Off-Target LID Posterior Mass Predicts Probe Gains in Spoken Language Diarization: A Pre-Registered Validation

Anonymous ACL submission

Abstract

We propose **off-target LID posterior mass** (fraction of posterior probabilities assigned to non-target languages) as a diagnostic measure for predicting downstream Spoken Language Diarization (SLD) probe improvements on pre-trained speech models (e.g., Whisper, MMS). Using a pre-registered evaluation set on unseen language pairs we show that probe gains on SLD are highly predictive from the proposed off-target LID posterior mass measure. Additionally, real-world validity of this diagnostic measure is tested on two held-out multilingual speech corpora, (1) DISPLACE (multilingual South Asian dataset) and (2) MUCS (Hindi–English code-switching), on which observed probe gains fall within the frozen 95% prediction interval (PI). The proposed diagnostic indicates that fine-tuning is most beneficial for utterances whose representations exhibit high posterior mass on non-target languages. Predictive power of off-target posterior mass extends across Whisper, MMS and traditional LID models with encoder specific relationship to downstream task gains.

1 Introduction

Spoken Language Diarization (SLD) is a multilingual speech processing task where the objective is to provide time-stamped monolingual segments from multi-language speech recordings (Mishra and Prasanna, 2024; Kalluri et al., 2024). Robust SLD models are necessary backbones for supporting an array of downstream speech tasks including multilingual automatic speech recognition, speech-to-text translation, and code-switching analysis (Liu et al., 2026). While SLD is analogous to speaker diarization, it has distinct challenges: languages from the same family share acoustic foundations (bilingual speakers code-switching produce phonetically overlapping segments) (Mishra and Prasanna, 2024; Wang et al., 2024), and real-world corpora typically contain large proportions of one or two primary languages with short, sparse minority-language turns (Liu et al., 2025; Frost et al., 2022). Typically, SLD models are trained in a two-stage manner, with extracting language identification (LID) posteriors from a pre-trained speech model such as Whisper (Radford et al., 2023) or MMS (Pratap et al., 2024) (zero-shot or fine-tuned on target corpus), followed by clustering (Kalluri et al., 2024; Lee et al., 2025; Kalda et al., 2024).

In this work, we propose and evaluate *off-target LID posterior mass* (defined in Sec 3) as a diagnostic measure that can act as a predictor of fine-tuning gain. Off-target LID posterior mass is the fraction of a multilingual LID posterior probability that falls outside the target language pair, measured on monolingual audio. Target language pair refers to the languages for which SLD model is trained. We hypothesize that off-target LID posterior mass (computed on pre-trained models without fine-tuning), predicts the Diarization Error Rate (DER) gain that a fine-tuned probe provides on bilingual data from that pair.

We validate this hypothesis across diverse language families through pre-registration. Using synthetic bilingual data from FLEURS (Conneau et al., 2023) we test three distinct pre-trained models (Whisper, MMS and VoxLingua-107 (Valk and Alumäe, 2021)) on eight pairs spanning seven language families: Niger-Congo, Bantu, Indo-Aryan, Austroasiatic, Romance, Japonic, and Austronesian. A linear regression of probe gain on off-target posterior mass gives a slope significantly greater than zero (β=20.5, p=0.003). To measure the predictive power of this regression model on unseen language-pairs, the model coefficients are frozen and eight held-out language families (distinct from training set) are evaluated. Seven of the eight fall inside the frozen 95% prediction interval for DER probe gains, exceeding our pre-registered pass criterion of $\geq 6/8$. Furthermore, we test this phenomenon on two real-world corpora:

DISPLACE (South Asian broadcast), and MUCS (Hindi–English code-switching) which have high and low off-target posterior mass respectively. On both, observed probe gains fall inside the prediction interval of the linear regression model.

Off-target mass can serve as a pre-experiment diagnostic for SLD system design (Figure 1; formal definition in §3). It is computed on monolingual audio, prior to fine-tuning of the segmentation model or clustering, and has the predictive power for whether a probe will deliver substantial DER improvement. This answers three questions that fine-tuning itself cannot: which language pairs need fine-tuning for the current model (and the languages which do not), the encoder which provides better zero-shot performance for a given language pair, and whether an unseen corpus could benefit from fine-tuning for a given model. The contributions of this work are as listed below:

- We identify off-target LID posterior mass as a diagnostic measure that predicts probe gain in SLD models in a training-free manner.
- We validate the predictive power of off-target posterior mass by a pre-registered evaluation set on (1) synthetically generated biligual corpora using FLEURS, (2) real-world multilingual speech datasets DISPLACE and MUCS.
- We attribute the observed predictive power of off-target posterior mass to fine-tuned models ability of redistributing off-target mass onto target language logits.

Our code and the frozen pre-registered regression coefficients are available at an anonymized repository.[1]

2 Related Work

Spoken Language Diarization (SLD). SLD is a multilingual speech processing task that assigns time-stamped language labels to multilingual speech recordings. Current SLD work draws heavily from speaker diarization advances, using cascaded pipelines that extract embeddings (x-vectors or self-supervised) over fixed-duration windows followed by clustering algorithms like agglomerative hierarchical clustering (AHC) (Mishra and Prasanna, 2024; Sell et al., 2018; Park et al., 2020) to provide diarized outputs. Alternatively, change-point methods use attention networks with temporal smoothing to locate code-switching boundaries before segment-level embedding extraction (Mishra et al., 2021). Another line of work poses SLD as supervised frame-level classification (Wang et al., 2024), made possible by the closed set of target languages (unlike the open-set speaker identities in speaker diarization). To better handle code-switching, end-to-end neural diarizers jointly predict segment boundaries and language classes without a separate clustering step (Mishra and Prasanna, 2024). A recent example is SAGE-LD (Lee et al., 2025), which trains an end-to-end diarizer on simulated code-switching corpora (SEAME (Lyu et al., 2010), MUCS (Diwan et al., 2021), MERLiON (Chua et al., 2023)) and reports 23–52% relative DER improvement over prior end-to-end methods (Watanabe et al., 2018; Fujita et al., 2019).

For SLD of Indian languages, the DISPLACE 2024 challenge (Kalluri et al., 2024) is a popular benchmark, with the best-performing model (Kalda et al., 2024) reaching 27.6% DER through LDA/PLDA post-processing of fine-tuned SSL features. While prior work focuses on improving DER directly — whether through new segmentation architectures or new feature pipelines — we instead ask *why* those gains arise, by linking off-target posterior mass to clustering DER on a pre-trained speech model checkpoints.

Representation geometry and fine-tuned probe performance. Speaker diarization has progressed from i-vectors (Sell and Garcia-Romero, 2014) through x-vectors (Dawalatabad et al., 2021) to SSL d-vectors with PLDA backends (Landini et al., 2022), each step aligning embedding geometry more closely with the clustering objective. Whereas in LID models, a different phenomenon is observed: while fine-tuning improves classification accuracy, the resulting embeddings do not necessarily group embeddings from similar languages in the same cluster (Frost et al., 2022), and posteriors from supervised LID models like VoxLingua107 (Valk and Alumäe, 2021) and MMS (Pratap et al., 2024) transfer poorly when utilized for clustering in SLD task.

