QUALITY BEFORE QUANTITY: WHAT GOVERNS THE GAIN FROM SELF-TRAINING IN LOW-RESOURCE SPEECH RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-training with pseudo-labels is a commonly used strategy for domain adaptation, yet practitioners have no evidence with which to predict how much benefit it will provide, or how aggressively to filter, before running it. We study confidence-filtered self-training on speaker-disjoint and fully cross-corpus target splits, using an external test corpus that is never pseudo-labelled, oracle ceilings obtained from true target labels, and ten seeds per condition with paired non-parametric tests. We perform the full threshold sweep twice, using two different pools of unlabelled spontaneous Hindi 6.5 hours sampled speaker-disjointly from the target corpus, and 30 hours from a different corpus altogether and then perform a second self-training round on top of them. For both pools, self-training improves word error rate by 6.5 points absolute (15% relative), improves an external benchmark it never saw by 2.3 and 3.1 points, does not harm source-domain accuracy, and yields a student that outperforms its own teacher, trained on 128 hours of labelled speech, on every seed. Sweeping the threshold over the two pools reveals the same structure in both: once retained volume is controlled, downstream error is perfectly rank-ordered by the error rate of the retained pseudo-labels, and the optimum is interior, falling at the same threshold in both pools even though one is five times the size of the other and the teachers differ by fifteen points in quality. The relationship is strongly concave: almost all of the benefit comes from removing the worst labels, and filtering severely enough to retain only a small fraction of the data reverses the benefit even though those are the cleanest labels in the study. The second self-training round saturates: the improved teacher creates labels 2.8 points cleaner overall but only 0.4 points cleaner at the operating threshold, because iteration repairs precisely the labels that filtering had already discarded. We argue that filtering and iteration are substitutes rather than complements, that the pre-registered slope transfers in ordering but not in magnitude across corpora, and that quality drives the result, with diminishing returns, until volume becomes the limiting factor.

1 INTRODUCTION

The enormous scale of multilingual pre-training, together with parameter-efficient adaptation methods such as LoRA (Hu et al., 2022), has made speech foundation models such as Whisper (Radford et al., 2023) the default starting point for new languages and domains. In low-resource settings the bottleneck is not compute but transcription: labelled audio is expensive, and even when it is available it is often scripted read speech, which is cheap to elicit and easy to transcribe. The speech these systems must ultimately recognise is spontaneous. This mismatch is significant and well reported: models achieving single-digit or low double-digit error rates on in-corpus read test sets degrade by a factor of two or more on spontaneous, accented or conversational audio (Kumar et al., 2026; Juvekar et al., 2026).

When no label exists for the target audio, the standard solution is self-training with pseudo-labels (Kahn et al., 2020; Xie et al., 2020; Park et al., 2020; Likhomanenko et al., 2021). The recipe is simple: a teacher trained on labelled source data transcribes unlabelled target audio, the pseudo-labels

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

are filtered, and a student is trained on the union. What is not straightforward is determining what to filter. Confidence thresholding is a compromise between quality and quantity (Sohn et al., 2020; Chen et al., 2023): the stricter the threshold, the more likely a retained label is to be accurate, but the fewer labels remain. In practice the threshold is determined by intuition or by tuning on the target test set, and the field lacks a quantitative account of what that choice buys.

This paper poses a more limited, answerable question: *given a measurement of pseudo-label quality and quantity, what determines the downstream gain?* We answer it with a controlled study on Hindi, adapting Whisper-small from RESPIN-S1.0 (Kumar et al., 2026) scripted, dialect-balanced read speech to Lahaja (Javed et al., 2024a), spontaneous multi-accent speech from 83 districts. Our protocol ensures that the typical confounds cannot produce our result: the target corpus is split speaker-disjointly, so that no test speaker is ever pseudo-labelled; an external corpus (FLEURS-Hindi, Conneau et al., 2023) is evaluated but never pseudo-labelled, so a gain there cannot be transductive; *oracle* runs using the true target transcripts quantify the ceiling; and every condition is run with ten seeds and compared using paired Wilcoxon signed-rank tests with Cliff's δ.

Our contributions are:

1. **A controlled measurement of self-training under read-to-spontaneous shift, replicated on two unlabelled pools.** One hundred and eighty LoRA fine-tuning runs: nine configurations $\times$ ten seeds $\times$ two adaptation pools. Self-training reduces held-out WER by 6.5 points in both ($0.4300 \to 0.3648$ and $0.4198 \to 0.3545$; $10/10$ seeds, $p = 0.002$, $\delta = 1.00$), improves FLEURS by 2.3 and 3.1 points, improves source-domain accuracy, and beats a teacher trained on 128.5 labelled hours on every seed. At that operating point it recovers 38.4% and 43.9% of the improvement that true target labels would buy, at no annotation cost.
2. **Replication of a quality–quantity structure with an interior optimum.** After controlling for retained volume, the retained pseudo-label error rate is perfectly rank-ordered with downstream WER in both pools (Spearman $\rho = +1.00$). Although the pools differ by $4.6\times$ in size and their teachers by 15 points in quality, the optimum is the same in both, $\theta = -0.15$, and in both the strictest threshold yields the cleanest labels in the study together with a worse model.
3. **A pre-registered prediction, and the lesson it teaches.** We fitted a slope on the first pool (0.73) and pre-registered a prediction for the second before running it. The ordering and direction transferred; the magnitude did not (predicted 3.7 points, observed 2.65). Segment-wise analysis shows why: the relationship is strongly concave in both pools, so a linear coefficient is a corpus-specific summary rather than a law.
4. **Iteration saturates, and we show why.** A second round, using the best round-one student as teacher, changes target WER by 0.0005 — 0.21 of the pooled seed standard deviation — and slightly degrades both other test sets. The pseudo-labels did improve substantially ($0.3647 \to 0.3364$ unfiltered, with degenerate hypotheses falling from 34 to 13), but at the operating threshold the improvement was only $0.3240 \to 0.3203$: iteration corrected the labels that filtering had already removed. Confidence filtering and iterative self-training are *substitutes*, not complements.
5. **Scope conditions and practical guidance.** Report the retained pseudo-label error rate and hours rather than the threshold, which is uninformative: two thresholds differing by 0.20 can differ by 0.0001 in retained error. Filter until the retained hours start to drop significantly, then stop; a second round is not worth its compute once the first is well filtered.

