A Systematic Comparison of Parameter-Efficient Fine-Tuning Techniques for Low-Resource Neural Machine Translation: Evidence from Indigenous Languages of the Americas

Drew Stackhouse, Justin DeBenedetto
Department of Computing Sciences, Villanova University
{dstackho,jdeben01}@villanova.edu

Abstract

We present the first systematic benchmark of parameter-efficient fine-tuning (PEFT) for low-resource neural machine translation (NMT) of indigenous languages of the Americas. We evaluate eight PEFT methods alongside full fine-tuning on NLLB-200-distilled-600M across 13 indigenous-to-Spanish language pairs spanning four resource tiers (357–125,008 training sentences). OFT (Orthogonal Finetuning) achieves the highest development-set chrF++ among PEFT methods (26.63) while training only 0.28% of parameters. LoRA (Low-Rank Adaptation) offers a strong efficiency–quality tradeoff (25.27 chrF++, 0.19%). On held-out test data, full fine-tuning ranks first (25.12) with OFT a close second (25.06; $p = 0.43$). VeRA (Vector-based Random Matrix Adaptation) and Prefix Tuning consistently underperform. These results demonstrate that PEFT is a viable alternative to full fine-tuning for indigenous-language NMT.

1 Introduction

Of the world's approximately 7,000 living languages, only a small proportion have sufficient digital resources to benefit from modern natural language processing (Joshi et al., 2020). The indigenous languages of the Americas are disproportionately affected. Hundreds of languages spanning dozens of language families are spoken across North, Central, and South America, yet most lack the parallel corpora, monolingual text, and standardized orthographies that contemporary NMT systems require (Mager et al., 2021). Machine translation has the potential to support language documentation, education, and access to information for these communities, but only if effective systems can be built from the limited data available.

Multilingual pretrained models have emerged as the dominant approach to low-resource NMT. Models such as NLLB-200 (NLLB Team et al., 2022) cover over 200 languages and can be fine-tuned on new language pairs with relatively small amounts of parallel data. However, full fine-tuning of these models by updating all 615 million parameters is computationally expensive, requires substantial GPU memory, and risks catastrophic forgetting of the pretrained representations that make transfer learning effective in the first place (Kirkpatrick et al., 2017). PEFT methods address these limitations by training only a small fraction of the model's parameters while keeping the pretrained weights frozen or nearly so (Houlsby et al., 2019).

The landscape of PEFT methods has expanded rapidly in recent years, with techniques ranging from low-rank weight decompositions (Hu et al., 2022) to orthogonal transformations (Qiu et al., 2023) to learned activation scaling (Liu et al., 2022). Each method makes different assumptions about how model weights should be adapted and each offers a different tradeoff between parameter efficiency and expressiveness. Despite this growing diversity, no systematic comparison of PEFT methods exists for low-resource indigenous-language NMT. Prior PEFT benchmarks have largely focused on high-resource or English-centric settings, leaving the question open of which methods are most effective when parallel data is measured in hundreds or thousands of sentences rather than millions.

This work addresses that gap. We make the following contributions:

1. The first comprehensive benchmark of eight PEFT methods for indigenous-language NMT, evaluated against a full fine-tuning baseline.

2. An analysis across 13 typologically diverse language pairs spanning four resource tiers (357–125,008 training sentences), revealing how data availability interacts with method effectiveness.

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 33–45
July 3, 2026

3. A quantification of the parameter-efficiency vs. translation-quality tradeoff, identifying Pareto-optimal methods.

4. Practical guidelines for practitioners selecting PEFT methods for underserved languages based on resource tier and compute budget.

2 Background

2.1 Low-Resource Neural Machine Translation

Neural machine translation systems depend on large quantities of parallel text, and their performance degrades substantially when this data is scarce. For low-resource language pairs this dependence creates tension as the languages most in need of translation tools are precisely those for which training data is hardest to obtain. The problem is compounded by morphological complexity. Many of the indigenous languages of the Americas are agglutinative or polysynthetic. They form long, meaning-dense words through the concatenation of many morphemes, which means that word-level models encounter a larger effective vocabulary and more severe data sparsity than they would for analytic languages. Standard tokenization schemes developed for European languages perform poorly on these morphological profiles. Automatic metrics calibrated on word boundaries underestimate translation quality for systems that handle them well.

Responses to these challenges have taken several forms. Transfer learning from high-resource languages, particularly via multilingual pretrained models, has become the dominant paradigm, exploiting the observation that representations learned across many languages share useful structure (Joshi et al., 2020). The AmericasNLP shared tasks (2021–2025) have motivated community efforts specifically for indigenous language MT, producing parallel corpora, shared evaluation frameworks, and a growing set of competitive baselines. However, most published systems treat fine-tuning strategy as a secondary concern, focusing instead on data augmentation, back-translation, or ensemble methods. The question of which fine-tuning approach is most appropriate for this setting has received comparatively little systematic attention.

2.2 Multilingual Pretrained Models for MT

Phase 1 of this study compares three multilingual encoder-decoder candidates representing distinct pretraining philosophies: mBART-50 (Tang et al., 2021), a denoising-objective model covering 50 languages; ByT5-small (Xue et al., 2022), a byte-level model robust to orthographic variation at the cost of longer sequences; and NLLB-200 (NLLB Team et al., 2022), trained specifically for translation across 200 languages and the only candidate that natively includes several of our target languages (Quechua, Guarani, Aymara).

2.3 Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning methods reduce the cost of adapting large pretrained models by training a small number of parameters while keeping the pretrained weights frozen or nearly so. For low-resource settings this both reduces memory and compute demands and limits the degree to which pretrained representations can be overwritten, partially mitigating catastrophic forgetting. We evaluate eight PEFT methods spanning four methodological families.

