF ₂	27.71	28.34 $\pm$ 0.20	[27.85, 28.83]
	TER	96.19	96.96 $\pm$ 1.88	[92.28, 101.64]
Augmented FT	BLEU	5.63	5.16 $\pm$ 0.12	[4.87, 5.45]
	chrF ₂	29.36	29.49 $\pm$ 0.27	[28.82, 30.16]
	TER	95.82	99.09 $\pm$ 0.36	[98.19, 99.99]

Note. Multi-seed estimates are based on three independently trained runs (seeds 42, 43, and 44). The 95% intervals are t-based intervals across the three runs (df = 2) and should be interpreted cautiously because of the small number of runs. The training datasets, including the deterministic 8,000-pair base set, were held fixed across seeds; therefore, these intervals reflect training-run variability rather than data-selection variability. For Augmented FT, the multi-seed runs used a reconstructed synthetic-data source split rather than the exact split used in the original single-run experiment because the original synthetic-data source split had not been archived.

For Direct FT, the originally reported single-run BLEU score (4.89) fell within the range observed across the three seeds, with a multi-seed mean of 4.92 ($SD = 0.13$). For Direct-9K, the original BLEU score (4.74) was likewise close to the multi-seed mean of 4.72 ($SD = 0.16$). Augmented FT maintained the highest mean BLEU among the three fine-tuning conditions, although its multi-seed mean (5.16, $SD = 0.12$) was lower than the originally reported single-run value of 5.628. This difference may partly reflect variation introduced when the synthetic-data components were reconstructed, because the exact source split used in the original Augmented FT run had not been archived. The across-seed mean BLEU scores preserved the same ordering among the fine-tuning conditions (Augmented FT > Direct FT > Direct-9K), and all three remained below the zero-shot NLLB-200 baseline.

The multi-seed analysis showed numerically higher mean TER than the original single-run TER for all three fine-tuning conditions. Mean chrF was also slightly higher across all three conditions, whereas BLEU did not show a uniform directional shift relative to the original single-run values. Overall, the multi-seed results preserved the mean BLEU ordering among the three fine-tuning conditions while also indicating some across-run variability in the estimated performance of each condition. For Augmented FT, comparison with the original single-run result is additionally complicated by reconstruction of the synthetic-data source split. The paired bootstrap results reported above remain specific to the original checkpoints and should therefore be interpreted separately from the across-run variability estimated by the multi-seed analysis.

To address the contribution of the individual augmentation components, a component-specific follow-up analysis was conducted for pivot-generated data, back-translation, and orthographic normalization, as summarized in Table

11. Because the resulting conditions differed in training-set size and, for the synthetic-data conditions, did not reproduce the exact 500-pair component quantities used in Augmented FT, the analysis is treated as a descriptive follow-up evidence rather than as a strict one-factor-removal experiment. The normalization-only condition used the same 8,000 directly mined pairs as Direct FT, with orthographic normalization applied to the Sango targets. The pivot-only and back-translation-only conditions each contained 9,000 pairs, comprising the 8,000-pair direct-data base plus 1,000 pairs generated using the respective synthetic-data procedure.

Table 11. Results of the Component-Specific Follow-up Analysis

Condition	Training Composition	BLEU	chrF ₂	TER
Direct FT (8K)	8K direct	4.92 ± 0.13	28.74 ± 0.07	98.55 ± 1.52
Normalization-only	8K direct Normalized	5.24	28.72	95.75
Direct-9K	9K direct	4.72 ± 0.16	28.34 ± 0.20	96.96 ± 1.88
Pivot-only	8K direct + 1K pivot-generated	5.31	29.06	106.03
Back-translation-only	8K direct + 1K back-translated	5.13	29.18	98.09
Augmented FT	8K direct + 500 pivot-generated + 500 back-translated	5.16 ± 0.12	29.49 ± 0.27	99.09 ± 0.36

Note. Direct FT, Direct-9K, and Augmented FT values represent means and standard deviations across three training seeds. The normalization-only, pivot-only, and back-translation-only conditions were evaluated as additional single-run follow-up conditions. Pivot-only and back-translation-only each used 1,000 pairs of the respective synthetic-data type rather than the 500-pair quantity used for each component within Augmented FT.