Prior works in Probing theory (Hewitt and Liang, 2019; Belinkov and Glass, 2019) indicate that high probe accuracy does not guarantee a representation useful for downstream tasks. In this work, we attempt to identify if it is possible to predict this probe gain through inductive biases of the pre-trained posterior, measured on monolingual audio. As shown in the subsequent sections, Off-target

[1] https://anonymous.4open.science/r/emnlp-2026-diarization-submission-5EF8/

posterior mass is able to act as this diagnostic measure.

3 Off-Target LID Posterior Mass

Given a pre-trained LID model and a set of target languages $\mathcal{T}$ (to build an SLD model), off-target mass (as computed in Fig 1) is the fraction of posterior probability the model assigns to non-target languages (in its LID posteriors), when the input utterance belongs to one of the target languages. For an utterance x from a language in $\mathcal{T}$ with N-way posterior p_x:

$$m(x) \;=\; 1 \;-\; \sum_{\ell \in \mathcal{T}} p_x(\ell). \qquad (1)$$

Here, $m(x)$ refers to the aggregated off-target posterior over all utterances in a corpus belonging to the target set of languages.

Utilizing this metric as a diagnostic tool in the context of SLD provides two distinct advantages. Firstly, it can be calculated on monolingual speech utterances without requiring bilingual or code-switched data. Additionally, it does not require fine-tuning of the pre-trained speech models. We use this measure as the predictor in cross-pair regression (§6.1) to forecast the performance gains from a fine-tuned probe on a bilingual corpus.

Hypothesis. Off-target posterior mass for a target language pair (languages of interest for SLD task) predicts the probe gain when fine-tuning a model on that language pair. This hypothesis is motivated from the observation that when the pre-trained model assigns high probability to the two target languages, the probe does not need to reassign the posterior mass, so the observed gain is small. When the model assigns probability across many off-target languages, the probe reassigns this mass onto the target languages, and the observed gain is relatively larger.

4 SLD Pipeline

All systems reported in subsequent sections use a common pipeline: voice activity detection (Silero Team, 2021; Bredin et al., 2020) followed by utterance segmentation into 10 s windows with 0.4 s shift (this follows baseline method proposed in DISPLACE benchmark (Kalluri et al., 2024)). Features were extracted from pre-trained models (§4.1); which were then used in cosine AHC clustering with threshold offset 0.015, and one iteration of VBx resegmentation (Landini et al., 2022) (Fa= 1.5, Fb= 2). Diarization Error Rate (DER) is computed with `dscore` (Ryant et al., 2019) (collar 0.25 s, no forgiveness).

4.1 Feature Extraction

From a pre-trained speech model, we extract two features for use in clustering. Each 10 s segment is encoded, zero-padded to the model's expected input length, and mean-pooled over encoder frames.

LID posterior (`dl`). The model's native language identification posterior, a distribution over the model's full set of supported languages. For Whisper this is LID posterior extracted from its decoder over 99 supported languages. When we fine-tune the encoder in this setting, this LID head is frozen. This ensures that any improvements in DER is attributed to improved model embeddings in the encoder.

Fine-tuned head probabilities (`head`). A linear layer with softmax, trained end-to-end with the encoder on the target language set (p(`en`), p(`hi`), p(`ml`) on DISPLACE; 2-class for FLEURS pairs).

The `dl` and `head` readouts share a checkpoint and differ only at inference. We instantiate this pipeline with three pre-trained models: Whisper `base` (74M parameters), MMS-LID-1024, and VoxLingua107. The FLEURS cross-pair regression (§6.1) fits a 2-class LR probe on each pair's FLEURS train split for all three encoders, with the encoder frozen. The DISPLACE and MUCS analyses (§6.2) use Whisper `base` with a 3-class probe over the corpus target languages. A within-DISPLACE cross-encoder comparison (Appendix F) extends the analysis to MMS-LID and VoxLingua107. Additional details on all model variants are provided in Appendix E.

5 Datasets

5.1 FLEURS

FLEURS (Conneau et al., 2023) is a multilingual read-speech corpus covering 102 languages, with parallel utterances translated from FLoRes-101. We construct synthetic bilingual evaluation data such that it yields off-target posterior mass values across different ranges i.e., low to high. For each language pair, we synthesize ≈50 bilingual test files by concatenating utterances from the two languages' held-out test splits, and train a 2-class Linear Regression (LR) probe on 400 monolingual train-split utterances (200 per language).

We use 16 FLEURS pairs in total: 8 training pairs spanning seven language groupings (Niger-

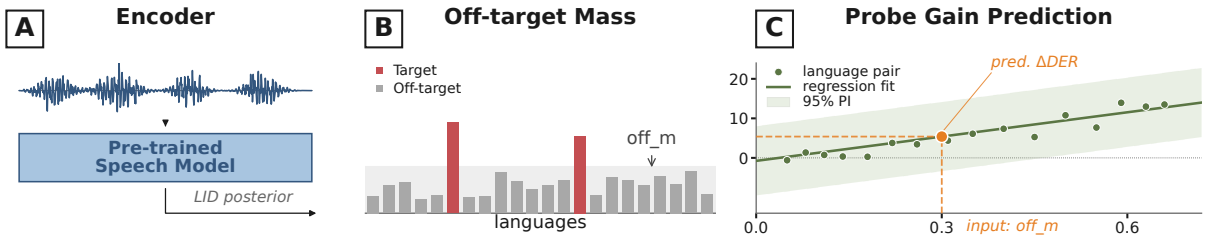

Figure 1: Off-target LID posterior mass as a training-free diagnostic for SLD probe gain. **(A)** A monolingual utterance is passed through a pre-trained speech model, which produces a multilingual LID posterior. **(B)** Target languages (red) carry most of the mass; off_m is the average fraction falling on off-target languages (gray). **(C)** Across FLEURS pairs (green dots), off_m linearly predicts the probe gain Δ in DER points, with the frozen 95 % prediction interval shaded. Orange marker: an input off_m of 0.3 yields a predicted gain $\hat{\Delta}$ of $\sim$5 pt. See §3 for the formal definition and §6.1 for predictive validation.

Congo, Bantu, Indo-Aryan, Austroasiatic, Romance, Japonic, Austronesian) and 8 pre-registered out-of-sample pairs (§6.1). Off-target mass on these pairs ranges from 2% to 71% (Appendix G), giving the regression a wide predictor range that DISPLACE and MUCS alone do not span.

5.2 Displace

The DISPLACE 2024 challenge corpus (Kalluri et al., 2024) is a multilingual conversational speech benchmark released for the Diarization of Speaker and Language in Multilingual Conversational Environments. It contains natural, far-field multi-speaker speech recordings collected in informal conversational settings (group discussions, interviews, and broadcast-style multi-party exchanges), with three target languages reflecting the multilingual context of Indian speech: English, Hindi, and Malayalam. Recordings contain both within-utterance and across-utterance code-switching, with most files featuring two or three speakers.