Low-Rank Reparameterization. LoRA (Hu et al., 2022) decomposes weight updates into a product of two low-rank matrices, initialized to zero so that training begins from the pretrained model's behavior; it has become the dominant PEFT paradigm and serves as our natural baseline. AdaLoRA (Zhang et al., 2023) adaptively reallocates the rank budget across weight matrices via SVD-based pruning. DoRA (Liu et al., 2024a) decomposes each weight into magnitude and direction, applying LoRA-style updates only to the directional component. VeRA (Kopiczko et al., 2024) shares a single pair of frozen random matrices across layers and learns only small per-layer scaling vectors, reducing trainable parameters dramatically.

Activation Scaling. IA^3 (Liu et al., 2022) learns three vectors per transformer layer that elementwise rescale keys, values, and feedforward activations, introducing fewer trainable parameters ($\sim$74K, 0.01% of base) than any other method in this benchmark.

Orthogonal Transformations. OFT (Qiu et al., 2023) constrains weight updates to be orthogonal transformations, preserving the pairwise angular relationships between neurons in the pretrained weight matrix. This constraint is motivated by the hypothesis that the relative geometry of learned representations matters more than their absolute positions, and preserving this geometry limits catastrophic forgetting. The orthogonal matrices are parameterized via a block-diagonal Cayley transform which ensures orthogonality throughout training without requiring projection steps. BOFT (Liu et al., 2024b) factorizes this orthogonal transformation using butterfly matrices, a structured sparse decomposition that allows information to propagate across all dimensions of the weight matrix, thus achieving a better expressiveness–parameter tradeoff than OFT's block-diagonal structure.

Prompt-Based Methods. Prefix Tuning (Li and Liang, 2021) prepends learnable continuous "virtual tokens" to the key and value matrices at every layer, modifying behavior by steering attention patterns rather than altering any pretrained weights directly. A small MLP reparameterizes the prefix during training and is discarded at inference.

2.4 Evaluation Metrics for MT

We adopt chrF++ (Popović, 2017) as the primary evaluation metric and report BLEU (Papineni et al., 2002) as a secondary metric for comparability with prior work. chrF++'s character-level n-gram F-score (augmented with word unigrams and bigrams) is less sensitive to tokenization boundaries than BLEU and rewards partial morphological matches, a critical property for agglutinative and polysynthetic languages where a single word may encode information that would span an entire clause in an analytic language. For the indigenous languages in this study, many of which exhibit productive morphology and lack standardized orthographies, character-level evaluation captures meaningful overlap that word-level metrics miss entirely; BLEU additionally correlates poorly with human judgments under morphological richness (Bapna and Firat, 2019). All analytical conclusions are drawn from chrF++. Metric computations use SacreBLEU (Post, 2018).

3 Methodology

Our experimental design follows a three-phase structure.[1] Phase 1 selects a base model from three multilingual pretrained candidates. Phase 2 benchmarks eight PEFT methods plus full fine-tuning on development data. Phase 3 evaluates the trained models on held-out test sets to assess generalization. All experiments are repeated with three random seeds (0, 1, 2) to estimate variance.

[1]Code, configuration files, and experiment scripts are available at https://github.com/drewstackhouse/peft-nmt-americas.

Language	**Family**	**Train**	**Dev**	**Tier**
Quechua	Quechuan	125,008	996	High
Wayuu	Arawakan	59,715	6,635	High
Guarani	Tupian	26,032	995	High
Awajun	Jivaroan	21,964	1,018	Med.
Nahuatl	Uto-Aztecan	16,063	672	Med.
Raramuri	Uto-Aztecan	14,720	995	Med.
Shipibo-K.	Panoan	14,592	996	Med.
Wixarika	Uto-Aztecan	8,966	994	Low
Bribri	Chibchan	7,508	996	Low
Aymara	Aymaran	6,531	996	Low
Otomi	Oto-Manguean	4,889	599	Low
Ashaninka	Arawakan	3,883	883	V-Low
Chatino	Oto-Manguean	357	499	V-Low

Table 1: Summary of the 13 indigenous-language-to-Spanish translation pairs. Tier boundaries: very-low (<5K), low (5K–10K), medium (10K–25K), high (>25K training sentences).

3.1 Languages and Data

We use parallel corpora for 13 indigenous-language-to-Spanish translation pairs drawn from the AmericasNLP shared-task datasets (Mager et al., 2021; Ebrahimi et al., 2022; Chiruzzo et al., 2024). The languages span 10 language families and exhibit substantial typological diversity, including agglutinative (Quechua, Aymara, Nahuatl), polysynthetic (Ashaninka, Guarani), and tonal (Chatino, Bribri, Otomi) profiles. Training set sizes range from 357 sentences (Chatino) to 125,008 sentences (Quechua), a 350-fold difference that motivates grouping languages into four resource tiers for analysis: *very-low* (<5K: Chatino, Ashaninka), *low* (5K–10K: Otomi, Aymara, Bribri, Wixarika), *medium* (10K–25K: Shipibo-Konibo, Raramuri, Nahuatl, Awajun), and *high* (>25K: Guarani, Wayuu, Quechua). These tiers are defined relative to the data available in this study; even the high tier (up to 125K sentences) remains far below what is typically available for high-resource languages such as French or German. All data are formatted as tab-separated parallel sentences with NLLB-style language codes as column headers. Table 1 summarizes the languages and data sizes.