The normalization-only condition achieved a BLEU score of 5.24, compared with the 8K Direct FT multi-seed mean of 4.92. Among the 9K-scale conditions, pivot-only and back-translation-only achieved BLEU scores of 5.31 and 5.13, respectively, both numerically higher than the Direct-9K multi-seed mean of 4.72. Thus, all three

component-specific conditions yielded higher BLEU scores than their corresponding scale-matched direct-data baselines under the tested configuration, although these comparisons do not establish the independent contribution of each component within the combined Augmented FT condition.

The pattern was not uniform across evaluation metrics. Notably, the pivot-only condition achieved the highest BLEU among the component-specific conditions (5.31) while also producing the highest TER in Table 11 (106.03). This divergence indicates that the relative performance of the component-specific conditions depends partly on the evaluation metric and cautions against interpreting the BLEU differences as a uniform improvement in translation quality.

The component-specific results should therefore be interpreted descriptively. First, the three follow-up conditions were evaluated as single runs, whereas Direct FT and Direct-9K are represented by multi-seed means, making the comparisons asymmetric with respect to training-run variability. Second, pivot-only and back-translation-only each used 1,000 pairs of the respective synthetic-data type, whereas Augmented FT contained 500 pivot-generated and 500 back-translated pairs. The component-specific conditions therefore do not reproduce the exact component quantities used in the combined Augmented FT setting. Accordingly, these experiments do not provide precise estimates of the independent contribution of each component within Augmented FT, nor do they establish whether the components interact additively, synergistically, or redundantly.

4. Discussion

4.1. Comparative Performance and Metric-Dependent Rankings

The system comparison revealed a clear separation between the two leading systems and the remaining approaches. Google Translate ranked first across all automatic metrics and

native-speaker rating dimensions, while zero-shot NLLB-200 consistently ranked second. Because Google Translate is a commercial black-box system, its advantage cannot be attributed to a particular architecture, dataset, or adaptation procedure. The strong performance of zero-shot NLLB-200 is nevertheless notable because it confirms the value of massively multilingual transfer for an extremely low-resource language [4]. French Pivot, Augmented FT, Direct FT, and SangoAI formed a lower performance range on BLEU and chrF, although the differences among French Pivot and the two primary fine-tuned systems were substantially smaller than the gap separating them from the two leading systems. These results provide a common benchmark for systems based on substantially different development strategies, rather than evidence that any one strategy is universally superior.

The rankings were not entirely consistent across evaluation metrics. SangoAI obtained the lowest BLEU and chrF scores but a lower, and therefore better, TER score than French Pivot, Direct FT, and Augmented FT. This metric-dependent reversal is compatible with the distinct properties of the three measures described in Section 2.3. BLEU evaluates modified word-level n-gram precision with a brevity penalty, chrF combines character n-gram precision and recall, and TER estimates the normalized number of edits required to transform a system output into the reference [41]–[43]. A translation may therefore require relatively few edits to approximate the reference while still exhibiting limited exact word- or character-level overlap. Orthographic variation in written Sango may further affect reference-based comparisons, particularly those based on exact surface correspondence [34], [37]. Consequently, SangoAI's relatively lower TER should be interpreted as a metric-specific result rather than as evidence of uniformly better translation quality. The native-speaker ratings likewise preserved the broad advantage of Google Translate and zero-shot

NLLB-200 but did not reproduce the complete automatic-metric ordering among the lower-ranked systems. These findings support the use of multiple automatic metrics together with native-speaker evidence and caution against treating any single reference-based metric as a complete measure of translation quality.