Fine-tuning Split. 34 broadcast audio files from the official training split (Kalluri et al., 2024), with no overlap with the evaluation set. The training distribution is heavily imbalanced toward English and Hindi: by speech duration, English accounts for $\sim$50%, Hindi $\sim$48%, and Malayalam $\sim$2% of the training data.

Evaluation Split. 32 broadcast audio files ($\sim$17 hours) from DISPLACE 2024 (Kalluri et al., 2024), with no overlap with the training split. File composition includes 3 monolingual, 16 bilingual, 12 trilingual and 1 quadrilingual. English and Hindi form the majority of speech duration whereas Malayalam appears in only 13 of 32 files and constitutes $\sim$12% of total speech by duration.

5.3 MUCS

MUCS (Diwan et al., 2021) is a Hindi–English code-switching corpus drawn from conversational speech. We use the 30-file evaluation set ($\sim$5.18 h) as a held-out cross-corpus test of the predictive law (§6.1). MUCS has a low off-target posterior mass value: pre-trained Whisper assigns only 16% of its posterior mass to non Hindi or English languages, compared to 44% on DISPLACE.

5.4 Statistical Analysis

Linear Regression. We fit ordinary least squares regressions of probe gain Δ (in DER points) on off-target mass, giving the predicted gain $\hat{\Delta}$:

$$\hat{\Delta} = \alpha + \beta \cdot \text{off_m}.$$

We report two-sided p-values on the slope (t-test against β=0), in-sample R^2, and leave-one-out Q^2. Confidence intervals on β come from a residual bootstrap (10K resamples). The 95% prediction interval used in §6.1 is computed from the n=8 training fit, then frozen before any held-out pair is evaluated.

DER comparisons. For DER comparisons on a single corpus, we use a paired bootstrap with files as the unit (10K resamples, 95% CIs, two-sided p-values). When a family of comparisons is pre-planned, we apply Holm correction within family.

Pre-registration. The Whisper FLEURS regression coefficients (α=-0.76, β=20.5), the prediction interval ($\pm$8.74 pt), and the pass rule ($\geq 6/8$ held-out pairs inside PI) were git-committed before any held-out pair was evaluated. The MMS and VoxLingua107 regressions were run post-hoc.

6 Results

We first validate the hypothesis on FLEURS (§6.1), where synthetically generated language pairs let us run a pre-registered out-of-sample test. We then check whether the prediction transfers to two real-world corpora, DISPLACE and MUCS (§6.2).

6.1 Off-Target Mass Predicts Probe Gain on FLEURS dataset

This section tests the hypothesis from §3: off-target posterior mass, measured on monolingual audio, predicts SLD probe gain on bilingual data. The test is pre-registered — the regression coefficients and pass criterion were frozen before any held-out pair was evaluated (details in §5.4).

Using Whisper as the pre-trained speech model,we test on 8 training language pairs from FLEURS (§5.1) spanning seven language families and off-target mass from 2% to 71%: Yoruba–Amharic (Niger-Congo), Swahili–Hausa (Bantu), Bengali–Tamil (Indo-Aryan), Vietnamese–Thai (Austroasiatic), French–German (Romance), Japanese–Korean (Japonic), Spanish–Portuguese (Romance), and Filipino–Indonesian (Austronesian). The downstream clustering pipeline uses cosine AHC. Per-pair DERs are in Appendix G.

Regression fit. A linear regression of probe gain Δ (in DER points) on off-target mass yields the predicted probe gain $\hat{\Delta}$:

$$\hat{\Delta} \;=\; -0.76 \;+\; 20.5 \cdot \text{off_m}$$

with β=20.5, R^2=0.80, p=0.003, and leave-one-out Q^2=0.42. Pairs with off-target mass below 10% (European, East Asian, Southeast Asian) show near-zero probe effect ($|\Delta|$<1.2 pt). Pairs with off-target mass above 45% (African, South Asian) show 8–20 point gains (p<0.001).

As a comparison to off-target posterior mass, we use baseline DER as an alternative diagnostic metric — baseline DER refers to metric measured from clustering on the pre-trained encoder's LID posterior, with no probe applied. The observed fit is similar (β=0.55, R^2=0.87) and the two diagnostic metrics are colinear within FLEURS as recordings are clean concatenations of monolingual speech, so baseline DER mostly reflects how well the encoder separates the target pair — which is exactly what off-target mass measures directly. However, off-target mass has two advantages. Firstly, it does not require running the downstream clustering pipeline, while baseline DER does. Second, on real-world speech corpora (MUCS) baseline DER is unable to provide accurate predictions unlike off-target posterior mass measure (§6.2), where high baseline DER reflects clustering difficulty rather than geometric alignment.

Pre-registered out-of-sample test. A good in-sample fit does not yet show that the regression model has predictive power on unseen language pairs. To test predictive validity, the training-fit coefficients were frozen, the 95% prediction interval (± 8.74 pt around $\hat{\Delta}$), and a pass rule ($\geq 6/8$ held-out pairs inside the interval), all before any held-out pair was evaluated.

We then evaluated eight new pairs spanning the range of off_m $\in [0.015, 0.65]$, with no language shared with the training set: Assamese–Nepali (Indo-Aryan), Persian–Pashto (Iranian), Marathi–Gujarati (Indo-Aryan), Malay–Filipino (Austronesian), Vietnamese–Khmer (Austroasiatic), Dutch–Swedish (Germanic), Arabic–Hebrew (Semitic), and Italian–Romanian (Romance).

Seven of eight pairs (87.5%) fall within the frozen 95 % prediction interval (Figure 3). Refitting on the combined 16 pairs yields β=19.0 (bootstrap 95 % CI $[10.5, 29.2]$, p=0.001), preserving the original slope's sign and confidence interval. Per-pair predictions, residuals, and influence diagnostics are in Appendix D.

Out of the eight unseen language tests pairs, Vietnamese–Khmer prediction of probe gain falls outside the prediction interval. Per-language follow-up shows that Vietnamese has 3% off-target mass while Khmer has 42% — an order of magnitude difference — but the off-target posterior mass computation for bilingual corpus mean averages this to 23% and the regression treats them as a single point. This could indicate that Asymmetric pairs may need a different predictor. One natural direction for future work is to use the per-language off-target mass of each language in a pair as separate variables, rather than collapsing them into a file-mean. Appendix D reports a sensitivity check using the larger of the two per-language values as the predictor.

Generalization across encoders. We fit the same regression on MMS-LID-1024 and VoxLingua107 using a matched 15-pair panel: the combined 16-pair set with Arabic–Hebrew dropped, because MMS-LID-1024 has only a macro Arabic label and no separate Egyptian Arabic neuron, so its

off-target mass for that pair is not comparable to Whisper's. Table 1 reports the regression fit for all three encoders; Figure 2 shows the MMS and VoxLingua regression plots.

Table 1: Cross-encoder regression on the same 15 FLEURS pairs. MMS's n=14 row drops `fil_ph+id_id` (a leverage outlier; Appendix F); VoxLingua's n=14 row drops `ms_my+fil_ph` (highest residual). MMS's n=14 slope falls inside Whisper's CI; VoxLingua's stays outside on the high side.