3.2 Phase 1: Base Model Selection

We compare three multilingual pretrained models as candidates for fine-tuning: NLLB-200-distilled-600M (NLLB Team et al., 2022), mBART-large-50-many-to-many (Tang et al., 2021), and ByT5-small (Xue et al., 2022). To control for the adaptation method, we apply LoRA with identical hyperparameters to each model and evaluate on a strati-

Method	Targets	Params	% Base
Full FT	all	615.1M	100.00
AdaLoRA	q, v	1,771K	0.29
OFT	q, k, v, o	1,714K	0.28
LoRA	q, v	1,181K	0.19
DoRA	q, v	665K	0.11
VeRA	q, v	616K	0.10
BOFT	q, k, v, o	553K	0.09
Prefix Tuning	–	492K	0.08
IA^3	k, v, wo	74K	0.01

Table 2: Trainable parameter counts and percentage of the 615M-parameter base model for each adaptation method. Methods are ordered by parameter count.

fied subset of four languages spanning the resource tiers: Chatino (very-low), Bribri (low), Nahuatl (medium), and Guarani (high). The model with the highest mean chrF++ across languages and seeds is selected for all subsequent experiments.

3.3 Model and Tokenizer Preparation

The selected base model is NLLB-200-distilled-600M, an encoder-decoder transformer with approximately 615 million parameters. Of the 13 source languages, three—Quechua, Guarani, and Aymara—are natively present in the NLLB-200 vocabulary. For the remaining 10 languages, we add a new language token to the tokenizer and initialize its embedding by copying from the Spanish (`spa_Latn`) token embedding. This initialization provides a reasonable starting point for the source language representation while requiring no additional training data. Embedding layers remain frozen during PEFT training.

3.4 PEFT Method Configurations

We evaluate eight PEFT methods spanning four methodological families, each configured following its original paper's recommendations. Full per-method hyperparameters are provided in Appendix A. Table 2 summarizes the trainable parameter counts.

3.5 Training Configuration

All experiments share a common training configuration. We train for a maximum of 15 epochs with an effective batch size of 64 (per-device batch size 16 with 4 gradient accumulation steps). The optimizer is AdamW with weight decay 0.01, warmup ratio 0.06, and label smoothing 0.1. The default learning rate is 1×10^{-3} for all PEFT methods except DoRA, which uses 5×10^{-4} for training stability; full fine-tuning uses 3×10^{-5}. Training uses mixed-precision bf16 arithmetic where supported.

We evaluate at the end of each epoch using greedy decoding (beam size 1) and save the model checkpoint that achieves the highest development-set chrF++. Early stopping with a patience of 5 epochs terminates training if no improvement is observed. The maximum sequence length is 128 tokens for both source and target.

3.6 Inference and Evaluation

At inference time, PEFT adapter weights are merged into the base model for all methods except Prefix Tuning, which retains its PEFT wrapper due to architectural incompatibility with weight merging. Translations are generated using greedy decoding (beam size 1) with a maximum output length of 128 tokens, consistent with the decoding strategy used during development-set evaluation.

We compute chrF++ (Popović, 2017) as the primary metric and BLEU (Papineni et al., 2002) as a secondary metric, both via the SacreBLEU implementation (Post, 2018). chrF++ uses word order 2 (i.e., word unigrams and bigrams as additional features beyond character n-grams).

3.7 Statistical Analysis

We assess statistical significance using paired bootstrap resampling (Koehn, 2004) with 10,000 resamples and a two-sided test. Each method comparison is based on 13 paired observations (one mean chrF++ score per language pair, averaged over three seeds). We adopt a significance threshold of $p < 0.05$.

4 Results

4.1 Phase 1: Base Model Selection

Table 3 compares the three candidate base models. NLLB-200-distilled-600M achieves the highest mean chrF++ (26.13) across the four evaluation languages, outperforming mBART-large-50 (24.32) by 1.81 points and ByT5-small (13.92) by 12.21 points. The gap between NLLB and mBART is moderate but consistent, while ByT5 lags substantially, likely due to its byte-level tokenization producing very long sequences that exceed the 128-token training limit. NLLB also trains fastest (0.47 hours/run vs. 0.59 for mBART and 0.70 for ByT5). Based on these results, we select NLLB-200-distilled-600M as the base model for all subsequent experiments.

Model	chrF++	BLEU	Hours
NLLB-200-600M	**26.13**	**7.78**	**0.47**
mBART-large-50	24.32	5.81	0.59
ByT5-small	13.92	1.21	0.70

Table 3: Phase 1 model selection. Mean chrF++ and BLEU across four languages (Chatino, Bribri, Nahuatl, Guarani) and three seeds, all using LoRA adaptation. Hours = mean training time per language-seed run.

Method	chrF++	BLEU	Hours
OFT	**26.63**	7.82	1.63
Full FT	26.13	**7.84**	0.96
LoRA	25.27	7.16	0.88
AdaLoRA	23.77	6.27	0.94
DoRA	23.50	6.10	1.29
BOFT	22.99	5.90	2.56
IA^3	21.03	5.00	0.86
Prefix Tuning	19.63	4.01	0.66
VeRA	18.69	4.14	0.97

Table 4: Phase 2 development-set results. Mean chrF++ and BLEU across 13 language pairs and 3 seeds. Hours = mean training time per language-seed run.

4.2 Phase 2: Development Set Performance

4.2.1 Overall Method Ranking

Table 4 presents the overall development-set results. OFT achieves the highest mean chrF++ (26.63), followed closely by full fine-tuning (26.13) and LoRA (25.27). The top three methods are separated by just 1.36 chrF++ points, while the gap between the best and worst methods (OFT vs. VeRA) spans 7.94 points. Notably, OFT surpasses full fine-tuning while training only 0.28% of the parameters, demonstrating that parameter efficiency need not come at the cost of translation quality.

The middle tier—AdaLoRA (23.77), DoRA (23.50), and BOFT (22.99)—achieves moderate performance, trailing LoRA by 2.3–3.6 points despite similar or fewer trainable parameters. The bottom tier—IA^3 (21.03), Prefix Tuning (19.63), and VeRA (18.69)—shows that the most parameter-efficient methods sacrifice substantial translation quality.