4.2. Adaptation-Related Out-of-Domain Degradation

The most consequential result was that zero-shot NLLB-200 significantly outperformed all three LoRA fine-tuning conditions on the out-of-domain FLORES-200 benchmark. Domain-specific fine-tuning has previously been shown to improve performance on an adaptation distribution while reducing performance elsewhere [9], and this loss has been examined in relation to catastrophic forgetting and the composition of adaptation data [10], [11]. The present results extend this concern to same-direction English-to-Sango evaluation using automatically mined data. They also reinforce the distinction between parameter preservation and behavioral preservation: although LoRA freezes the pretrained weights and learns low-rank updates [26], all three adapted conditions produced substantially weaker out-of-domain behavior than the unadapted backbone. The training and validation loss trajectories reported in the Appendix showed no evidence of gross optimization instability under the tested configuration, although they do not establish that the selected hyperparameters were optimal. Nevertheless, the present study uses the more descriptive term adaptation-related out-of-domain degradation rather than diagnosing catastrophic forgetting, because adaptation-domain improvement was not tested for the same final checkpoints.

The result is also consistent with the developer-reported evidence for the full-parameter MEYNG French-to-Sango model, which improved on an adaptation-distribution sample but fell below its NLLB baseline on a FLORES-200 sample [8]. The convergence of the two patterns suggests that the

observed degradation may not be unique to LoRA, in line with controlled evidence that parameter-efficient adaptation does not necessarily prevent forgetting more effectively than full fine-tuning [33]. However, the MEYNG results involve a different translation direction, training set, and evaluation procedure and therefore do not constitute a controlled LoRA-full-parameter comparison. Similarly, the religious-domain intrusions identified in the native-speaker evaluations illustrate the semantic form that some failures took but do not identify their cause. Prior research shows that religious corpora and domain composition can materially affect low-resource MT [17], [21], but the present study neither quantified the domain composition of the mined corpus nor measured the frequency of religious intrusions across the full test set.

4.3. Effects and Limits of the Combined Augmentation Procedure

The matched-control analysis distinguishes the effect of data composition from a simple increase in training quantity. Direct-9K did not achieve significantly higher BLEU than Direct FT despite including 1,000 additional directly mined pairs and receiving a proportionally larger optimizer-step budget. This nonsignificant result should not be interpreted as evidence that the two conditions were equivalent; rather, it indicates that the tested increase in similarly mined data produced no detectable out-of-domain BLEU benefit. Large mined corpora can expand low-resource coverage, but corpus size does not by itself guarantee alignment quality, domain diversity, or suitability for a particular language pair [4], [13]. In the present setting, modestly increasing the quantity of data drawn from the same mined pool was therefore insufficient to recover the zero-shot model’s FLORES-200 performance.

By contrast, Augmented FT achieved significantly higher BLEU than the quantity- and step-matched Direct-9K condition. Because the two conditions

contained the same number of training pairs and used the same optimizer-step budget, this difference indicates a composition-related advantage for the combined augmentation procedure rather than an advantage attributable only to data volume or optimization exposure. This result is consistent with the established value of back-translation [14] and with evidence that the effectiveness of parameter-efficient adaptation depends on the specific method and data condition [27], [28]. The component-specific follow-up analysis further provided descriptive evidence that the normalization-only condition achieved a higher BLEU score than the 8K Direct FT baseline, while the pivot-only and back-translation-only conditions achieved higher BLEU scores than the 9K Direct-9K baseline. However, these follow-up results should be interpreted cautiously because the component-specific conditions were evaluated as single runs and the pivot-only and back-translation-only conditions used 1,000 synthetic pairs each rather than the 500-pair quantities used for each component in Augmented FT. Accordingly, the follow-up analysis does not establish the independent contribution of each component within the combined condition or determine whether their effects are additive, synergistic, or redundant. Moreover, Augmented FT remained significantly lower than zero-shot NLLB-200 in BLEU. The combined augmentation procedure therefore reduced part of the out-of-domain BLEU loss relative to the matched direct-data control but did not restore the pretrained model's zero-shot performance.

4.4. Methodological Implications, Limitations, and Practical Recommendations

The findings have several implications for low-resource MT evaluation. First, an adapted multilingual model should be compared with the corresponding strong zero-shot backbone; otherwise, a positive score or an improvement over a weak baseline may obscure adaptation-related

regression. Second, adaptation-distribution validation should be complemented by an independent out-of-domain benchmark. Third, augmentation claims should be tested against controls matched for both data quantity and optimization exposure. Fourth, system comparisons should combine consistently reported automatic metrics, statistical significance testing, and native-speaker judgments where feasible [2], [4], [15], [40], [45]. The SangoAI collection process also demonstrates that evaluations involving rate-limited APIs require resumable collection, sentence-level logging, and verification of complete output coverage to ensure that system comparisons remain reproducible.