Encoder	n	β	R^2	Q^2	**95 % CI**
Whisper	15	+18.97	0.53	0.36	[+10.3, +29.2]
MMS-LID	15	+60.74	–	0.06	–
MMS-LID	14	+18.09	0.78	0.72	[+14.8, +25.0]
VoxLingua107	15	+64.95	0.81	0.68	[+39.9, +80.6]
VoxLingua107	14	+63.21	0.85	0.66	[+36.9, +79.6]

the same positive monotonic relationship for all three encoders, with encoder-dependent slope. The relationship is positive for all three encoders, with encoder-dependent slope. MMS's slope lands inside Whisper's CI only after dropping a single high-leverage pair; VoxLingua's is roughly 3× steeper. Both alternative encoders produce narrower off-target-mass ranges than Whisper on these pairs; absolute ranges and gain bounds are included in Appendix F.

Takeaway. On Whisper, off-target posterior mass predicts probe gain to within a calibrated interval, and the pre-registered out-of-sample test passes. MMS and VoxLingua107 show the same direction: higher off-target mass indicates higher probe gain. The slope differs by encoder though: MMS lands inside Whisper's CI only after dropping one outlier pair, and VoxLingua's slope is about 3× steeper. So the direction generalizes, but the prediction interval is encoder-specific.

6.2 Transfer to Real-World Corpora: DISPLACE and MUCS

We test whether the FLEURS-fit regression transfers to two real-world corpora outside the training set: DISPLACE (multilingual South Asian dataset, high off-target posterior mass) and MUCS (Hindi–English code-switching dataset, low off-target posterior mass). Linear probe gain predictions for both datasets fall within the pre-registered 95 % prediction interval (Figure 3, purple and orange diamonds).

DISPLACE has a high off-target-mass range. The pre-trained Whisper model assigns only 60% of its posterior on the three target languages (English, Hindi, Malayalam). The other 40% is assigned to non-target languages (Table 2), and majority of this non-target mass (26%) is assigned to acoustically related South Asian languages: Telugu, Tamil, Bengali, Urdu, Marathi, Kannada, and Nepali. This pattern is observed across different parameter scales of Whisper models (Appendix C), and prior work on LID for Indian languages reports the same pattern (Kommineni et al., 2025).

Table 2: Posterior mass decomposition and DER on DISPLACE eval (32 files, 47,828 subsegments). `baseline_dl` is pre-trained Whisper's 99-dim posterior; `displace_head` is the 3-class softmax probe. Task = en+hi+ml; S.Asian = te+ta+bn+ur+mr+kn+ne.

System	**Dim**	**Task**	**S.Asian**	**Other**	**DER**
baseline_dl	99	60.3%	26.0%	13.7%	35.48
displace_head	3	100%	0%	0%	**33.41**

DISPLACE has 3 target languages (en, hi, ml), so we compute off-target mass as $1 - P(\text{en+hi+ml})$, averaged over the 32 eval files. This gives off_m=0.44. The frozen FLEURS regression predicts a probe gain of +8.3 pt. We observe +2.07 pt (Table 2, `displace_head` vs. `baseline_dl`), inside the 95 % PI (Figure 3, purple diamond).

The observed gain is driven by supervised training rather than dimension restriction. Restricting the pre-trained 99-dim posterior to the three target dimensions without retraining degrades DER, whereas a logistic-regression probe trained on top of the 99-dim posterior with DISPLACE labels recovers most of the head's gain (Appendix B).

MUCS has a low off-target posterior mass measure (0.16). Hindi and English are well represented in Whisper's LID posterior, so MUCS's off-target mass is only 16% — compared to 44% on DISPLACE and 70% on the high-mass FLEURS pairs (Yoruba–Amharic, Swahili–Hausa). The frozen regression predicts +2.5 pt; we observe +0.14 pt (p=0.251, n=30 files), inside the PI (Figure 3, orange diamond).

Results on MUCS dataset reflect the impact of different diagnostic metrics (baseline DER vs off-target posterior mass). The alternative regression using baseline DER as the diagnostic metric (§6.1) yields a predicted gain of +21.4 pt, far above the observed +0.14 pt: MUCS's high baseline DER comes from clustering densely code-

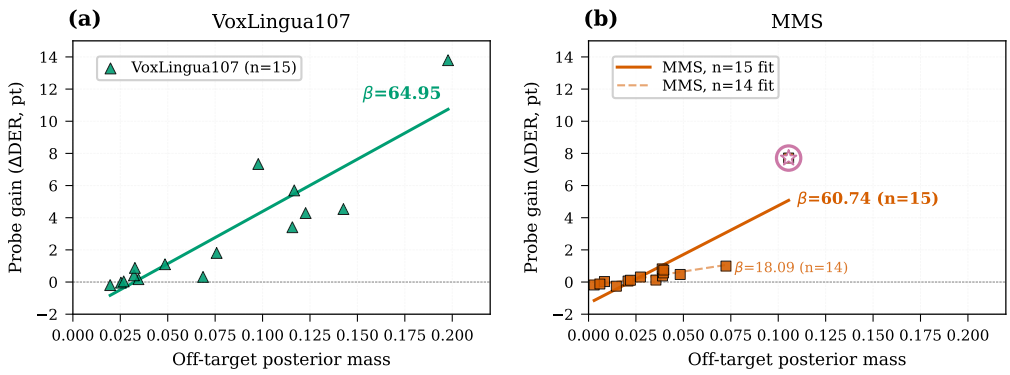

Figure 2: Cross-encoder slope comparison on the same 15 FLEURS pairs. **(a)** VoxLingua107: probe gain rises monotonically with off-target mass at β=64.95. **(b)** MMS-LID: the n=15 fit (solid, β=60.74) is dominated by a single high-leverage pair (Filipino–Indonesian, circled); the n=14 refit with that pair dropped (dashed) yields β=18.09, inside Whisper's CI. The same positive monotone relationship holds for all three encoders; quantitative slopes and drop-one diagnostics in Table 1 and Appendix F.

switched speech, not from poor alignment between Whisper's posterior and the Hindi–English target pair. Off-target mass tracks the second source only and is therefore the diagnostic metric that transfers from FLEURS to MUCS.

Takeaway. Predictions from the frozen regression coefficients fall within the pre-registered prediction interval on both held-out FLEURS pairs and on two real-world corpora (DISPLACE, MUCS) that were not used to fit the regression.

DISPLACE (off_m=0.44, predicted +8.3, observed +2.07) tests transfer to a high fraction of off-target-mass dataset. On this real world dataset, the deviation between the predicted and observed gains is large; we acknowledge this limitation and note that it is consistent with the substantial residual variance (Q^2=0.38) of the regression. This could be owing to the fact that the current model takes into account only the proportion of off-target posterior mass whereas SLD probe gains would have impact from other sources such as recording conditions, language data imbalance (such as extremely low Malayalam data in Displace train set).