Full fine-tuning achieves the highest BLEU (7.84) despite ranking second in chrF++, reflecting chrF++'s greater sensitivity to character-level overlap in morphologically rich languages.

4.2.2 Per-Language Results

OFT wins the most language pairs on the development set, achieving the highest chrF++ for 9 of 13 languages. Full fine-tuning wins 3 languages (Aymara, Wayuu, Quechua—all in the high tier), and VeRA wins 1 (Guarani, by a negligible margin). No single method dominates across all languages, but the top-three methods (OFT, Full, LoRA) are remarkably consistent, each appearing in the top 3 for at least 10 of 13 language pairs. Lower-ranked methods show high variance—Prefix Tuning, for instance, achieves 23.75 chrF++ on Chatino (outperforming LoRA) but only 12.21 on Raramuri.

Method	chrF++	BLEU
Full FT	**25.12**	**7.18**
OFT	25.06	6.27
LoRA	22.70	5.32
BOFT	21.46	4.76
AdaLoRA	21.27	5.03
DoRA	20.94	4.76
IA^3	20.19	4.57
Prefix Tuning	19.57	3.80
VeRA	18.43	3.86

Table 5: Phase 3 test-set results. Mean chrF++ and BLEU across 13 language pairs and 3 seeds.

4.2.3 Performance by Resource Tier

Performance varies substantially across resource tiers. In the *very-low* tier (<5K sentences), OFT leads by a wide margin (23.98 chrF++), outperforming the second-best method (full fine-tuning, 20.69) by 3.29 points. This advantage narrows in the *low* tier (OFT 26.04 vs. full 25.57) and *medium* tier (OFT 24.94 vs. full 24.64). In the *high* tier (>25K sentences), full fine-tuning takes the lead (32.51 vs. OFT 31.45), suggesting that full parameter updates become advantageous when sufficient data are available.

LoRA consistently ranks third across all tiers, maintaining a 1–2 point gap behind the leader. The bottom-tier methods (Prefix Tuning, VeRA) show the steepest performance degradation as data decreases: VeRA drops from 27.94 chrF++ in the high tier to 11.07 in the very-low tier, a 16.87-point decline.

4.3 Phase 3: Test Set Performance

4.3.1 Overall Test Results

Table 5 presents the held-out test results. Full fine-tuning achieves the highest mean chrF++ (25.12), narrowly surpassing OFT (25.06). LoRA remains third (22.70), followed by BOFT (21.46) and AdaLoRA (21.27). The bottom three methods maintain their rankings: IA^3 (20.19), Prefix Tuning (19.57), and VeRA (18.43).

On the test set, full fine-tuning wins 8 of 13 language pairs, with OFT winning the remaining 5. OFT retains its advantage on very-low-tier languages (Chatino, Ashaninka) and several low-tier pairs, while full fine-tuning dominates the medium and high tiers.

4.3.2 Dev-to-Test Generalization

Method rankings are largely stable between development and test sets. Six of nine methods retain their dev-set rank on the test set; the largest rank change is BOFT, which rises from 6th to 4th. The top-two swap (OFT $\rightarrow$ 2nd, Full $\rightarrow$ 1st) reflects a difference of only 0.06 chrF++ on test, well within noise.

Dev-to-test chrF++ drops vary considerably across methods. LoRA (-2.57), DoRA (-2.56), and AdaLoRA (-2.51) show the largest degradation, suggesting some overfitting to development data. Full fine-tuning (-1.01) and OFT (-1.57) degrade moderately. Prefix Tuning (-0.06) and VeRA (-0.27) show minimal dev-to-test gaps, though this stability reflects consistently low performance rather than robust generalization.

4.4 Parameter Efficiency Analysis

Figure 1 shows that parameter count is a poor predictor of translation quality: OFT (1.71M parameters) outperforms AdaLoRA (1.77M) by 2.86 chrF++ despite nearly identical budgets, and IA^3 (74K) outperforms VeRA (616K) by 2.33 points with 8$\times$ fewer parameters, confirming that adaptation mechanism matters more than parameter count. Evaluated against three objectives simultaneously—maximize chrF++, minimize trainable parameters, and minimize training time—seven of nine methods are Pareto-optimal, each taking a unique position on the tradeoff frontier: OFT at the quality front, IA^3 at the minimum-parameter front, Prefix Tuning at the minimum-time front, and LoRA, Full FT, DoRA, and BOFT at intermediate positions. Only AdaLoRA and VeRA are fully dominated: LoRA strictly beats AdaLoRA on all three dimensions, and IA^3 strictly beats VeRA; practitioners have no reason to prefer either method.

4.5 Statistical Significance

Paired bootstrap resampling (10,000 resamples, two-sided, $\alpha = 0.05$) was applied to both the development and test sets; full results are in Appendix B. On the development set, 32 of 36 pairwise comparisons are significant. The four non-significant pairs are: Full vs. OFT ($p = 0.097$), AdaLoRA vs. DoRA ($p = 0.052$), IA^3 vs. Prefix Tuning ($p = 0.135$), and Prefix Tuning vs. VeRA ($p = 0.229$).

On the test set, 29 of 36 comparisons are significant. Full vs. OFT becomes even less significant ($p = 0.425$), further supporting the central finding that there is no statistically significant difference between the two methods. Three additional pairs lose significance on test: AdaLoRA vs. BOFT ($p = 0.291$), IA^3 vs. Prefix Tuning ($p = 0.293$), and AdaLoRA vs. Prefix Tuning ($p = 0.111$), along with BOFT vs. Prefix Tuning ($p = 0.074$) and DoRA vs. Prefix Tuning ($p = 0.136$). These new non-significant pairs cluster around Prefix Tuning, reflecting that the bottom tier of methods are genuinely close in absolute performance and their small differences are overwhelmed by noise on held-out data.