2 Related Work

Self-training is an established technique in semi-supervised ASR. Early methods trained a student from scratch on labelled and pseudo-labelled data, sometimes over several rounds, and generally used a language model during pseudo-label generation (Kahn et al., 2020; Park et al., 2020). A known drawback is the slow decoding time of the language model, and the acoustic model can overfit to the LM's text distribution (Rouditchenko et al., 2024). Continuous variants resolve both issues: slimIPL keeps a dynamic cache of unlabelled samples and regenerates hard pseudo-labels using greedy decoding without a language model (Likhomanenko et al., 2021), while momentum pseudo-labelling employs a teacher whose weights are an exponential moving average of the student's (Higuchi et al., 2021); both avoid the collapse that follows from regenerating labels too aggressively. Berrebbi et al. (2023) demonstrate that in low-resource settings the initial supervised phase these

methods require can itself lead to overfitting, and that additional pre-training can have a negative impact at test time. We are intentionally less elaborate than these methods one teacher, one filtering step, offline pseudo-labels because our object of study is the filtering decision, not the training schedule.

In semi-supervised classification, fixed-threshold approaches such as FixMatch (Sohn et al., 2020) admit unlabelled samples whose confidence exceeds a threshold τ. Chen et al. (2023) formalise the resulting tension: raising the threshold makes retained pseudo-labels more likely to be correct but directly reduces how many are retained; their remedy is a soft, distribution-aware weighting. In ASR, filtering is typically applied to utterance-level confidence scores for instance length-normalised shallow-fusion scores, with the threshold relaxed gradually across self-training iterations (Singh et al., 2023) or handled by automatic thresholding that adapts to the confidence distribution (Zhu et al., 2023); in either case the threshold trades coverage against label noise. Recent work on ASR domain adaptation combines pseudo-labelling with multi-stage filtering and data selection (Rangappa et al., 2025; Park & Hain, 2025; Damianos et al., 2025), and Hwang et al. (2022) investigate self- and semi-supervised learning for large-scale ASR domain adaptation. What this literature lacks, and what we provide, is a measurement of how downstream error changes with the retained pseudo-label error rate when the retained volume is held fixed; the two axes are almost always varied together, which makes their contributions unidentifiable.

LoRA (Hu et al., 2022) is now the default adapter for Whisper. Song et al. (2024) achieve near-full-fine-tuning quality on Whisper-small with language-specific LoRA matrices of rank 32, and LoRA-adapted Whisper has been reported for many low-resource languages and domains (Wang et al., 2025). Relevant to our setting, Juvekar et al. (2026) report precisely the mismatch we target: fine-tuning multilingual ASR models for low-resource languages generally improves read speech while degrading performance on spontaneous audio. We use a standard LoRA recipe and hold it fixed across all runs, so that differences between runs are attributable to data rather than to adapter tuning.

RESPIN-S1.0 (Kumar et al., 2026) provides read speech in nine Indian languages with dialect labels that are acoustically meaningful: multimodal dialect identification reaches roughly 80% accuracy on a subset of the corpus (Amartyaveer et al., 2025). Lahaja (Javed et al., 2024a) is a spontaneous Hindi benchmark, IndicVoices (Javed et al., 2024b) supplies natural and extempore speech at scale, and FLEURS (Conneau et al., 2023) provides a small read benchmark for cross-lingual comparison. Previous pseudo-labelling research on Hindi has focused on scaling (Bhogale et al., 2024), and data augmentation is a complementary low-resource remedy (Bartelds et al., 2023); self-supervised fine-tuning has likewise been applied to other Indian languages (Vaddepalli & Neelakandan, 2026). To our knowledge there has been no published downstream work on RESPIN Hindi, whose follow-ups concentrate on dialect identification and on the other eight languages.

3 METHOD

We are given a labelled source corpus $\mathcal{S} = \{(x_i, y_i)\}$ of read speech, an unlabelled target pool $\mathcal{U} = \{x_j\}$ of spontaneous speech, and a transcription budget that permits using only a small subset of $\mathcal{S}$. The objective is a model that performs well on the target domain. At no time does the method have access to any target transcripts.