On the other hand, MUCS (off_m=0.16, predicted +2.5, observed +0.14) tests the low off-target posterior mass regime. The probe yields near-zero gain on MUCS due to heavy Hindi–English code-switching, but this observation still falls within the frozen prediction interval. Together, the two real-world tests span the low- and high-mass regimes, supporting off-target posterior mass as a reliable predictor of SLD probe gains across both.

7 Discussion

Off-target mass is not an artifact of model capacity. To demonstrate that the observed predictability of off-target mass is not solely attributed to the parameter count of the model, we provide an ablation with whisper variants of different sizes, refer to Appendix C for detailed results. When varying the parameter count from (74M → 1.5B), using variants (`base` 74M, `small` 244M, `large-v3` 1.5B), the share of posterior mass that Whisper posteriors assign to the three target languages increases slightly from 60.3% to 67.8% (averaged over all 47,828 DISPLACE subsegments; per-file averages in Table 3). This indicates that Off-target posterior mass CIs are significantly greater than zero across every model scale (Appendix A). As the model size increases, posterior mass is concentrated around a smaller set of languages, but not necessarily assigned to the target languages: for `Whisper large-v3`, Kannada (non-target Dravidian) carries more posterior mass than Hindi (target). The probe's advantage over the posterior LID is correspondingly stable across scales (+3.1 pt at `small`, +4.4 pt at `large-v3`). Off-target mass is therefore a structural property of how a multilingual LID head assigns probability across a closed language set, not a consequence of limited capacity — which is why it can be measured on a pre-trained encoder and still predict downstream gain.

Diagnostic measure as SLD probe-gain magnitude indicator. The diagnostic metric partitions corpora into three operating regimes (Figure 3):

- off_m $<$ 0.10: near-zero probe gain expected.

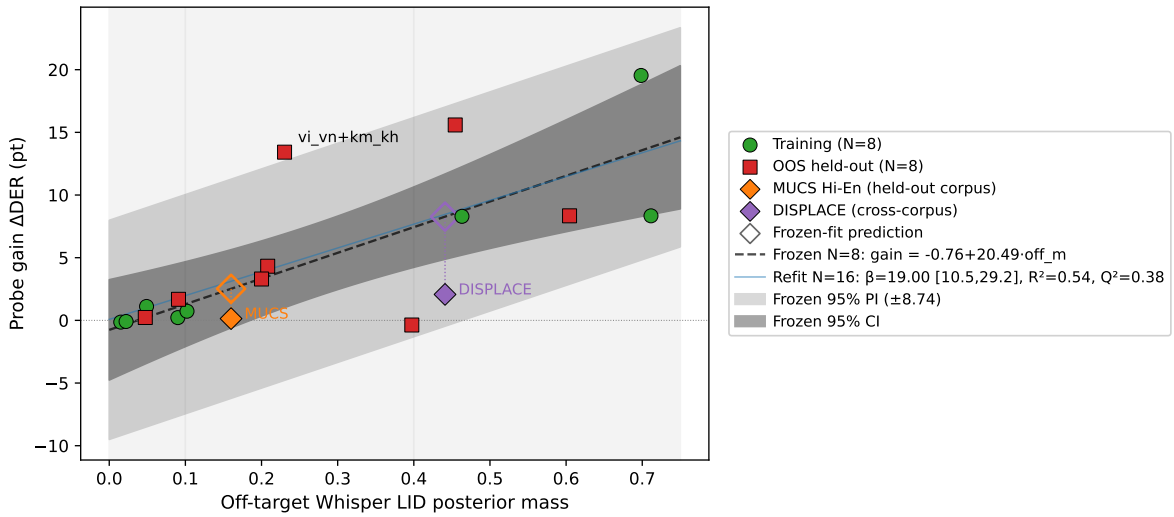

Figure 3: Combined n=16 FLEURS regression: probe gain vs. off-target mass for the 8 training pairs (green circles) and the 8 pre-registered out-of-sample pairs (red squares). Shaded band is the frozen n=8 95 % prediction interval (± 8.74 pt around $\hat{\Delta} = -0.76 + 20.5$ off_m); 7/8 OOS pairs fall within the band. The single miss (`vi_vn+km_kh`) is an asymmetric pair: per-language follow-up gives off_m$_{\text{vi}}$=0.032 vs. off_m$_{\text{km}}$=0.421. The orange diamond marks the held-out MUCS corpus (off_m=0.16; filled = observed $+0.14$ pt, hollow = predicted $+2.5$ pt). The purple diamond marks DISPLACE (off_m=0.44; filled = observed $+2.07$ pt, hollow = predicted $+8.3$ pt). Both held-out corpora fall inside the PI. For DISPLACE, off-target mass is computed over its 3 target languages (en, hi, ml) rather than 2.

- $0.10 \leq$ off_m < 0.40: moderate probe gain expected.
- off_m ≥ 0.40: large probe gain expected.

MUCS and the European FLEURS pairs fall in the first regime as Whisper assigns larger proportions of posterior to the target languages. This could be owing to Whispers' pre-training data consisting larger proportions of European origin languages. DISPLACE and the African FLEURS pairs fall in the third. Across different encoder models, MMS-LID and VoxLingua107 exhibit similar monotonic relationship for probe gain prediction, but their operating ranges vary compared to Whisper (off_m < 0.11 for MMS, < 0.20 for VoxLingua) lying entirely within Whisper's low-to-moderate regime. Two questions of practical interest for SLD system design are therefore answered by off-target mass, prior to fine-tuning: (1) which encoder produces the largest off-target mass on the target language pair, and (2) the corresponding SLD probe gain that can be expected from fine-tuning on the corpus.

8 Conclusion

We introduced off-target LID posterior mass as a training-free diagnostic measure for predicting probe gains in Spoken Language Diarization (SLD). The proposed measure is the fraction of a multilingual LID posterior that a pre-trained speech model places outside the target languages when the input is monolingual audio from one of the target languages. We hypothesized that off-target posterior mass, computed without fine-tuning or clustering, predicts the diarization error rate (DER) reduction a fine-tuned probe will deliver on bilingual data from the same pair.

We validated this through pre-registration on FLEURS: a linear regression fit on training pairs was frozen along with its prediction interval, and the majority of held-out language pairs fell inside the pre-registered interval, exceeding our pass criterion. The frozen prediction also transferred to two held-out real-world corpora, DISPLACE and MUCS. The direction of the relationship generalizes across Whisper, MMS-LID and VoxLingua107, though the slope is encoder-specific. The practical implication is that off-target posterior mass lets a practitioner answer, before any fine-tuning, three questions that fine-tuning alone cannot: identifying the language pairs that need fine-tuning for a given model, which encoder is best suited to a given target language pair, and the expected DER reduction through fine-tuning on a new corpus.