These conclusions must be considered alongside the study's limitations. Although the three principal fine-tuning conditions were evaluated across three training seeds, the training datasets were held fixed, and the multi-seed results therefore do not capture data-selection variability. The component-specific follow-up conditions provide only descriptive evidence because they were evaluated as single runs and did not reproduce the exact component quantities used in Augmented FT. The native-speaker evaluation also remained exploratory: the initial and checkpoint-matched evaluations each involved only 18 sentences and one evaluator, and the two sets of ratings were reported separately because they involved different evaluators and, for Direct FT and Augmented FT, different checkpoints and presentation procedures. The domain composition of the mined training corpus was not exhaustively characterized, and the occurrence of religious-domain intrusions was not systematically assessed across the complete FLORES-200 set. The study also did not evaluate the final LoRA checkpoints on a matched adaptation-domain test set or include a successfully trained full-parameter English-to-Sango baseline. Finally, the present results concern a relatively small adaptation regime of 8,000-9,000 sentence pairs and should not be

assumed to generalize to substantially larger fine-tuning datasets. Practically, an adapted Sango model should therefore not replace a stronger zero-shot model solely because it has undergone language-specific fine-tuning. Developers should conduct out-of-domain regression tests, audit mined data for alignment and domain concentration, retain a zero-shot fallback when adaptation reduces general performance, and use blinded evaluations by multiple native speakers before deployment in high-stakes or general-purpose settings.

IV. Conclusions

This study compared six English-to-Sango machine translation systems under a common evaluation framework and examined the out-of-domain effects of LoRA-based adaptation in an extremely low-resource language setting. Google Translate achieved the strongest overall automatic-evaluation performance, followed by zero-shot NLLB-200. All three LoRA-adapted conditions showed significantly lower BLEU scores than the zero-shot baseline in the paired-bootstrap comparisons on FLORES-200, indicating that the tested language-specific adaptation conditions did not preserve the pretrained model’s general-domain performance in the English-to-Sango setting. This finding is particularly relevant to low-resource MT, where limited access to high-quality parallel data often increases reliance on automatically mined corpora and may make adaptation outcomes sensitive to data quality and domain composition. The quantity- and step-matched comparison further showed that Augmented FT achieved significantly higher BLEU than Direct-9K, whereas adding an equivalent amount of additional directly mined data produced no detectable improvement over Direct FT. Thus, under the tested configuration, the combined use of pivot-generated data, back-translation, and

orthographic normalization provided an advantage over an equal-sized directly mined training set, although it remained insufficient to recover the zero-shot baseline. The exploratory native-speaker evaluations broadly supported the automatic-evaluation pattern, while the qualitative analysis also identified cases of severe semantic distortion, including unrelated religious-domain intrusions, within the evaluated 18-sentence subset.

Several limitations should be considered when interpreting these findings. First, although the three principal fine-tuning conditions were repeated across three training seeds, the training datasets themselves were held fixed across runs. In particular, the base 8,000-pair adaptation set was constructed as a deterministic first slice of the ordered cleansed pool. The reported across-seed variability therefore reflects stochastic variation in model training but does not capture variability arising from alternative selections of the adaptation data. Moreover, since the multi-seed analysis was based on only three training seeds, the corresponding t-score based confidence intervals should be interpreted cautiously. The 500 Sango sentences used for back-translation were also selected through deterministic slicing, and the ordering of the deduplicated English sources used for pivot generation was not explicitly controlled. These procedures may have introduced corpus-order or data-selection effects that were not examined in the present study.

