# Fine-tuning Whisper Across 81 Languages

**Shivam Singh**
University of California, San Diego
shs046@ucsd.edu

**Alex Warstadt**
University of California, San Diego
awarstadt@ucsd.edu

## 1 Introduction

Whisper (Radford et al., 2023) is a family of multilingual automatic speech recognition (ASR) models trained on 680,000 hours of web audio spanning 99 languages. Out of the box, it performs well on high-resource languages like English, Spanish, and French, but its accuracy drops sharply on lower-resource languages. For example, word error rates (WER) exceed 50% for languages like Bengali, Tamil, and Georgian, and reach 100% for Khmer, Burmese, and Lao.

Prior work showed that fine-tuning Whisper on monolingual ASR data can reduce WER by as much as 70% for some languages (Liu et al., 2024), but this approach has yet to be scaled beyond a handful of languages. We fill this gap by fine-tuning Whisper large-v3 on each of the 81 languages in the FLEURS benchmark (Conneau et al., 2023). We find that fine-tuning improves speech recognition for all 81 languages tested, reducing WER by nearly 30% on average. However, the size of the improvement varies widely, and we show that the language's writing system is the best predictor of how much fine-tuning helps. We further demonstrate that this script-family pattern is mediated by tokenizer fertility: Whisper's BPE compression ratio (characters per token) correlates with WER reduction at Spearman $\rho \approx -0.78$, and the dependence persists after partialing out per-language pretraining hours. We intend to release model weights upon publication, providing, to the best of our knowledge, the largest fleet of open-source language-specialized ASR models.

## 2 Methods

We fine-tune Whisper large-v3 independently for each of the 81 FLEURS languages. Every language uses the same pipeline: load a fresh copy of the pretrained model, train on the FLEURS training split, evaluate on the validation split, and save the checkpoint with the lowest WER. Early stopping triggers after 3 epochs without improvement or 12 epochs total. Audio is resampled to 16 kHz, capped at 30 s, and trained with AdamW, gradient clipping at 1.0, and mixed precision on a single A100. We use separate learning rates for the encoder and decoder, setting the encoder learning rate to $0.3\times$ the decoder learning rate (this ratio was consistent across both hyperparameter sweeps).

| Hyperparameter | Config A | Config B |
|---|---|---|
| Learning rate | $1.18 \times 10^{-5}$ | $6.48 \times 10^{-6}$ |
| Grad. accumulation | 24 | 16 |
| LR scheduler | Exponential | Cosine |
| LR decay rate | 0.9 | 0.95 |
| Encoder LR ratio | 0.3 | |
| Warmup steps | 150 | |
| Max epochs | 12 | |
| Early stopping | Patience 3 | |

Table 1: Hyperparameter configurations. Values below the divider are shared across configs.

Two hyperparameter configurations were used (Table 1), found by Bayesian sweeps on Hindi and German. These two languages were chosen as endpoints of Whisper's coverage spectrum—Hindi (Devanagari, mid-coverage) and German (Latin, high-coverage)—to test whether a single set of hyperparameters generalizes across script families. The two configurations correlate at $\rho = 0.996$ on the resulting 81-language WER vector, indicating that our conclusions are robust to hyperparameter choice within a reasonable range.

## 3 Results

Figure 1 plots zero-shot WER from pretrained Whisper large-v3 against fine-tuned WER for each target language under both configurations. Every point falls below the diagonal, meaning fine-tuning helps across the board. Under Config A, the mean WER drops from 34.8% to 24.6%, and 50 out of 81

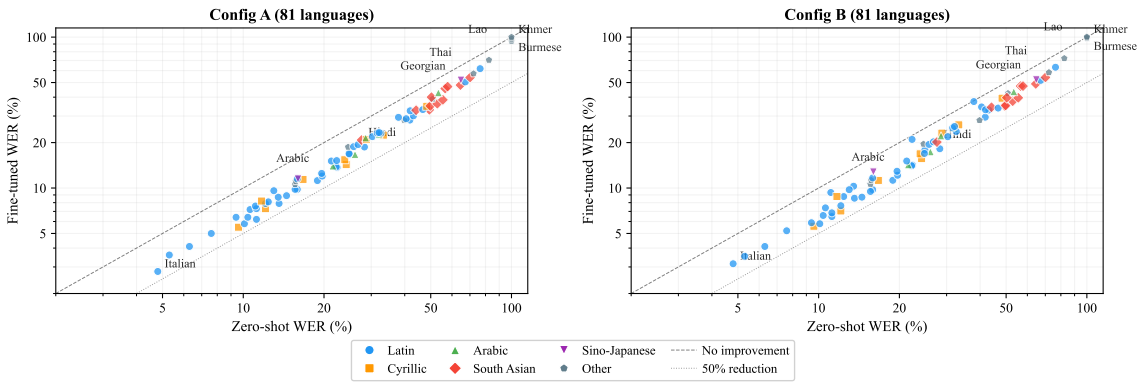


Figure 1: Zero-shot vs. fine-tuned WER under Config A (left) and Config B (right). Points are colored by script family. The dashed line marks no improvement; the dotted line marks 50% WER reduction. The two configurations produce nearly identical patterns (Spearman $\rho = 0.996$).