Figure 1 summarises the pipeline. A teacher f_T, obtained by LoRA fine-tuning the foundation model on $\mathcal{S}$, greedily decodes each $x_j \in \mathcal{U}$, yielding a hypothesis $\hat{y}_j$ and a sequence-level confidence

$$c_j \;=\; \frac{1}{|\hat{y}_j|} \sum_{t=1}^{|\hat{y}_j|} \log p_{f_T}(\hat{y}_{j,t} \mid \hat{y}_{j,<t}, x_j)\,, \tag{1}$$

the mean token log-probability. The set retained at threshold θ is

$$\mathcal{U}_\theta \;=\; \{\,(x_j, \hat{y}_j) \;:\; c_j \geq \theta \;\wedge\; \mathrm{sane}(\hat{y}_j)\,\}, \tag{2}$$

where $\mathrm{sane}(\cdot)$ rejects degenerate hypotheses (below). A student is then trained from the *unadapted* foundation model on $\mathcal{S}_{\text{sub}} \cup \mathcal{U}_\theta$, where $\mathcal{S}_{\text{sub}} \subset \mathcal{S}$ is a fixed-size labelled subset.

On out-of-domain audio Whisper can fall into repetition loops, and a single looping hypothesis can introduce hundreds of spurious tokens. Independently of confidence we reject any hypothesis

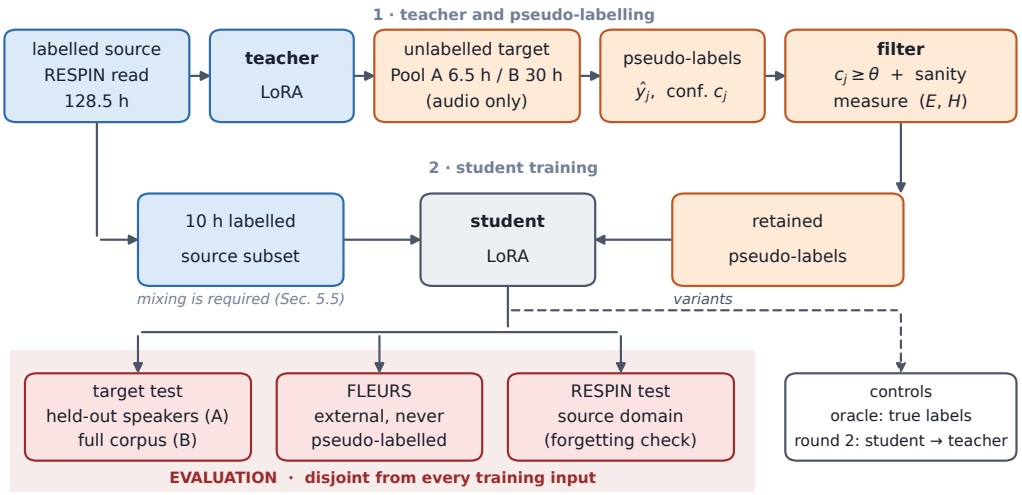

Figure 1: The pipeline. A teacher LoRA-tuned on the full labelled source corpus pseudo-labels an unlabelled target pool; hypotheses are filtered by confidence and sanity checks, and we record the retained pseudo-label error rate E and retained hours H. A fresh student is trained on the retained pseudo-labels mixed with a 10 h labelled source subset. Nothing inside the evaluation band is ever pseudo-labelled or trained on: in Pool A the target pool and target test set are speaker-disjoint, in Pool B they come from different corpora, and FLEURS is external to both. Two controls are not shown: the *oracle* replaces the pseudo-labels with the pool's true transcripts, and *round 2* promotes the best student to teacher and repeats the cycle.

whose 4-gram repetition ratio exceeds 0.30, whose fraction of Devanagari letters falls below 0.80, or whose character rate falls outside $[2, 30]$ characters per second. These filters eliminated 34 of 3,220 hypotheses and are applied in every self-training condition, including the unfiltered one, so that they cannot account for differences between conditions.

Two quantities we measure. For each threshold we record the retained *volume* $H(\theta) = \sum_{j \in \mathcal{U}_\theta} d_j$ (hours) and, because the target transcripts of the adaptation split are withheld rather than discarded, the retained *pseudo-label error rate* $E(\theta) = \mathrm{WER}(\{\hat{y}_j\}_{\mathcal{U}_\theta}, \{y_j\}_{\mathcal{U}_\theta})$. The true labels are never used in training in any condition except the *oracle* ceiling; $E(\theta)$ is a measurement instrument, and in deployment it would be estimated rather than computed. The central empirical question of this paper is how downstream WER depends on the pair (E, H).

4 Experimental Setup

RESPIN-S1.0 Hindi (Kumar et al., 2026): 128.5 h / 85,800 utterances of scripted read speech, five dialects, balanced by dialect, domain and gender, mean utterance duration 5.4 s, with speaker-disjoint dev and test splits. **Target (test).** Lahaja (Javed et al., 2024a): 12.32 h / 6,152 utterances, 132 speakers, 83 districts, spontaneous, 16 kHz. The corpus `text` field contains only pure Devanagari; the `normalized` field carries bracketed loanword annotations on 64% of rows and would inflate WER spuriously. **External.** FLEURS hi_in test (Conneau et al., 2023): 418 utterances, 1.34 h, never pseudo-labelled.

The whole study is conducted twice. *Pool A (in-corpus, 6.55 h).* Lahaja is split by speaker: half the speakers (3,220 utterances) form the unlabelled pool and the disjoint half is the test set, so no test speaker is ever pseudo-labelled. *Pool B (cross-corpus, 30.0 h).* The pool is instead drawn from IndicVoices Hindi (Javed et al., 2024b), 17,391 utterances of natural and extempore speech, and all of Lahaja becomes the test set. Pool B is the stronger design — it shares no speaker, session or corpus with the evaluation data and doubles the test set — and Pool A is retained because it permits the speaker-disjoint comparison and a smaller volume regime. The teacher is markedly weaker on Pool B

(0.5138 pseudo-label WER, against 0.3647 on Pool A), and its degenerate-hypothesis rate is higher (447 of 17,391 rejected, against 34 of 3,220).