Second, the component-specific follow-up experiments provide only descriptive evidence regarding the roles of orthographic normalization, pivot-generated data, and back-translation. The pivot-only and back-translation-only conditions each used 1,000 synthetic pairs, whereas Augmented FT contained 500 pairs from each source, and these follow-up conditions were evaluated in single runs rather than across multiple training seeds. They therefore do not constitute strict one-factor-removal ablations and cannot establish the independent, additive, or

interactive contribution of each augmentation component.

Third, the native-speaker evaluation remained exploratory. The initial evaluation and the checkpoint-matched re-evaluation each involved only 18 source sentences and a single native Sango-speaking evaluator. Because Evaluator 1 was unavailable for the checkpoint-matched evaluation, Evaluator 2 conducted the second evaluation; consequently, the two sets of ratings were reported separately rather than pooled, and inter-rater reliability could not be assessed. In addition, the initial evaluation was not blinded and included earlier Direct FT and Augmented FT checkpoints, whereas the second evaluation used masked, randomized outputs from the final seed-42 checkpoints. Although the qualitative analysis identified religious-domain intrusions within the evaluated subset, their frequency was not systematically assessed across the complete FLORES-200 test set.

Fourth, the topical and domain composition of the automatically mined training corpus was not exhaustively characterized. Although direct overlap and near-duplicate checks found no overlap between the final training sets and FLORES-200 under the procedures described in Section 2.1, the present study cannot determine which specific properties of the adaptation data were responsible for the observed out-of-domain degradation. Future work should therefore examine independently sampled adaptation sets, larger numbers of training seeds, and more systematic characterization of training-data quality and domain composition.

Finally, the fine-tuning conditions examined here involved relatively small adaptation sets of 8,000–9,000 pairs, corresponding to approximately 0.24–0.26% of the cleansed candidate pool. Prior work has reported that larger volumes of selective fine-tuning data can intensify forgetting [33]. Accordingly, the present findings should be interpreted as evidence from a small-scale

adaptation regime and should not be assumed to generalize to substantially larger fine-tuning datasets.

Despite these limitations, the study demonstrates the importance of evaluating low-resource adaptation against a strong zero-shot backbone on an independent out-of-domain benchmark. Under the relatively small-scale adaptation regime examined here, LoRA fine-tuning of NLLB-200 on automatically mined English-Sango parallel data did not improve general-domain translation performance over the zero-shot model. At the same time, the matched-control and component-specific follow-up analyses indicate that data composition can influence the extent of adaptation-related out-of-domain degradation: the combined augmentation condition outperformed an equal-sized directly mined condition, although none of the tested fine-tuning conditions recovered the zero-shot baseline. These findings suggest that language-specific adaptation should not be assumed to improve translation quality solely because additional parallel data are available; its benefits should instead be established through controlled evaluation on both the intended target domain and independent out-of-domain benchmarks. Future research should examine larger and independently sampled adaptation datasets, broader multi-seed designs, and more systematic characterization of training-data quality and domain composition. It should also investigate whether alternative parameter-efficient adaptation methods, including Orthogonal Finetuning [46], MLAS-LoRA [29], and DoRA [31], can improve English-to-Sango translation while better preserving the multilingual capabilities of the pretrained model.

ACKNOWLEDGEMENT

Generative AI was used during manuscript preparation for language refinement, formatting, and checking the formatting of references. The authors independently verified the existence and accuracy of all cited sources, together with all analyses and interpretations, and take full responsibility for the final manuscript.