Limitations

While the identified diagnostic metric off-target posterior mass is able to predict the probe gain that can be obtained for SLD across different model architectures, the observed regression parameters show variance across model families indicating that the magnitude of gain observed is model specific with limited cross model generalizability. Predictive experiments on FLEURS datasets are performed on synthetically generated bilingual files, which may not accurately reflect spontaneous code-switching in real-world scenarios, which may impact the observed predictive power of off-target posterior mass. This work provides validation for off-target posterior mass on a specific architecture of SLD models that have a two-stage pipeline (as this architecture is widely used) with a concatenation of segmentation model and clustering. Future works should consider analyzing more recent architectures such as E2E models. While off-target posterior mass provides predictive power, probe gains for SLD are impacted by other factors such as SNR of recording conditions, demographics of speakers and near-/far-field recordings, which we do not take into account in this work. These factors in conjunction with off-target posterior mass should be investigated further in subsequent works.

References

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: Neural building blocks for speaker diarization. In *Proc. ICASSP 2020*, pages 7124–7128.

Victoria Yi Han Chua, Hexin Liu, Leibny Paola Garcia Perera, Fei Ting Woon, Jinyi Wong, Xiangyu Zhang, Sanjeev Khudanpur, Andy W. H. Khong, Justin Dauwels, and Suzy J. Styles. 2023. MERLIon CCS challenge: An english-mandarin code-switching child-directed speech corpus for language identification and diarization. In *Proc. Interspeech 2023*, pages 5133–5137.

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

Nauman Dawalatabad, Mirco Ravanelli, François Grondin, Jenthe Thienpondt, Brecht Desplanques, and Hwidong Na. 2021. Ecapa-tdnn embeddings for speaker diarization. In *Proc. Interspeech 2021*, pages 3560–3564.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, and Karthik Sankaranarayanan. 2021. MUCS 2021: Multilingual and code-switching ASR challenges for low resource Indian languages. In *Proc. Interspeech 2021*, pages 2446–2450.

Geoffrey Frost, Emily Morris, Joshua Jansen van Vüren, and Thomas Niesler. 2022. Fine-tuned self-supervised speech representations for language diarization in multilingual code-switched speech. In *Artificial Intelligence Research (SACAIR 2022)*, volume 1734 of *Communications in Computer and Information Science*, pages 246–259. Springer.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019. End-to-end neural speaker diarization with self-attention. In *Proc. ASRU 2019*, pages 296–303.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proc. EMNLP*, pages 4238–4248.

Joonas Kalda, Tanel Alumäe, Martin Lebourdais, Hervé Bredin, Séverin Baroudi, and Ricard Marxer. 2024. TalTech-IRIT-LIS speaker and language diarization systems for DISPLACE 2024. In *Proc. Interspeech 2024*, pages 1635–1639.

Shareef Babu Kalluri, Prachi Singh, Pratik Roy Chowdhuri, Apoorva Kulkarni, Shikha Baghel, Pradyoth Hegde, Swapnil Sontakke, Deepak K T, S. R. Mahadeva Prasanna, Deepu Vijayasenan, and Sriram Ganapathy. 2024. The second DISPLACE challenge: DIarization of SPeaker and LAnguage in conversational environments. In *Proc. Interspeech 2024*, pages 1630–1634.

Aditya Kommineni, Rajat Hebbar, Sarah Petrosyan, Pranali Khobragade, Sudarsana Kadiri, Miguel Arce Rentería, Jinkook Lee, and Shrikanth Narayanan. 2025. Can speech foundation models effectively identify languages in low-resource multilingual aging populations? *JASA Express Letters*, 5(9).

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71:101254.

Sangmin Lee, Woongjib Choi, Jihyun Kim, and Hong-Goo Kang. 2025. SAGE-LD: Towards scalable and generalizable end-to-end language diarization via simulated data augmentation. *arXiv:2510.00582*.

Hexin Liu, Haoyang Zhang, Qiquan Zhang, Xiangyu Zhang, Dongyuan Shi, Eng Siong Chng, and Haizhou Li. 2026. Code-switching speech recognition under the lens: Model- and data-centric perspectives. *IEEE Transactions on Audio, Speech and Language Processing*, 34:1853–1865.

Hexin Liu, Xiangyu Zhang, Haoyang Zhang, Leibny Paola Garcia-Perera, Andy W. H. Khong, Eng Siong Chng, and Shinji Watanabe. 2025. Aligning speech to languages to enhance code-switching speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4712–4725.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2010. SEAME: A mandarin-english code-switching speech corpus in south-east asia. In *Proc. Interspeech 2010*, pages 1986–1989.

Jagabandhu Mishra, Abhinav Agarwal, and S. R. Mahadeva Prasanna. 2021. Spoken language diarization using an attention based neural network. In *National Conference on Communications (NCC)*, pages 1–6.

Jagabandhu Mishra and S. R. Mahadeva Prasanna. 2024. Implicit self-supervised language representation for spoken language diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3393–3407.

Tae Jin Park, Kyu J. Han, Manoj Kumar, and Shrikanth Narayanan. 2020. Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. *IEEE Signal Processing Letters*, 27:381–385.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(1):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *Proc. ICML*.

Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. 2019. The second DIHARD diarization challenge: Dataset, task, and baselines. In *Proc. Interspeech 2019*, pages 978–982.

Gregory Sell and Daniel Garcia-Romero. 2014. Speaker diarization with plda i-vector scoring and unsupervised calibration. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 413–417. IEEE.

Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, and Sanjeev Khudanpur. 2018. Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge. In *Proc. Interspeech 2018*, pages 2808–2812.

Silero Team. 2021. Silero VAD: pre-trained enterprise-grade voice activity detector.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a dataset for spoken language recognition. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658.

Yujia Wang, Hexin Liu, and Leibny Paola Garcia. 2024. Bridging child-centered speech language identification and language diarization via phonetics. In *Proc. Interspeech 2024*, pages 5148–5152.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-End Speech Processing Toolkit. In *Proc. Interspeech 2018*, pages 2207–2211.

A Off-Target Posterior Mass Ablations for Whisper Variants

Table 3 (§2) reports the bootstrap 95% CIs for posterior mass at Whisper base, small, and large-v3. The off-target CIs are significantly greater than zero for all whisper variants, indicating that the off-target mass is not an artefact of small models. At per-language granularity, larger models concentrates posterior mass on a smaller set of languages, but not entirely on the target language of interest The effective number of languages occupying the posterior LID reduces as model size increases (EffN : $6.69 \rightarrow 5.10 \rightarrow 3.45$) (Table 3)

B Projection Baselines

Two projection alternatives test whether the softmax head's advantage over the pre-trained LID posterior arises from the projection or from supervised training.

Direct dimension extraction (`direct_3dim`). We extract the three task-relevant dimensions (`en`, `hi`, `ml`; indices 18, 31, 55 in Whisper's sorted language list) directly from the pre-trained 99-dim posterior and renormalize them onto the 3-simplex. The result is *worse* than the 99-dim baseline: 37.71% DER, 2.23 pt above `baseline_dl`. The failure mode is class-proportion misranking: renormalizing inflates Malayalam from 25% to ≈42% of the posterior mass and shifts Hindi from 12% to ≈20%. The duration-weighted eval-set composition is ∼50% English, ∼38% Hindi, ∼12% Malayalam, so renormalization makes Malayalam the

largest predicted class even though it is the smallest in reality. Restricting to three dimensions without supervised training does not suffice.