We also perform one additional iteration of self-training on Pool A: the student with the lowest target WER (seed 55, 0.3626) becomes the new teacher, re-labels the same unlabelled pool, and ten fresh students are trained at the best round-one threshold. Pool, test set, budget, recipe and seeds are all held constant, so this isolates pseudo-label quality as the only varying factor — the cleanest such comparison in the study.

Whisper-small (244M) with LoRA (r=32, α=64, on q and v projections; 3.54M trainable parameters, 1.44% of the model), learning rate 10^{-3}, 10% warmup, batch 16, 3 epochs, fp16. The labelled subset is 10 h drawn at random from the RESPIN pool using the run's seed. All hyperparameters are identical across all conditions; only the training data varies. The teacher is a LoRA model trained on the entire 128.5 h labelled pool.

Nine configurations $\times$ ten seeds ($11, 22, \ldots, 110$): `baseline_respin` (10 h labelled only); five thresholds $\theta \in \{-0.70, -0.50, -0.30, -0.15, -0.05\}$; `st_nofilter` (all hypotheses, sanity filters only); `st_pseudo_only` (filtered pseudo-labels, no labelled data); and `oracle` (10 h labelled + the adaptation split with its true transcripts). Seeds vary both the labelled subset and training stochasticity.

We report WER on held-out Lahaja speakers, FLEURS, and RESPIN test. All references are NFC-normalised with punctuation and bracketed annotations removed. Comparisons are paired by seed and tested with the two-sided Wilcoxon signed-rank test; with n=10 the smallest attainable p-value is 0.002. We report Cliff's δ (Cliff, 1993) and seed-win counts, and apply Holm correction within each test set across the eight comparisons against the baseline.

5 Results

5.1 Self-training closes a large part of the read-to-spontaneous gap, in both pools

Tables 1 and 2 give the two sweeps. On this material zero-shot Whisper-small is unusable, reaching 1.29 and 1.21 WER because of hallucination loops on spontaneous Hindi. Fine-tuning on 10 labelled hours reaches 0.4300 / 0.4198. Adding unlabelled target audio through confidence-filtered self-training reaches 0.3648 / 0.3545 at the best threshold: a 6.5-point absolute, 15% relative reduction in both pools, on 10/10 seeds with $p = 0.002$ and $\delta = 1.00$.

Three properties hold in both, quoted here at the best-WER threshold $\theta = -0.15$. The gain is *not transductive*: FLEURS contains no pseudo-labelled sample yet still gains by 2.3 and 3.1 points, and in Pool B the adaptation audio shares no corpus with the test set at all. There is *no catastrophic forgetting*: source-domain WER improves in both ($0.2123 \to 0.2033$; $0.2136 \to 0.1892$); the largest source-domain gain in Pool B, $0.2136 \to 0.1844$, occurs at $\theta = -0.70$. And the student *beats its own teacher* on the target domain four of five thresholds in Pool A and five of five in Pool B, each on 10/10 seeds, against a teacher trained on 128.5 labelled hours. Six to thirty hours of unlabelled in-domain audio is worth more, for the target domain, than 118 additional hours of labelled out-of-domain audio.

The *oracle* condition the same adaptation utterances with their true transcripts reaches 0.2602 / 0.2712. At $\theta = -0.15$, self-training therefore recovers 38.4% / 43.9% of the improvement that perfect target labels would buy, at zero annotation cost.

5.2 Quality governs the gain with diminishing returns, until quantity binds

Restricting to conditions that retain at least 90% of the maximum volume, downstream WER is perfectly rank-ordered by the retained pseudo-label error rate E in both pools: $\rho = +1.00$ on Pool A ($n = 5$, $H \in [6.06, 6.55]$, exact two-sided $p = 0.017$) and $\rho = +1.00$ on Pool B ($n = 4$, $H \in [28.6, 30.0]$, $p = 0.083$). Including the $\theta = -0.15$ condition of Pool B, which retains 75% of volume, preserves perfect monotonicity at $n = 5$ ($p = 0.017$). Ordinary least squares over these points gives slopes of 0.73 (Pool A) and 0.21 (Pool B); restricting Pool B further, to the four conditions retaining $\geq 90\%$ of volume, raises its slope to 0.56. That the coefficient depends this

Table 1: **Pool A: in-corpus adaptation** (6.55 h of unlabelled Lahaja, speaker-disjoint from the test half). WER, mean ± std over ten seeds. E = retained pseudo-label error rate, H = retained hours. All comparisons against the baseline have $p = 0.002$, $\delta = \pm 1.00$, $10/10$ seeds (Wilcoxon, Holm-corrected) except RESPIN at $\theta{=}-0.70$ ($p = 0.084$, $9/10$).