REFERENCES

- [1] S. Bird and D. Chiang, “Machine Translation for Language Preservation,” in *Proceedings of COLING 2012: Posters*, pp. 125–134, 2012.
- [2] W. Nekoto et al., “Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2144–2160, 2020. DOI: 10.18653/v1/2020.findings-emnlp.195.
- [3] W. J. Samarin, “French and Sango in the Central African Republic,” *Anthropological Linguistics*, Vol. 28, No. 3, pp. 379–387, 1986.
- [4] NLLB Team, “Scaling Neural Machine Translation to 200 Languages,” *Nature*, Vol. 630, pp. 841–846, 2024. DOI: 10.1038/s41586-024-07335-x.
- [5] I. Caswell, “110 New Languages Are Coming to Google Translate,” Google, June 27, 2024. <https://blog.google/products-and-platforms/products/translate/google-translate-new-languages-2024/>
- [6] Google Translate Help, “What’s New in Google Translate: More Than 100 New Languages,” June 27, 2024. <https://support.google.com/translate/answer/15139004>
- [7] M. Wenezoui, “Vocabulary-Augmented Prompting for Sango—Production African Language AI without a Parallel Corpus,” *Hugging Face Community Articles*, May 6, 2026. <https://huggingface.co/blog/MEYNG/sango-vocabulary-augmented-prompting>
- [8] M. Wenezoui, “MEYNG/nllb-sango-finetuned-600m,” *Hugging Face Model Card*, May 22, 2026.
- [9] P. Dakwale and C. Monz, “Fine-Tuning for Neural Machine Translation with Limited Degradation across In- and Out-of-Domain Data,” in *Proceedings of Machine Translation Summit XVI: Research Track*, pp. 156–169, 2017.
- [10] B. Thompson, J. Gwinnup, H. Khayrallah, K. Duh, and P. Koehn, “Overcoming Catastrophic Forgetting during Domain Adaptation of Neural Machine Translation,” in *Proceedings of NAACL-HLT 2019*, pp. 2062–2068, 2019. DOI: 10.18653/v1/N19-1209.
- [11] D. Saunders and S. DeNeefe, “Domain Adapted Machine Translation: What Does Catastrophic Forgetting Forget and Why?” in *Proceedings of EMNLP 2024*, pp. 12660–12671, 2024. DOI: 10.18653/v1/2024.emnlp-main.704.
- [12] A. Bapna and O. Firat, “Simple, Scalable Adaptation for Neural Machine Translation,” in *Proceedings of EMNLP-IJCNLP 2019*, pp. 1538–1548, 2019. DOI: 10.18653/v1/D19-1165.
- [13] K. Heffernan, O. Çelebi, and H. Schwenk, “Bitext Mining Using Distilled Sentence Representations for Low-Resource Languages,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2101–2112, 2022. DOI: 10.18653/v1/2022.findings-emnlp.154.
- [14] R. Sennrich, B. Haddow, and A. Birch, “Improving Neural Machine Translation Models with Monolingual Data,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 86–96, 2016. DOI: 10.18653/v1/P16-1009.
- [15] L. Susanto, R. Diandaru, A. Krisnadhi, A. Purwarianti, and D. T. Wijaya, “Replicable Benchmarking of Neural Machine Translation (NMT) on Low-Resource Local Languages in Indonesia,” in *Proceedings of the First Workshop in South East Asian Language Processing*, pp. 100–115, 2023. DOI: 10.18653/v1/2023.sealp-1.8.
- [16] Allen Institute for AI, “allenai/nllb,” dataset, Hugging Face, 2022. <https://huggingface.co/datasets/allenai/nllb>
- [17] Ž. Agić and I. Vulić, “JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3204–3210, 2019. DOI: 10.18653/v1/P19-1310.
- [18] A. Akinfaderin, “HausaMT v1.0: Towards English–Hausa Neural Machine Translation,” in *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pp. 144–147, 2020. DOI: 10.18653/v1/2020.winlp-1.38.
- [19] C. C. Emezue and F. P. B. Dossou, “FFR v1.1: Fon–French Neural Machine Translation,” in *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pp. 83–87, 2020. DOI: 10.18653/v1/2020.winlp-1.21.
- [20] L. Martinus, J. Webster, J. Moonsamy, M. Shaba, Jr., R. Moosa, and R. Fairon, “Neural Machine Translation for South Africa’s Official Languages,” arXiv:2005.06609, 2020. DOI: 10.48550/arXiv.2005.06609.
- [21] D. I. Adelani et al., “The Effect of Domain and Diacritics in Yorùbá–English Neural Machine Translation,” in *Proceedings of Machine Translation Summit XVIII: Research Track*, pp. 61–75, 2021.
- [22] S. Issaka et al., “The African Languages Lab: A Collaborative Approach to Advancing Low-Resource African NLP,”