Logistic-regression probe (`proper_lr`). LR probe trained on the 99-dim Whisper posterior with DISPLACE official train labels reaches 34.71% DER, within 1.3 pt of `displace_head` (p=0.125). A 3-class softmax mapping on top of a frozen pre-trained encoder recovers most of the head's gain; the residual reflects the benefit of joint encoder-head training rather than an architectural difference.

C Whisper Scale Generalization

This appendix provides per-scale Whisper results supporting the "off-target mass is not an artifact of model capacity" argument in §7. We compare the softmax head against the 99-dim posterior backend on the DISPLACE eval at three Whisper scales: `base` (74M parameters), `small` (244M), and `large-v3` (1.5B).

Head advantage across scales. The head retains its advantage over the 99-dim posterior at every scale, with similar magnitude: $\Delta\approx3.1$ pt at `small` and $\Delta\approx4.4$ pt at `large-v3`. The `small` and `large-v3` runs are single-seed; we report direction and magnitude rather than significance, and flag multi-seed validation at these scales as future work.

Larger whisper variants do not produce improvements in absolute DER. Increasing encoder model size does *not* reduce DER of the fine-tuned models: `small_head` 33.07%, `base_head` 33.41%, `large-v3_head` 32.95% — all within 0.6 pt and statistically indistinguishable (`small` vs. `base` p=0.272; `large-v3` vs. `base` p=0.222, Δ=$+0.50$ pt).

Larger whisper variants do not produce noticeably larger on-target language posteriors. Posterior mass on the three target languages rises modestly with scale ($0.603\rightarrow0.632\rightarrow0.678$ from `base` to `small` to `large-v3`; Table 3). While increase in model size results in increased posterior mass over the target languages, this does not translate to reduction in baseline DER. However, fine-tuning is able to provide improvements across the encoder model scale.

D Pre-Registered OOS Validation Detail

We elaborate the pre-registered out-of-sample validation summarised in §6.1 with per-pair predictions, the combined n=16 refit, influence diagnostics, and the rationale for single OOS miss.

Table 3: Whisper `detect_language` posterior mass on the DISPLACE official eval (32 files), with bootstrap 95% CIs across files (10K resamples). Task = $P(\mathrm{en})$ + $P(\mathrm{hi})$ + $P(\mathrm{ml})$; SA-near = $P(\mathrm{bn}, \mathrm{ta}, \mathrm{te}, \mathrm{ur}, \mathrm{mr}, \mathrm{kn}, \mathrm{ne})$. Off-target CIs exclude zero at every scale. Whisper scales: base/small/large-v3 correspond to 74M/244M/1.5B parameters.

Whisper scale	Task mass	SA-near	Other	EffN
base	0.559 [0.50, 0.62]	0.272 [0.22, 0.32]	0.169 [0.14, 0.20]	6.69
small	0.565 [0.49, 0.64]	0.306 [0.24, 0.37]	0.128 [0.10, 0.15]	5.10
large-v3	0.604 [0.52, 0.69]	0.287 [0.21, 0.37]	0.108 [0.08, 0.14]	3.45

Table 4: Pre-registered out-of-sample test on eight held-out FLEURS pairs. Predicted Δ from frozen n=8 coefficients (α=-0.76, β=20.5); 95 % PI half-width ±8.74 pt. Hit = |residual| within frozen PI. 7/8 within PI exceeds the pre-registered $\geq 6/8$ threshold.

Pair	**off_m**	**Pred. Δ**	**Obs. Δ**	**Resid.**	**Hit**
as_in+ne_np	0.605	+11.63	+8.34	−3.29	✓
fa_ir+ps_af	0.454	+8.55	+15.59	+7.04	✓
mr_in+gu_in	0.397	+7.38	−0.37	−7.75	✓
ms_my+fil_ph	0.208	+3.50	+4.32	+0.82	✓
vi_vn+km_kh	**0.230**	+3.96	+13.42	+9.46	×
nl_nl+sv_se	0.200	+3.34	+3.30	−0.04	✓
ar_eg+he_il	0.091	+1.10	+1.68	+0.58	✓
it_it+ro_ro	0.047	+0.21	+0.23	+0.02	✓

Combined n=16 refit. Refitting on the combined n=16 pairs yields β=19.0 (bootstrap 95 % CI [10.5, 29.2], p=0.001, in-sample R^2=0.54, Q^2=0.38); the sign and magnitude of β are unchanged relative to the frozen n=8 fit, and the original confidence interval contains the refit slope.

Influence diagnostics. No pair has Cook's distance >1 in the combined fit (maximum D=0.47, `sw_ke+ha_ng`), confirming that the high-mass anchors that dominated the training fit (D=1.7 at n=8) lose their leverage as the off-target-mass range is filled in.

The OOS miss (vi_vn + km_kh) arises from asymmetry in off-target language mass in selected language pair. The single OOS pair that fell outside the frozen 95 % prediction interval (`vi_vn+km_kh`) is the structural refinement that pre-registration surfaced: an in-sample fit alone could not have flagged it. For `vi_vn+km_kh` the residual is $+9.5$ pt at off_m=0.23 (pre-

dicted $+4.0$ pt, observed $+13.4$ pt); a per-language follow-up shows file-mean off-target mass of 0.032 ± 0.020 on Vietnamese-only segments versus 0.421 ± 0.065 on Khmer-only segments—an order-of-magnitude asymmetry the bilingual file-mean of 0.230 averages away. Substituting $\max(\text{off_m}_{\text{vi}}, \text{off_m}_{\text{km}})$=0.421 into the frozen regression raises the prediction from $+4.0$ to $+7.9$ pt, closing roughly half of the residual; the remaining $+5.6$ pt is consistent with asymmetric pairs interacting non-additively, the under-resourced side drawing posterior mass toward the well-resourced target rather than dispersing across all 99 dimensions. The refinement the pre-registered miss identifies—moving from a bilingual file-mean to a per-language or per-segment-conditional statistic—is a prediction-form revision that an in-sample fit alone could not have motivated; we leave its evaluation to future work.

$\max$(**off_m**) **sensitivity check.** As a post-hoc sensitivity check, we evaluated $\max(\text{off_m}_{\text{lang1}}, \text{off_m}_{\text{lang2}})$ as an alternative predictor across all 16 pairs. It improves in-sample fit (n=8 R^2=0.90 vs. 0.80; LOO Q^2=0.73 vs. 0.42) but does not improve out-of-sample hit rate ($6/8$ vs. $7/8$ within each predictor's own 95 % PI). The lower in-sample residual SD under $\max$(off_m) reflects fitting to the same training data rather than improved generalization. Per-segment-conditional or per-language regressors remain the more promising direction for asymmetric pairs.

E Description of System Variants

All systems evaluated in this paper are summarized in Table 5.

