Optimization of Voice Translation Systems for Indigenous Languages: Retraining the NLLB-200 Model for the Quechua–Spanish Pair

Mitzuko Davis Quispe Callañaupa
mitzukodavis@gmail.com

Max Erixon Toledo Bernal
maxtoledo142@gmail.com

Ronil Nilo Torres Bautista
roni36608@gmail.com

Patrick Michael Pumaccahua Huallpa
patrickpuccahua@gmail.com

Abstract

This article describes the fine-tuning and incremental retraining process of the massive NLLB-200 model applied to the Quechua (Chanka and Collao variants) and Spanish language pair. Using a curated dataset of 22,891 parallel pairs, a robust cleaning strategy and optimized training for consumer hardware (NVIDIA RTX 3060) were implemented. The results demonstrate a progressive improvement in the BLEU metric, reaching a competitive state for translation tasks in low-resource scenarios, in line with the challenges posed by the IWSLT 2026 shared task.

1 Introduction

Machine translation of indigenous languages presents unique challenges due to the scarcity of high-quality parallel data and the complex agglutinative morphology of Quechua. This work falls within the context of the "Low Resources" track of IWSLT 2026, which aims to promote translation technology for dialects lacking supervised data at scale. The main objective is to evaluate the capacity of a massively pre-trained model to adapt to a specific domain of text-to-text translation derived from speech transcripts.

2 Methodology

2.1 Data and Cleaning

An initial dataset of 22,891 records with voice and text metadata was used. The robust cleaning process consisted of:

- **Quality filtering:** Removal of NaN values, excessive whitespace, and correction of malformed characters.
- **Deduplication:** Duplicates were removed based on the source column (Quechua) to avoid overfitting, resulting in a final dataset of 21,926 parallel pairs.
- **Split:** An 80% partition was applied for training, 10% for validation, and 10% for final testing.

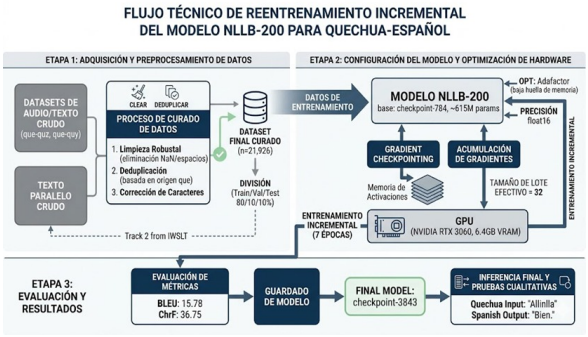

Figure 1: Fine-Tuning Process pipeline.

2.2 Model Architecture

The NLLB-200 (Costa-jussà et al., 2022) (No Language Left Behind) model was selected as the basis, specifically using checkpoint-784 from previous training to allow for incremental learning. The model has approximately 615 million parameters.

2.3 Training Setup

To handle hardware limitations (6.4 GB VRAM), several memory optimization techniques were implemented:

- **Gradient Checkpointing:** Reduction of memory usage by 25–30%.
- **Gradient Accumulation:** Configured in 8 steps with a batch size of 4, simulating an effective batch size of 32.
- **Precision:** Use of `float16` for fast and efficient inference.
- **Optimizer:** Adafactor (Shazeer and Stern, 2018), chosen for its low memory footprint.

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 316–317

3 Results and Evaluation

The retraining was successfully completed over 7 epochs, reaching a final save point at checkpoint-3843.

3.1 Quality Metrics

The results on the validation set showed a trend of constant improvement:

- **Final BLEU:** 15.78
- **Final ChrF:** 36.75

Table 1 summarizes the evolution of the main metrics across training epochs.

Epoch	Train Loss	Val. Loss	BLEU	ChrF
1	17.94	2.19	13.19	34.28
7	15.20	1.97	15.78	36.75

Table 1: Training and validation metrics across epochs.

4 Discussion

Qualitative tests demonstrate that the model is capable of effectively translating short phrases and greetings (e.g., *"Allinlla"* → "Good"; *"Napayku"* → "Greetings"). However, in complex sentences the prediction can diverge significantly from the reference, suggesting the need to incorporate data augmentation techniques or the use of larger pre-trained language models if hardware resources allow (Vaswani et al., 2017; Papineni et al., 2002).

According to the official IWSLT 2026 evaluation, our system (`velo.st.unconstrained.primary`) was evaluated in the Unconstrained track, achieving a BLEU score of 8.9 and a chrF2 score of 39.9. The performance drop from our validation metrics (BLEU 15.78) to the official test set highlights a common challenge in low-resource machine translation: while the model memorizes and translates simpler in-domain structures effectively, it struggles to generalize to the more complex or unseen sentences of the test set. Furthermore, when observing other submissions in the same track (such as the quespa systems, which achieved BLEU scores up to 27.2), our results suggest that our current pipeline is prone to overfitting. This confirms that to properly leverage the NLLB-200 architecture under strict hardware limitations, future work must prioritize robust data augmentation strategies to improve generalization.

5 Conclusion

This work demonstrates that fine-tuning NLLB-200 on a carefully cleaned Quechua–Spanish dataset yields measurable gains in BLEU and ChrF even under strict memory constraints. Future work will explore data augmentation and larger model variants to further close the performance gap in this low-resource setting (Ortega et al., 2020; Adelani et al., 2026).

References

David Ifeoluwa Adelani, Víctor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sébastien Bratières, Marina Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the IWSLT campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, USA. Association for Computational Linguistics.

Marta R. Costa-jussà and 1 others. 2022. No Language left Behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

John E. Ortega and 1 others. 2020. Overcoming resistance: The case of Quechua and Spanish machine translation. In *Proceedings of the 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas*.

Kishore Papineni and 1 others. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *arXiv preprint arXiv:1805.09843*.

Ashish Vaswani and 1 others. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.