- arXiv:2510.05644, 2025. DOI: 10.48550/arXiv.2510.05644.
- [23] B. Nyalang and B. Debbarma, "Towards High-Quality Machine Translation for Kokborok: A Low-Resource Tibeto-Burman Language of Northeast India," arXiv:2604.19778, 2026. DOI: 10.48550/arXiv.2604.19778.
- [24] P. R. Singh, K. Prasad, M. Zaki, and P. Wasnik, "In-Domain African Languages Translation Using LLMs and Multi-Armed Bandits," in *Proceedings of the Sixth Workshop on African Natural Language Processing*, pp. 167–175, 2025. DOI: 10.18653/v1/2025.africanlp-1.26.
- [25] Y. Song, L. Li, C. Lothritz, S. Ezzini, L. Sleem, N. Gentile, R. State, T. F. Bissyandé, and J. Klein, "Are Small Language Models the Silver Bullet to Low-Resource Languages Machine Translation?" in *Proceedings of the Ninth Workshop on Technologies for Machine Translation of Low-Resource Languages*, pp. 1–26, 2026. DOI: 10.18653/v1/2026.loresmt-1.1.
- [26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proceedings of the Tenth International Conference on Learning Representations*, 2022.
- [27] D. Stackhouse and J. Debnedetto, "A Systematic Comparison of Parameter-Efficient Fine-Tuning Techniques for Low-Resource Neural Machine Translation: Evidence from Indigenous Languages of the Americas," in *Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas*, pp. 33–45, 2026. DOI: 10.18653/v1/2026.americasnlp-6.4.
- [28] T. Su, X. Peng, S. Thillainathan, D. Guzmán, S. Ranathunga, and E.-S. Lee, "Unlocking Parameter-Efficient Fine-Tuning for Low-Resource Language Translation," in *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4217–4225, 2024. DOI: 10.18653/v1/2024.findings-naacl.263.
- [29] T. Dong, B. Li, J. Liu, S. Zhu, and D. Xiong, "MLAS-LoRA: Language-Aware Parameters Detection and LoRA-Based Knowledge Transfer for Multilingual Machine Translation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vol. 1*, pp. 15645–15660, 2025. DOI: 10.18653/v1/2025.acl-long.762.
- [30] P. Acharya, H. Mondal, D. Saha, D. Das, and S. Bandyopadhyay, "JU-NLP: Improving Low-Resource Indic Translation System with Efficient LoRA-Based Adaptation," in *Proceedings of the Tenth Conference on Machine Translation*, pp. 1201–1209, 2025. DOI: 10.18653/v1/2025.wmt-1.95.
- [31] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "DoRA: Weight-Decomposed Low-Rank Adaptation," in *Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research*, Vol. 235, pp. 32100–32121, 2024.
- [32] J. Eschbach-Dymanus, F. Essenberger, B. Buschbeck, and M. Exel, "Exploring the Effectiveness of LLM Domain Adaptation for Business IT Machine Translation," in *Proceedings of the 25th Annual Conference of the European Association for Machine Translation, Vol. 1*, pp. 610–622, 2024.
- [33] D. Liu and J. Niehues, "Conditions for Catastrophic Forgetting in Multilingual Translation," in *Proceedings of the 5th Workshop on Multilingual Representation Learning*, pp. 347–359, 2025. DOI: 10.18653/v1/2025.mrl-main.23.
- [34] M. Diki-Kidiri, *Dictionnaire orthographique du sängö*. Bangui: Buku ti Beafrika (BBA), 1998.
- [35] W. J. Samarin, "Sango," in *The Survey of Pidgin and Creole Languages: Vol. III. Contact Languages Based on Languages from Africa, Asia, Australia, and the Americas*, S. M. Michaelis, P. Maurer, M. Haspelmath, and M. Huber, Eds. Oxford University Press, pp. 13–24, 2013.
- [36] J. A. Walker and W. J. Samarin, "Sango Phonology," in *Phonologies of Asia and Africa (Including the Caucasus)*, Vol. 2, A. S. Kaye and P. T. Daniels, Eds. Winona Lake, IN: Eisenbrauns, pp. 861–880, 1997.
- [37] M. Diki-Kidiri, "La réforme de l'orthographe officielle du sanngo en Centrafrique," *C@LEA—Cadernos de Aulas do LEA*, Special Issue, pp. 14–19, 2022.
- [38] C. R. Taber, *French Loan Words in Sango: A Statistical Analysis of Incidence*. Hartford, CT: Hartford Seminary Foundation, 1964.
- [39] C. Thornell, "Established French Loanwords in Sango: A Pilot Study," *Working Papers, Lund University, Department of Linguistics*, Vol. 44, pp. 159–177, 1995.
- [40] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, 2018. DOI: 10.18653/v1/W18-6319.
- [41] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. DOI: 10.3115/1073083.1073135.
- [42] M. Popović, "chrF: Character n-Gram F-Score for Automatic MT Evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, 2015. DOI: 10.18653/v1/W15-3049.
- [43] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231, 2006.
- [44] N. Shazeer and M. Stern, "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost," in *Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research*, Vol. 80, pp. 4596–4604, 2018.
- [45] P. Koehn, "Statistical Significance Tests for Machine Translation