Across both sets, the non-significance of Full vs. OFT is the most consequential result: there is no statistically significant difference in translation quality between OFT and full fine-tuning, while OFT trains fewer than 0.3% of the parameters. This is the central finding of the study.

4.6 Error Analysis

Aggregate metrics like chrF++ and BLEU summarize translation quality as a single number, but two methods with similar scores can exhibit qualitatively different failure modes. To characterize these differences, we compute four diagnostic metrics over the test-set translations: *repetition*, *source copy ratio*, *length ratio*, and *entity preservation F1*.

Repetition measures self-repetition as $1 - (\text{unique n-grams}/\text{total n-grams})$, macro-averaged over $n \in \{1, 2, 3\}$ (range $[0, 1]$; lower is better). **Source copy ratio** computes the multi-set intersection of hypothesis tokens with source tokens, divided by hypothesis length (range $[0, 1]$; lower is better), capturing the degree to which the model copies source text rather than translating. **Length ratio** divides hypothesis length by reference length (ideal = 1.0); values below 1 indicate under-generation and values above 1 indicate over-generation, aggregated via the median at the sentence level to resist outliers. **Entity F1** extracts numeric entities from hypothesis and reference via regex and computes F1 over their multi-set overlap (range $[0, 1]$; higher is better), restricted to the subset of sentences whose reference contains at least one entity.

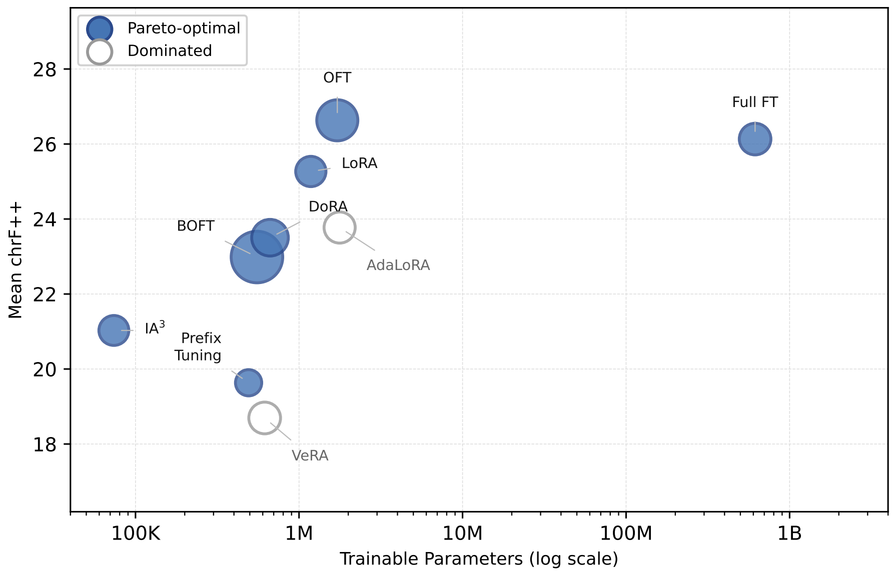

Figure 1: Efficiency–quality tradeoff across all nine methods (development set). Each bubble is one method; bubble area is proportional to mean training hours per language-pair run (BOFT largest at 2.6 h; Prefix Tuning smallest at 0.7 h). Filled circles are Pareto-optimal across three objectives simultaneously (maximize chrF^{++}, minimize trainable parameters, minimize training time); hollow circles (AdaLoRA, VeRA) are strictly dominated on all three. Full FT sits at 615M parameters (far right); all PEFT methods fall below 2M.

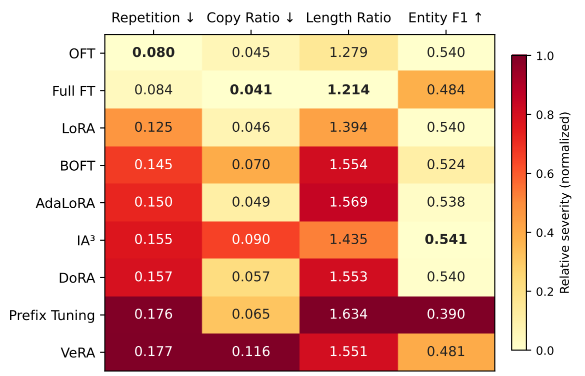

Figure 2: Heatmap of diagnostic metrics by method. Color intensity reflects relative severity (normalized per column); yellow indicates better performance and red indicates worse. Cell values are raw (un-normalized) scores.

Full diagnostic metrics are reported in Appendix C, and Figure 2 summarizes the failure profiles. OFT and full fine-tuning produce the least repetition (0.080 and 0.084 respectively) and lowest source copy ratios (0.045 and 0.041), confirming that their chrF++ advantage reflects genuinely cleaner translations rather than superficial n-gram matching. Full fine-tuning also achieves the length ratio closest to 1.0 (1.214), while all other methods over-generate by 28–63%.

However, aggregate chrF++ rankings do not predict all dimensions of quality. Full fine-tuning achieves the best chrF++ but the *third-worst* entity F1 (0.484), behind only Prefix Tuning (0.390) and VeRA (0.481), indicating that it drops or distorts numeric content more often than most PEFT methods. Conversely, IA^3—which ranks 7th in chrF++—achieves the *best* entity F1 (0.541), suggesting that its minimal activation-scaling intervention preserves factual content more faithfully than methods that modify weight matrices directly. VeRA exhibits the worst repetition (0.177) and source copy ratio (0.116), consistent with its low chrF++ and suggesting that its shared-random-matrix parameterization struggles to learn a genuine translation function under low-resource conditions.