Condition	θ	E	H (h)	Lahaja	FLEURS	RESPIN
Zero-shot	—	—	—	1.2928	0.6485	0.8443
Teacher (128.5 h)	—	—	—	0.3853	0.2854	0.1209
`baseline`	—	—	0	0.4300 ±.0117	0.3475 ±.0043	0.2123 ±.0021
`nofilter`	—	.3647	6.55	0.3970 ±.0084	0.3248 ±.0037	0.2035 ±.0020
$\texttt{st}_{-0.70}$	−0.70	.3495	6.54	0.3777 ±.0102	0.3252 ±.0039	0.2057 ±.0099
$\texttt{st}_{-0.50}$	−0.50	.3494	6.54	0.3719 ±.0038	**0.3235** ±.0019	**0.2024** ±.0024
$\texttt{st}_{-0.30}$	−0.30	.3472	6.52	0.3691 ±.0021	0.3245 ±.0044	0.2038 ±.0019
$\texttt{st}_{-0.15}$	−0.15	.3240	6.06	**0.3648** ±.0020	0.3242 ±.0030	0.2033 ±.0021
$\texttt{st}_{-0.05}$	−0.05	.2125	2.14	0.3741 ±.0019	0.3321 ±.0022	0.2064 ±.0019
`pseudo_only`	−0.70	.3495	6.54	0.3970 ±.0036	0.3716 ±.0021	0.3214 ±.0020
`oracle`	—	.0000	6.55	0.2602 ±.0011	0.3120 ±.0029	0.2037 ±.0015

Table 2: **Pool B: cross-corpus adaptation** (30 h of unlabelled IndicVoices; the test set is *all* of Lahaja, which shares no corpus with the adaptation pool). All comparisons against the baseline have $p = 0.002$, $\delta = \pm 1.00$, $10/10$ seeds.

Condition	θ	E	H (h)	Lahaja	FLEURS	RESPIN
Zero-shot	—	—	—	1.2106	0.6485	0.8443
Teacher (128.5 h)	—	—	—	0.3741	0.2854	0.1209
`baseline`	—	—	0	0.4198 ±.0074	0.3473 ±.0043	0.2136 ±.0021
`nofilter`	—	.5138	30.00	0.3830 ±.0055	0.3133 ±.0040	0.1846 ±.0025
$\texttt{st}_{-0.70}$	−0.70	.4797	29.74	0.3608 ±.0025	0.3123 ±.0029	**0.1844** ±.0017
$\texttt{st}_{-0.50}$	−0.50	.4770	29.57	0.3586 ±.0020	**0.3122** ±.0036	0.1855 ±.0023
$\texttt{st}_{-0.30}$	−0.30	.4637	28.62	0.3565 ±.0017	0.3133 ±.0037	0.1854 ±.0016
$\texttt{st}_{-0.15}$	−0.15	.3999	22.45	**0.3545** ±.0015	0.3166 ±.0055	0.1892 ±.0018
$\texttt{st}_{-0.05}$	−0.05	.2248	4.75	0.3594 ±.0013	0.3285 ±.0024	0.2010 ±.0018
`pseudo_only`	−0.70	.4797	29.74	0.3739 ±.0046	0.3346 ±.0027	0.2383 ±.0031
`oracle`	—	.0000	30.00	0.2712 ±.0018	0.3016 ±.0035	0.1959 ±.0019

strongly on where the volume cut is drawn is itself informative, and motivates the concavity analysis below.

The relationship is concave, not linear. The fitted slopes vary by a factor of three between pools, and within Pool B by a factor of 2.7 depending on the volume cut-off, which shows why a single coefficient is not the appropriate summary. Moving from the unfiltered pool to $\theta = -0.30$ removing the worst labels converts quality into accuracy at a rate of 1.27 (Pool A) and 0.65 (Pool B) points per point. Moving further, from $\theta = -0.30$ to $\theta = -0.15$, converts at only 0.19 and 0.03. The majority of the gain comes from discarding poor pseudo-labels; any further cleaning buys very little. A linear fit therefore represents a corpus-specific mixture of these two regimes, not a transferable constant.

In both pools the strictest threshold, $\theta = -0.05$, produces the cleanest pseudo-labels in the study ($E = 0.2125$, 0.2248) together with a worse model than $\theta = -0.15$. It retains 34% (Pool A) and 14% (Pool B) of the pool. The optimum is interior and falls at the same threshold in both runs, even though the pool size differs by $4.6\times$, the teacher quality by 15 points, and the adaptation corpus entirely.

Computed over all six conditions, Spearman(E, WER) and Spearman(H, WER) are numerically identical in both pools ($+0.657$, $p = 0.156$): loosening the threshold adds hours and errors together, so the two axes are rank-equivalent and their contributions cannot be separated. The only way to isolate quality is to restrict to a volume-matched regime. We consider this a methodological issue for