F Cross-Encoder Validation Detail

Within-DISPLACE probe construction. We applied the same 3-class probe construction used for Whisper to two dedicated LID systems (VoxLingua107, MMS-LID-1024) on the DISPLACE official eval. Both are significantly worse than Whisper's baseline (all p<0.05); the probe construction yields no comparable absolute gain (MMS: Δ=+0.27 pt, p=0.967). The probe construction depends on the encoder representation; weaker representations are not rescued by supervised probability fitting alone.

Cross-pair regression on MMS and VoxLingua107. We additionally fit the cross-pair off-target-mass→probe-gain regression (§6.1) on MMS and VoxLingua107 over the same matched 15-pair FLEURS panel: the combined n=16 set with Arabic–Hebrew dropped, because MMS-LID-1024 collapses all Arabic varieties (Egyptian, MSA, Maghrebi, Levantine) into a single macro `ara` neuron and therefore has no comparable off-target mass for the Egyptian Arabic side of that pair. For each pair we extracted the encoder's native multilingual LID softmax over the canonical 50 sorted FLEURS test-split utterances per language (matching the Whisper protocol), computed mean off-target softmax mass, and ran the same 2-class probe vs. DL clustering comparison used for Whisper. Table 1 in §6.1 reports the resulting fits.

MMS leverage diagnostic. At n=15, the MMS slope (β=+60.7) is roughly three times Whisper's, p<0.001, with Spearman ρ=+0.93 (p<10^{-6}). However, leave-one-out cross-validation gives Q^2=0.06, indicating the fit does not generalize: the pair `fil_ph+id_id` is the lone MMS pair whose off-target mass exceeds 0.10 (0.1055) and the lone pair whose probe gain exceeds 1 pt (+7.71 pt), with the next-largest MMS gain being `fr_fr+de_de` at +0.99 pt. Removing this single point yields β_{MMS}=+18.09 (R^2=0.78, Q^2=0.72, p<0.001, bootstrap CI [+14.8, +25.0]), inside Whisper's n=15 bootstrap CI [+10.3, +29.2].

VoxLingua diagnostic. Unlike MMS, VoxLingua's slope is stable to drop-one: β=+64.95 at n=15 (R^2=0.81, Q^2=0.68, p<0.001, bootstrap CI [+39.9, +80.6], Spearman ρ=+0.91, p<10^{-5}); the highest-residual pair is `ms_my+fil_ph` (off_m=0.098, gain +7.34 pt), and dropping it

Table 5: Systems evaluated on the official evaluation set (32 files). All use Whisper `base` unless noted. For each fine-tuned base, `*_dl` and `*_head` rows share the same checkpoint and differ only in inference-time readout (see §4.1).

Label	Features	Training
baseline_dl	99-dim post.	None
displace_dl	99-dim post.	DISPLACE (34)
displace_head	3-dim softmax	DISPLACE (34)
frozen_baseline_head	3-dim softmax	Head only
displace_head_logits	3-dim logits	DISPLACE (34)
direct_3dim	3-dim renorm.	None
proper_lr	3-dim LR proj.	LR on DISP train

yields β=+63.21 (R^2=0.85, Q^2=0.66, bootstrap CI $[+36.9, +79.6]$). VoxLingua's coupling between off-target mass and probe gain is therefore genuinely steeper than Whisper's, not a single-point artefact.

Operating-range contrast. The three encoders produce very different off-target-mass ranges on these pairs. Whisper's n=15 FLEURS pairs span off_m $\in [0.015, 0.711]$ (mean 0.279, median 0.198) and probe gains $\in [-0.37, +19.54]$ pt; MMS's pairs span $[0.003, 0.106]$ (mean 0.034, median 0.036) and probe gains $\in [-0.26, +7.71]$ pt; VoxLingua's pairs span $[0.020, 0.198]$ (mean 0.069, median 0.069) and probe gains $\in [-0.19, +13.80]$ pt. Both alternative encoders concentrate posterior mass on their target languages much more tightly than Whisper does, leaving little headroom for a probe to recover.

The within-DISPLACE null and the regression-level slope agreement are consistent under one mechanism: The relationship is positive for all three encoders, with encoder-dependent slope. MMS's slope lands inside Whisper's CI only after dropping a single high-leverage pair; VoxLingua's is roughly 3× steeper.

Pre-registration scope. The Whisper cross-pair regression was pre-registered (commit hashes omitted for anonymity; will be restored in the camera-ready version). The MMS and VoxLingua regressions reported above were conducted post-hoc, after the pre-registered Whisper OOS validation completed; they test cross-encoder generalization rather than pre-registered transfer.

G FLEURS In-Domain Regression Detail

We elaborate the in-domain FLEURS regression summarised in §6.1 with per-pair DERs and the auxiliary regression statistics omitted from the main text.

The `head_logits` readout. The ablation uses a third readout from the fine-tuned 3-class head: `head_logits`, the head's three output values *before* the softmax. Comparing `head` and `head_logits` isolates the softmax step from the dimensionality change. The `head` readout (softmax outputs) and `head_logits` share the same trained weights and differ only at inference.

The estimated slope of β=20.5 implies that a 0.1 increase in off-target mass is associated with a 2.05-point larger probe advantage on average; substantial residual variance remains (s=3.4 pt).

Table 6: FLEURS in-domain generalization across eight language pairs spanning seven language groupings: Niger-Congo (Yoruba), Bantu (Swahili), Indo-Aryan (Bengali), Austroasiatic (Vietnamese), Romance (French/Spanish/Portuguese), Japonic (Japanese), and Austronesian (Filipino/Indonesian). Probe trained on each pair's own train split. DL = pre-trained Whisper 99-dim posteriors; Probe = 2-class LR. $\Delta = \mathrm{DL} - \mathrm{Probe}$ (positive = probe wins). Off-tgt = fraction of Whisper posterior mass outside target language dims. Baseline DL DER is reported as a reference quantity; the predictor used for the regression in this section is off-target posterior mass.

Pair	DL (%)	Probe (%)	Off-tgt	Δ (pt)
yo_ng + am_et	32.51	24.17	71%	$+8.3^{***}$
sw_ke + ha_ng	37.91	18.37	70%	$+19.5^{***}$
bn_in + ta_in	19.01	10.71	46%	$+8.3^{***}$
vi_vn + th_th	11.49	11.27	9%	+0.2
fr_fr + de_de	11.08	9.97	5%	$+1.1^{\dagger}$
ja_jp + ko_kr	8.74	8.83	2%	−0.1
es_419 + pt_br	8.61	8.75	2%	−0.1
fil_ph + id_id	7.27	6.54	10%	+0.7
Pooled (n=373)	17.08	12.33		$+4.8^{***}$

$^{***}p<0.001$; $^{\dagger}p=0.001$; paired bootstrap.

The bootstrap 95 % confidence interval for β is $[11.6,\ 29.8]$, and the rank correlation between off-target mass and probe gain is Spearman ρ=0.86 (p=0.011).

Use of AI Assistants

We used large language models (LLMs) for two purposes: (i) language polishing and paraphrasing of author-written content, and (ii) generating code for experiments and table formatting. No LLM was used to generate research ideas or scientific content. All LLM-produced text and code was reviewed and verified by the authors, who take full responsibility for the manuscript.