- Evaluation,” in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, 2004.
- [46] Z. Qiu, W. Liu, A. Weller, and B. Schölkopf, “Orthogonal Finetuning Made Scalable,” in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31946–31963, 2025. DOI: 10.18653/v1/2025.emnlp-main.1627.

Authors



Rakhun Kim received the B.A. degree in English Education from Hankuk University of Foreign Studies, Korea, in 2008, and the M.Ed. and Ph.D. degrees in English Education from Seoul National University,

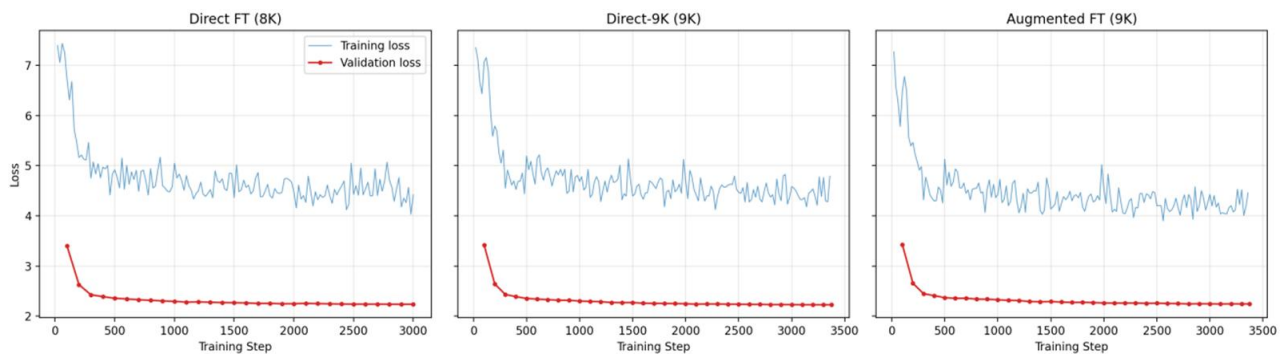
Korea, in 2012 and 2022, respectively. Dr. Kim is currently a lecturer in the Department of English Education at Hongik University. His research interests include AI-assisted language learning, natural language processing for educational applications, machine translation, and the development of intelligent systems for language learning and assessment.



Axel Deo Gratias Ndema Yetomane is an M.A. student of the Department of English Cultural Studies from Konkuk University. He is currently pursuing an M.A. degree in English and Cultural Studies.

His research interests include English linguistics, the sociocultural backgrounds of the Sango language, AI-based language research, and machine translation.

Appendix. Training and Validation Loss Trajectories for the Three Main Fine-Tuning Conditions



Note. Validation loss was evaluated on the 200-pair held-out in-distribution set every 100 optimizer steps and was used only for diagnostic monitoring, not for checkpoint selection or early stopping. Direct FT was trained for 3,000 optimizer steps, whereas Direct-9K and Augmented FT were trained for 3,375 optimizer steps. All three conditions used the same LoRA configuration described in Section 2.2.