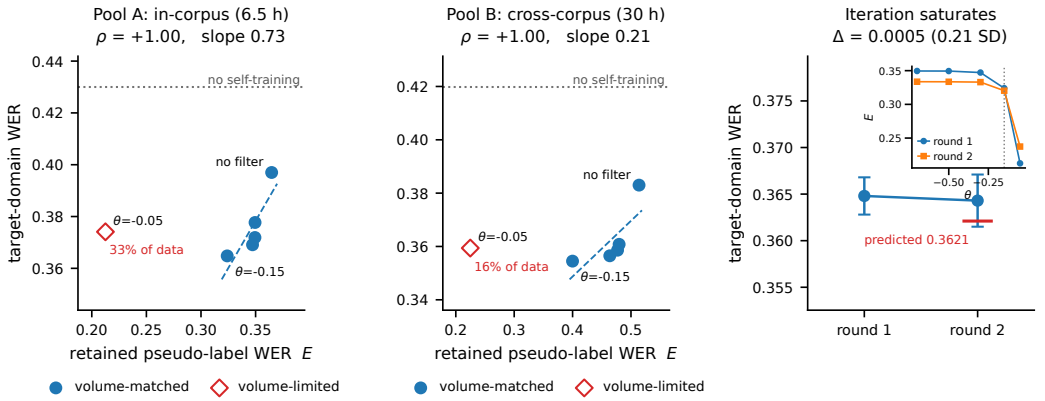

Figure 2: **Left, centre:** target-domain WER against the retained pseudo-label error rate E for the two unlabelled pools; the three unlabelled points between $\theta = -0.15$ and no filtering are, in order of increasing E, $\theta = -0.30$, -0.50 and -0.70. Volume-matched conditions (circles) are perfectly rank-ordered by E in both; the volume-limited conditions (diamonds) have the cleanest labels in the study and worse models. Dotted line: no self-training. **Right:** a second self-training round changes target WER by less than a quarter of a seed standard deviation, against the improvement the Pool A slope predicts (red). **Inset:** retained pseudo-label error by threshold in the two rounds — the new teacher is much better at loose thresholds and barely better at the operating point ($\theta = -0.15$, dotted), which is why iteration saturates.

the wider literature: when a threshold is reported without the resulting (E, H) pair, the mechanism is left unidentified.

5.3 A PRE-REGISTERED PREDICTION, AND WHAT TRANSFERRED

Before running Pool B we recorded a prediction based on the slope fitted on Pool A: applying the fitted slope 0.73 to the Pool B quality range ($\Delta E = 0.5138 - 0.4637 = 0.0501$ between the unfiltered condition and $\theta = -0.30$) predicts a downstream difference of 3.7 points, with monotone ordering across the volume-matched conditions and a reversal at $\theta = -0.05$.

All of the qualitative content of the prediction ordering, reversal, and the location of the optimum was reproduced with complete fidelity ($\rho = +1.00$). The quantitative content was not: the observed difference was 2.65 points, 72% of the prediction, and the slope fitted on Pool B's volume-matched conditions is 0.21–0.56 depending on the cut-off, against Pool A's 0.73. Because the two pools use different test sets (a held-out speaker half against the whole corpus), only the slope and the ordering are testable; we report the miss rather than refitting, because it delimits the claim. What transfers is the structure, not the coefficient.

5.4 A SECOND ROUND SATURATES

The quality–gain structure of Section 5.2 suggests a reason why iterative self-training should help: a better teacher should produce cleaner pseudo-labels, which should in turn produce a better student. The first half of that chain held. Promoting the best round-one student (0.3626) to teacher reduced unfiltered pseudo-label error from 0.3647 to 0.3364, and the number of degenerate hypotheses rejected by the sanity filters fell from 34 to 13.

The second half did not. Ten round-two students at $\theta = -0.15$ reach 0.3643 ± 0.0028, against round one's 0.3648 ± 0.0020: a difference of 0.0005, or 0.21 pooled seed standard deviations, and both other test sets are marginally worse (FLEURS 0.3278 against 0.3242; RESPIN 0.2056 against 0.2033). The round-two students do not outperform the round-one model that taught them.

The mechanism is visible in the retained-quality curves (Fig. 2, inset). The teacher improved most where confidence is low: unfiltered error fell by 2.8 points, but at the operating threshold $\theta = -0.15$ it fell by only 0.4 points ($0.3240 \rightarrow 0.3203$). *Iteration repaired exactly the labels that filtering had*

already discarded. Filtering and iteration are therefore substitutes rather than complements: they eliminate the same errors, and once the first is done well the second has little left to contribute. We note that this also makes the round-two comparison underpowered by construction at $\Delta E = 0.0037$ the Pool A slope predicts a 0.27-point gain, which is below the seed noise so we report the saturation and its mechanism rather than treating this as a test of the slope.

5.5 Scope conditions

External-domain behaviour is dominated by volume, not quality. On FLEURS the volume-matched conditions of Pool B span only 0.3122–0.3133 while their Lahaja values span 0.3565–0.3830; in Pool A the FLEURS spread is likewise 1.7 points against Lahaja's 3.2. In both pools the only FLEURS outlier is $\theta = -0.05$ (0.3321, 0.3285), the volume-limited condition. Pseudo-label *quality* thus affects mainly the domain the labels were drawn from, whereas the volume collapse is visible everywhere.

Labelled source data is required, and more so when the pool is small. Training on pseudo-labels alone degrades the source domain in both pools, but the magnitude depends sharply on pool size: -10.9 points with 6.5 h (Pool A: $0.2123 \rightarrow 0.3214$) and -2.5 points with 30 h (Pool B: $0.2136 \rightarrow 0.2383$), both $0/10$ seeds, $p = 0.002$. In Pool A it is also worse than the baseline on FLEURS. The labelled source data anchors the student against the teacher's systematic errors, and that anchor matters most when the unlabelled pool is too small to provide its own regularisation.

Standard deviations are low across the board (≤ 0.012) and no run diverged. Every comparison against the baseline reaches $p = 0.002$ with $\delta = \pm 1.00$ except one source-domain comparison in Pool A ($\theta = -0.70$, $p = 0.084$, $9/10$ seeds).