Failure modes intensify as training data decreases. In the very-low resource tier (<5K sentences), repetition scores roughly triple compared to the high tier for most methods: AdaLoRA rises from 0.105 to 0.427, and BOFT from 0.123 to 0.404. Length ratios become extreme, with AdaLoRA (2.99), BOFT (2.85), and DoRA (2.83) producing hypotheses nearly three times the reference length—a signature of repetitive over-generation. OFT degrades most gracefully: its very-low repetition (0.216) and length ratio (1.59) remain the best in the tier, which explains the 3-point chrF++ margin it holds over all other methods in that setting. The full breakdown by resource tier is provided in Appendix D.

5 Conclusion

5.1 Summary of Findings

This study compared eight PEFT methods and full fine-tuning for neural machine translation of 13 indigenous languages of the Americas into Spanish. Four findings stand out.

First, OFT and full fine-tuning do not differ significantly in translation quality ($p = 0.097$), with OFT training only 0.28% of the base model's parameters. On the development set, OFT leads (26.63 chrF++); on the held-out test set, full fine-tuning leads marginally (25.12 vs. 25.06). This demonstrates that orthogonal transformations preserve pretrained knowledge effectively under data scarcity.

Second, PEFT method effectiveness interacts strongly with data availability. OFT excels in the very-low tier (<5K sentences), leading full fine-tuning by 3.29 chrF++ points on dev and 3.00 points on test. Full fine-tuning takes the lead in the high tier (>25K sentences), where sufficient data mitigate overfitting. LoRA offers consistent third-place performance across all tiers.

Third, parameter count alone does not predict translation quality. OFT (1.71M parameters) outperforms AdaLoRA (1.77M) by 2.86 chrF++ points despite similar parameter budgets, and IA^3 (74K parameters) outperforms VeRA (616K) by 2.33 points despite having 8× fewer parameters. The adaptation mechanism matters more than the number of trainable parameters.

Fourth, error analysis reveals that aggregate scores mask method-specific failure profiles. Full fine-tuning achieves the best chrF++ but the third-worst entity preservation F1 (0.484), while IA^3 shows the opposite pattern (7th in chrF++, best entity F1 at 0.541). This suggests that method selection should consider which dimensions of translation quality matter most for a given application, not just a single aggregate score.

5.2 Practical Recommendations

For practitioners working with low-resource indigenous languages, we offer the following guidance:

- **Base model:** NLLB-200-distilled-600M provides the strongest starting point among the models tested, particularly for languages not in its pretraining vocabulary.
- **Default PEFT method:** OFT is recommended as the default choice, achieving the best or near-best quality across resource levels with a 359× parameter reduction.
- **Speed-constrained settings:** LoRA offers a strong alternative with training time half that of OFT and consistently strong performance.
- **Very-low-tier languages:** OFT is particularly advantageous when training data is extremely scarce (<5K sentences), outperforming all alternatives by a substantial margin.
- **Methods to avoid:** VeRA and Prefix Tuning consistently underperform, ranking 8th–9th across conditions. BOFT requires the most training time while achieving below-median quality.

5.3 Future Work

Future work should explore bidirectional translation (Spanish to indigenous languages), larger base models (NLLB-200-1.3B, 3.3B), per-method hyperparameter optimization (e.g., rank sweeps for LoRA, target module ablations for OFT), combinations of PEFT methods, data augmentation via back-translation for very-low tiers, and human evaluation with indigenous language community members.

Limitations

Target module configurations differ across PEFT methods, following each method's original paper recommendations rather than a single standardized configuration. This means methods differ in both adaptation mechanism and scope of modified parameters, which is realistic for practitioners but prevents a pure apples-to-apples comparison of mechanisms.

Learning rates differ across conditions: 1×10^{-3} for most PEFT methods, 5×10^{-4} for DoRA, and 3×10^{-5} for full fine-tuning. These rates reflect necessary tuning for training stability but introduce an additional variable.

No per-method hyperparameter search was conducted (e.g., rank sweeps for LoRA, target module ablations). All methods use a single recommended configuration; results therefore reflect default performance rather than each method's ceiling.

Statistical power is limited by $n = 13$ language pairs for the paired bootstrap test. While 32 of 36 pairwise comparisons reach significance, only large systematic differences are detectable at this sample size.

ByT5-small (∼300M parameters) is substantially smaller than NLLB (∼615M) and mBART (∼611M), making the Phase 1 comparison not purely architecture-controlled.

Evaluation relies on automatic metrics only (chrF++, BLEU). Human evaluation of translation quality, adequacy, and fluency would provide complementary insight, particularly for morphologically complex languages where automatic metrics may not capture meaning preservation.

All translations are unidirectional (indigenous language to Spanish). The reverse direction poses different challenges (generation in morphologically rich languages) and may yield different method rankings.

Greedy decoding (beam size 1) was used throughout for consistency. Beam search could alter relative method rankings, particularly for methods that produce higher-entropy output distributions.

Only one base model scale was tested (600M distilled). Results may not transfer to larger NLLB variants (1.3B, 3.3B), where the relative advantage of PEFT over full fine-tuning could differ.

Ethical Statement

All parallel corpora used in this work were drawn from the AmericasNLP shared-task datasets, which were compiled in collaboration with indigenous language communities and released for research purposes. We have not collected, redistributed, or modified any community data, and we follow the terms under which these datasets were made available.

The goal of this research is to lower the computational barrier to building NMT systems for underserved languages rather than to produce deployment-ready translation tools. The chrF++ scores reported here, while meaningful for benchmarking, reflect the inherent difficulty of low-resource MT: translations generated by these systems may be fluent but inaccurate. Any practical use of models trained on these data should involve review by speakers of the target language community.