languages reach a fine-tuned WER below 25%.

The writing system is a strong predictor of improvement. Latin and Cyrillic scripts see 30–45% WER reductions, reaching single-digit error rates for Italian (2.8%), Spanish (3.6%), and Russian (5.5%). Brahmic scripts (South Asian languages) improve ≈25% relative but remain at 20–54% WER. Languages with unique or logographic scripts benefit least: Thai drops only 14% (82.4%→70.5%), Burmese 3%, and Khmer not at all.

| Script | N | Mean FT | Rel. ↓ | FT Range |
|---|---|---|---|---|
| Latin | 44 | 16.2% | 32.8% | 2.8–61.8% |
| Cyrillic | 9 | 15.6% | 34.3% | 5.5–34.8% |
| Arabic | 5 | 23.5% | 29.0% | 14.0–42.8% |
| Brahmic | 11 | 39.2% | 25.7% | 20.8–54.0% |
| Sino-Japanese | 2 | 31.7% | 24.0% | 11.5–51.9% |
| Other | 10 | 52.7% | 18.1% | 10.6–100.1% |

Table 2: Fine-tuned WER by script family (Config A). Rel. ↓ = mean relative WER reduction from zero-shot.

Table 2 summarizes these trends by script family. This pattern holds even for poorly represented language families: Vietnamese (Austroasiatic, Latin script) reaches 7.6% WER, while Thai (Kra–Dai, unique script) stays at 70.5%. Similarly, Lingala (Niger-Congo, Latin) reaches 15.2%, while Georgian (Kartvelian, unique script) only reaches 57.3%. Config B produces nearly identical results: mean WER 25.6% vs. 24.6% for Config A, Spearman $\rho \approx 0.996$ ($p < 10^{-74}$) between the two, and a mean absolute difference of just 1.1 pp.

**Tokenizer compression ratio predicts fine-tuning headroom.** We directly link the script-family pattern to tokenizer behavior. For each language, we compute Whisper's BPE compression ratio (Unicode characters per token) on the FLEURS test transcripts. Across our 81 languages, this compression ratio correlates with relative WER reduction at Spearman $\rho \approx -0.78$: scripts that Whisper's tokenizer splits into many short fragments (low chars/token) gain less from fine-tuning. The same finding holds when controlling for Whisper's per-language pretraining duration via partial Spearman correlation (partial $\rho \approx -0.79$), indicating that the effect is not simply "less pretraining data → less benefit" but specifically reflects source-tokenizer fit.

## 4 Discussion

Our results replicate and scale up Liu et al. (2024), confirming that monolingual fine-tuning consistently improves Whisper across all 81 FLEURS languages, and suggest a simple solution to the curse of multilinguality (Conneau et al., 2020): recover strong monolingual performance while leveraging multilingual pretraining via fine-tuning.

The writing-system predictor, together with the compression ratio correlation, points to tokenization as Whisper's main bottleneck. Whisper's BPE vocabulary is trained predominantly on English text (Radford et al., 2023), so Latin and Cyrillic languages share many subword tokens with the existing vocabulary, and fine-tuning need only adjust acoustic-to-token mappings. In contrast, a Bengali sentence requiring 15 tokens with a dedicated tokenizer may require 40+ with Whisper's vocabulary, making learning harder and inference slower. For the roughly 20–30 languages where fine-tuned

WER remains above 30%, tokenizer adaptation or script-aware pretraining may be needed.

**Future work.** In ongoing work, we are extending this analysis to evaluate vocabulary-replacement and related tokenizer-adaptation methods as a direct remedy for poorly tokenized scripts (manuscript in preparation). Preliminary results suggest that swapping Whisper's tokenizer for a language-specific BPE recovers substantial headroom on the highest-fertility scripts, including languages that vanilla fine-tuning leaves above 50% WER.

## References


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

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riber, Ankur Bapna, and 1 others. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of Whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.