6 Discussion

Our results indicate that the quantities the threshold controls are a much better thing to report than the threshold itself. The conditions at $\theta = -0.70$ and $\theta = -0.50$ differ by 0.20 in threshold, by 0.0001 in retained error rate and by less than a minute of audio in retained volume, yet differ by 0.6 WER points downstream. The threshold is an uninformative label for the intervention; the pair (E, H) is not. We therefore recommend that self-training papers report both, estimating E where true target labels are unavailable, and that filtering be tightened only while H is approximately maintained. Operationally: sweep from loose to strict and stop at the last threshold before the retained hours drop significantly. In both pools that rule selects $\theta = -0.15$, the empirical optimum.

Chen et al. (2023) argue that fixed thresholding can never overcome the quantity–quality compromise, and recommend soft, distribution-aware weighting instead. Our measurements are consistent with that framing and also clarify it. In the volume-conserving regime the trade-off is not prominent: quality improvements are almost costless, and a fixed threshold performs a useful service. The trade-off only comes into play where the threshold starts to constrain volume, and that is also where the fixed threshold becomes ineffective. This yields a specific prediction: soft-weighting techniques should be most effective in the strict-threshold regime and less effective in the loose one. To our knowledge this has not been tested in ASR.

Conventionally both confidence filtering and iterative self-training are recommended, and they are often used in combination; our measurements indicate that they compete for the same gain. The improved teacher fixed the low-confidence hypotheses which a well-chosen threshold had already excluded and left the retained set essentially unchanged. A practitioner who filters near the optimum should therefore not expect a second round to pay for the compute it uses, whereas a practitioner who cannot filter (no confidence estimate, or a pool too small to discard from) should expect iteration to recoup part of what filtering would have provided. Continuous pseudo-labelling methods (Likhomanenko et al., 2021; Berrebbi et al., 2023) regenerate labels many times and could therefore be most useful in the unfiltered regime; this is testable and, to our knowledge, untested.

Our pre-registered prediction (Section 5.3) reproduced the ordering but was off by 28% in magnitude. The segment-wise analysis explains this: the quality–gain curve is concave, with a steep initial regime (removing the worst labels) and a nearly flat final one (further purification), so any linear fit is a weighted average of the two, with weights determined by where the corpus's confidence distribution

places its mass. A teacher with a heavier tail of bad hypotheses as on Pool B, where pseudo-label error is 15 points worse spends more of its range in the steep regime. Predicting the magnitude therefore requires modelling the confidence distribution and not merely its induced error rate, which we regard as the natural next step rather than a defect of the measurement.

Limitations. This is one language, one model size and one source corpus, with two target pools. The concavity and the interior optimum are repeated across pools spanning a wide range of size, corpus and teacher quality, but whether the *form* of the relationship transfers across languages and model scales is untested; our protocol is set up to be replicated mechanically for that purpose. The confidence measure is a mean token log-probability, which is imperfectly calibrated (Wu et al., 2025); a better quality estimator would shift points along the E axis and, by our results, should yield predictable improvements in the steep regime. Finally, we consider a single offline round; with continuous pseudo-labelling (Likhomanenko et al., 2021; Berrebbi et al., 2023) labels are regenerated during training and the picture may shift, though this also makes E a moving quantity and harder to measure.

7 Conclusion

Rather than assuming what confidence filtering buys in self-training under read-to-spontaneous domain shift, we quantified it, across two unlabelled pools differing in size, corpus and teacher quality, plus a second self-training round. Self-training improves held-out WER by 6.5 points in both pools, extends to an external corpus it has never seen, does not adversely affect the source domain, and outperforms a teacher trained on twenty times as much labelled data. Once retained volume is fixed, downstream error is exactly rank-ordered by the error rate of the retained pseudo-labels in both pools, with strongly diminishing returns and an interior optimum at the same threshold in both. What transfers across pools is this structure and not the slope: our pre-registered quantitative prediction was 28% optimistic. And a second round adds nothing measurable, because the improved teacher repairs the labels that filtering has already removed filtering and iteration are substitutes. Quality before quantity, with diminishing returns, and only until quantity runs out.

AI use statement

In this work, we used generative AI tools as a coding assistant for implementing the experimental pipeline (data loading, pseudo-label generation and filtering, the training and evaluation loops, and the statistical analysis). We take responsibility for the final content of this work, including all text, claims and artifacts.

Ethics statement

The corpora used here are publicly released for research with speaker consent obtained by their creators; we add no new recordings and attempt no speaker identification. Improving ASR for spontaneous, dialect-diverse Hindi is intended to reduce the accessibility gap faced by speakers whose speech is poorly served by systems trained on read corpora. We note a risk specific to the method studied: self-training amplifies whatever biases the teacher holds, because pseudo-labels are least reliable for the speakers the teacher already serves worst, which can widen rather than narrow subgroup performance gaps. We release per-district results so that this can be audited.

Reproducibility statement

All corpora are public: RESPIN-S1.0 (registration required), Lahaja (CC BY 4.0) and FLEURS. Section 4 specifies the speaker-disjoint split seed, the LoRA configuration, the reference-normalisation procedure, the ten seeds, and the statistical tests; Section 3 gives the confidence measure and the exact degenerate-hypothesis filter constants. Pseudo-labels and data selections are frozen to disk before any training, so every run is exactly reproducible from the released artifacts; the prediction of Section 5.3 was recorded before Pool B was run; per-run hypotheses and metrics are included in the supplementary material. The training recipe is identical across all one hundred and ninety runs (two pools plus one iteration), no hyperparameter is tuned per condition or per pool, and the confidence

threshold for the default system was fixed a priori at Whisper's conventional -0.7 before the sweep was run.