We recognize that the development of NLP tools for endangered and minority languages carries both promise and risk. At best, such tools support language documentation, education, and access to information. At worst, they can be used to displace human translators, misrepresent community voices, or create a false impression of language vitality. We encourage future work in this space to be conducted in partnership with the communities whose languages are involved.

Acknowledgments

This work used the Augie High-Performance Computing cluster, funded by award NSF 2018933, at Villanova University.

References

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 1538–1548. Association for Computational Linguistics.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G. Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, and 15 others. 2022. Findings of the second AmericasNLP competition on speech-to-text translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 6282–6293. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 388–395. Association for Computational Linguistics.

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. 2024. VeRA: Vector-based random matrix adaptation. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, volume 1, pages 4582–4597. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024a. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and David Ha. 2024b. Parameter-efficient orthogonal finetuning via butterfly factorization. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2021)*, pages 202–217. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Celebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, and Al Youngblood. 2022. No language left behind: Scaling human-centered machine translation.

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

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT 2017)*, pages 612–618. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT 2018)*, pages 186–191. Association for Computational Linguistics.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. 2023. Controlling text-to-image diffusion by orthogonal finetuning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics (TACL)*, 10:291–306.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. AdaLoRA: Adaptive budget allocation for parameter-efficient finetuning. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.

A PEFT Method Configurations

Each method is configured following the recommendations in its original paper rather than imposing a single standardized configuration across all methods. We consider this the most informative comparison for practitioners as it reflects the performance each method achieves under its intended operating conditions.

LoRA: rank 8, $\alpha = 16$, dropout 0.1, targeting query and value projections (q, v).

AdaLoRA: initial rank 12, target rank 8, $\alpha = 16$, dropout 0.1, targeting q and v, with orthogonality regularization weight 0.5.

DoRA: rank 4, $\alpha = 16$, dropout 0.1, targeting q and v, with weight decomposition into magnitude and direction components enabled.

VeRA: rank 256, targeting q and v, with shared frozen random matrices and learned per-layer scaling vectors initialized to 0.1.

IA^3: targeting key, value, and feedforward output projections (k, v, wo), with scaling vectors initialized to 1.0 (identity).

OFT: rank 32, targeting all attention projections (q, k, v, o), with block-diagonal Cayley parameterization initialized to the identity transformation.

BOFT: block size 4, butterfly factor 1, targeting q, k, v, and o projections.

Prefix Tuning: 20 virtual tokens with MLP reparameterization during training.

Full fine-tuning serves as the baseline, updating all 615 million parameters.

A noteworthy design choice is that low-rank methods (LoRA, AdaLoRA, DoRA, VeRA) target only query and value projections, while orthogonal methods (OFT, BOFT) target all four attention projections. This follows the respective authors' recommendations and reflects a real-world practitioner scenario, although it means the methods differ in mechanism and scope.

B Pairwise Bootstrap Significance Tests

Tables 6 and 7 report all 36 pairwise bootstrap significance tests on the development and test sets respectively. Each test uses 10,000 paired resamples over 13 language pairs, two-sided, at $\alpha = 0.05$. ΔchrF++ = chrF++(Method A) − chrF++(Method B); Method A is always the higher-scoring method so $\Delta > 0$ throughout. Rows are sorted by ΔchrF++ descending within each significance tier; a horizontal rule separates significant ($p < 0.05$) from non-significant pairs.

Method A	Method B	ΔchrF++	p	Sig.
OFT	VeRA	7.94	<0.001	✓
Full FT	VeRA	7.44	<0.001	✓
OFT	Prefix Tuning	7.00	<0.001	✓
LoRA	VeRA	6.57	<0.001	✓
Full FT	Prefix Tuning	6.50	<0.001	✓
LoRA	Prefix Tuning	5.63	<0.001	✓
OFT	IA^3	5.60	<0.001	✓
Full FT	IA^3	5.10	<0.001	✓
AdaLoRA	VeRA	5.08	<0.001	✓
DoRA	VeRA	4.81	<0.001	✓
BOFT	VeRA	4.29	<0.001	✓
LoRA	IA^3	4.24	<0.001	✓
OFT	BOFT	3.65	<0.001	✓
Full FT	BOFT	3.15	<0.001	✓
OFT	DoRA	3.13	<0.001	✓
OFT	AdaLoRA	2.86	<0.001	✓
Full FT	DoRA	2.63	<0.001	✓
AdaLoRA	IA^3	2.74	<0.001	✓
Full FT	AdaLoRA	2.36	<0.001	✓
LoRA	BOFT	2.28	<0.001	✓
DoRA	IA^3	2.47	<0.001	✓
IA^3	VeRA	2.33	<0.001	✓
BOFT	IA^3	1.96	<0.001	✓
LoRA	DoRA	1.77	<0.001	✓
LoRA	AdaLoRA	1.49	<0.001	✓
OFT	LoRA	1.36	<0.001	✓
Full FT	LoRA	0.86	<0.001	✓
AdaLoRA	Prefix Tuning	4.14	0.002	✓
DoRA	Prefix Tuning	3.87	0.003	✓
BOFT	Prefix Tuning	3.35	0.003	✓
AdaLoRA	BOFT	0.79	0.016	✓
DoRA	BOFT	0.52	0.017	✓
AdaLoRA	DoRA	0.27	0.052	
OFT	Full FT	0.50	0.097	
IA^3	Prefix Tuning	1.39	0.135	
Prefix Tuning	VeRA	0.94	0.229	

Table 6: All 36 pairwise bootstrap significance tests, development set (Phase 2). Sorted by ΔchrF++ descending within each tier.

C Aggregate Error Analysis

Table 8 reports four diagnostic metrics averaged across all 13 language pairs and 3 seeds on the test set. Metric definitions are given in Section 4.6.