REFERENCES

Amartyaveer, Saurabh Kumar, Sumit Sharma, Sathvik Udupa, Sandhya Badiger, Abhayjeet Singh, G. Deekshitha, Jesuraja Bandekar, Savitha Murthy, and Prasanta Kumar Ghosh. Improving dialect identification in indian languages using multimodal features from dialect informed asr. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. doi: 10.1109/ICASSP49660.2025.10889099.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Dan Berrebbi, Ronan Collobert, Navdeep Jaitly, and Tatiana Likhomanenko. Continuous pseudo-labeling from the start. In *International Conference on Learning Representations (ICLR)*, 2023.

Kaushal Santosh Bhogale, Deovrat Mehendale, Niharika Parasa, Tahir Javed, Pratyush Kumar, Mitesh M Khapra, et al. Empowering low-resource language asr via large-scale pseudo labeling. *arXiv preprint arXiv:2408.14026*, 2024.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Norman Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993.

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

Dimitrios Damianos, Georgios Paraskevopoulos, and Alexandros Potamianos. Msda: Combining pseudo-labeling and self-supervision for unsupervised domain adaptation in asr. In *Proc. Interspeech*, pp. 3863–3867, 2025.

Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori. Momentum pseudo-labeling for semi-supervised speech recognition. In *Proc. Interspeech*, pp. 726–730, 2021.

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

Dongseong Hwang, Ananya Misra, Zhouyuan Huo, Nikhil Siddhartha, Shefali Garg, David Qiu, Khe Chai Sim, Trevor Strohman, Françoise Beaufays, and Yanzhang He. Large-scale asr domain adaptation using self- and semi-supervised learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6627–6631, 2022.

Tahir Javed, Janki Nawale, Sakshi Joshi, Eldho George, Kaushal Bhogale, Deovrat Mehendale, and Mitesh M Khapra. Lahaja: A robust multi-accent benchmark for evaluating hindi asr systems. *arXiv preprint arXiv:2408.11440*, 2024a.

Tahir Javed et al. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. In *Findings of the Association for Computational Linguistics (ACL)*, 2024b.

Kush Juvekar, Kavya Manohar, Aditya Srinivas Menon, Arghya Bhattacharya, and Kumarmanas Nethil. Vividh-asr: A complexity-tiered benchmark and optimization dynamics for robust indic speech recognition. *arXiv preprint arXiv:2605.13087*, 2026.

Jacob Kahn, Ann Lee, and Awni Hannun. Self-training for end-to-end speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7084–7088, 2020.

Saurabh Kumar, Abhayjeet Singh, Jesuraj Bandekar, Savitha Murthy, Sumit Sharma, Sandhya Badiger, Sathvik Udupa, Amala Nagireddi, Srinivasa Raghavan KM, Rohan Saxena, et al. Respin-s1. 0: A read speech corpus of 10000+ hours in dialects of nine indian languages. *Advances in Neural Information Processing Systems*, 38, 2026.

Tatiana Likhomanenko, Qiantong Xu, Jacob Kahn, Gabriel Synnaeve, and Ronan Collobert. slimipl: Language-model-free iterative pseudo-labeling. In *Proc. Interspeech*, pp. 741–745, 2021.

Chanho Park and Thomas Hain. Semi-supervised learning for automatic speech recognition with word error rate estimation and targeted domain data selection. In *Proc. Interspeech*, pp. 3663–3667, 2025.

Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le. Improved noisy student training for automatic speech recognition. In *Proc. Interspeech*, pp. 2817–2821, 2020.

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

Pradeep Rangappa, Andres Carofilis, Jeena Prakash, Shashi Kumar, Sergio Burdisso, Srikanth Madikeri, Esaú Villatoro-Tello, Bidisha Sharma, Petr Motlicek, Kadri Hacioglu, Shankar Venkatesan, Saurabh Vyas, and Andreas Stolcke. Efficient data selection for domain adaptation of asr using pseudo-labels and multi-stage filtering. In *Proc. Interspeech*, pp. 4928–4932, 2025.

Andrew Rouditchenko, Ronan Collobert, and Tatiana Likhomanenko. Av-cpl: Continuous pseudo-labeling for audio-visual speech recognition. *arXiv preprint arXiv:2309.17395*, 2024.

Satwinder Singh, Feng Hou, and Ruili Wang. A novel self-training approach for low-resource speech recognition. In *Proc. Interspeech*, pp. 1588–1592, 2023. doi: 10.21437/Interspeech.2023-540.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. Lora-whisper: Parameter-efficient and extensible multilingual asr. In *Proc. Interspeech*, 2024.

Srivarthin Vaddepalli and Renjith Prabhavathi Neelakandan. Multilingual self-supervised fine-tuning for low-resource telugu automatic speech recognition. *IEEE Access*, 14:55517–55534, 2026.

Pu Wang, Shinji Watanabe, and Hugo Van Hamme. Ssvd: structured svd for parameter-efficient fine-tuning and benchmarking under domain shift in asr. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2025.

Jiaqi Wu, Junbiao Pang, and Qingming Huang. Uncertainty-aware long-tailed weights model the utility of pseudo-labels for semi-supervised learning. *arXiv preprint arXiv:2503.09974*, 2025.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Han Zhu, Dongji Gao, Gaofeng Cheng, Daniel Povey, Pengyuan Zhang, and Yonghong Yan. Alternative pseudo-labeling for semi-supervised automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3320–3330, 2023.