D Error Analysis by Resource Tier

Table 9 provides the full breakdown of diagnostic metrics by method and resource tier, aggregated across languages within each tier and across 3 seeds.

Method A	**Method B**	**ΔchrF++**	p	**Sig.**
Full FT	VeRA	6.69	<0.001	✓
OFT	VeRA	6.64	<0.001	✓
Full FT	Prefix Tuning	5.55	<0.001	✓
OFT	Prefix Tuning	5.49	<0.001	✓
Full FT	IA^3	4.93	<0.001	✓
OFT	IA^3	4.87	<0.001	✓
LoRA	VeRA	4.27	<0.001	✓
Full FT	DoRA	4.17	<0.001	✓
OFT	DoRA	4.12	<0.001	✓
Full FT	AdaLoRA	3.85	<0.001	✓
OFT	AdaLoRA	3.80	<0.001	✓
Full FT	BOFT	3.66	<0.001	✓
OFT	BOFT	3.60	<0.001	✓
LoRA	Prefix Tuning	3.12	<0.001	✓
BOFT	VeRA	3.03	<0.001	✓
AdaLoRA	VeRA	2.84	<0.001	✓
DoRA	VeRA	2.52	<0.001	✓
LoRA	IA^3	2.50	<0.001	✓
Full FT	LoRA	2.42	<0.001	✓
OFT	LoRA	2.37	<0.001	✓
IA^3	VeRA	1.77	<0.001	✓
LoRA	DoRA	1.75	<0.001	✓
LoRA	AdaLoRA	1.43	<0.001	✓
BOFT	IA^3	1.27	<0.001	✓
AdaLoRA	IA^3	1.07	<0.001	✓
LoRA	BOFT	1.24	0.004	✓
DoRA	IA^3	0.75	0.005	✓
BOFT	DoRA	0.52	0.026	✓
AdaLoRA	DoRA	0.32	0.043	✓
BOFT	Prefix Tuning	1.89	0.074	
AdaLoRA	Prefix Tuning	1.69	0.111	
DoRA	Prefix Tuning	1.37	0.136	
Prefix Tuning	VeRA	1.15	0.144	
BOFT	AdaLoRA	0.19	0.291	
IA^3	Prefix Tuning	0.62	0.293	
Full FT	OFT	0.06	0.425	

Table 7: All 36 pairwise bootstrap significance tests, test set (Phase 3). Sorted by ΔchrF++ descending within each tier. Note that Full FT and OFT are the final non-significant pair ($p = 0.425$), with Full FT leading OFT by only 0.06 chrF++ on held-out data.

Method	Repetition	Copy Ratio	Length Ratio	Entity F1
OFT	**0.080**	0.045	1.279	0.540
Full FT	0.084	**0.041**	**1.214**	0.484
LoRA	0.125	0.046	1.394	0.540
BOFT	0.145	0.070	1.554	0.524
AdaLoRA	0.150	0.049	1.569	0.538
IA^3	0.155	0.090	1.435	**0.541**
DoRA	0.157	0.057	1.553	0.540
Prefix Tuning	0.176	0.065	1.634	0.390
VeRA	0.177	0.116	1.551	0.481

Table 8: Error analysis: diagnostic metrics averaged across 13 language pairs and 3 seeds on the test set. Bold indicates best value per column (lowest for Repetition and Copy Ratio, closest to 1.0 for Length Ratio, highest for Entity F1).

Method	Tier	Repetition	Copy Ratio	Length Ratio	Entity F1
Full FT	high	0.049	0.035	1.043	0.510
OFT	high	0.077	0.035	1.122	0.552
LoRA	high	0.101	0.040	1.178	0.584
AdaLoRA	high	0.105	0.038	1.161	0.590
BOFT	high	0.123	0.040	1.192	0.520
VeRA	high	0.129	0.033	1.109	0.535
DoRA	high	0.131	0.041	1.205	0.596
Prefix Tuning	high	0.143	0.037	1.225	0.421
IA^3	high	0.146	0.037	1.180	0.569
OFT	low	0.044	0.055	1.240	0.615
Full FT	low	0.060	0.037	1.228	0.542
BOFT	low	0.062	0.083	1.279	0.603
LoRA	low	0.072	0.046	1.320	0.587
AdaLoRA	low	0.076	0.063	1.287	0.607
DoRA	low	0.077	0.077	1.288	0.596
IA^3	low	0.083	0.105	1.282	0.606
Prefix Tuning	low	0.091	0.094	1.465	0.475
VeRA	low	0.127	0.150	1.428	0.531
Full FT	medium	0.043	0.051	1.139	0.430
OFT	medium	0.051	0.047	1.281	0.484
LoRA	medium	0.102	0.054	1.396	0.493
BOFT	medium	0.114	0.092	1.451	0.494
AdaLoRA	medium	0.119	0.049	1.448	0.484
IA^3	medium	0.122	0.132	1.347	0.499
DoRA	medium	0.131	0.058	1.441	0.494
VeRA	medium	0.195	0.155	1.552	0.422
Prefix Tuning	medium	0.222	0.069	1.951	0.283
OFT	very-low	0.216	0.036	1.589	0.487
Full FT	very-low	0.268	0.035	1.595	0.433
Prefix Tuning	very-low	0.305	0.038	1.950	0.388
VeRA	very-low	0.310	0.094	2.456	0.418
LoRA	very-low	0.315	0.040	1.862	0.473
IA^3	very-low	0.377	0.056	2.301	0.450
BOFT	very-low	0.404	0.047	2.852	0.435
DoRA	very-low	0.406	0.043	2.827	0.436
AdaLoRA	very-low	0.427	0.041	2.990	0.429

Table 9: Diagnostic metrics by method and resource tier (test set). Within each tier, methods are sorted by repetition (ascending).