Navigating the Reality Gap: On-Device Continual Adaptation of ASR for Clinical Telephony

Darshil Chauhan[1], Adityasinh Solanki[1], Vansh Patel[1], Kanav Kapoor[1], Ritvik Jain[2], Aditya Bansal[2], Pratik Narang[1], Dhruv Kumar[1]
[1]BITS Pilani, Pilani Campus, India
[2]Qure.ai, India
f20230609@pilani.bits-pilani.ac.in

Abstract

Automatic Speech Recognition (ASR) can ease clinical documentation in resource-constrained regions, but deployment is hindered by a "Reality Gap" between laboratory performance and noisy, real-world clinical telephony, compounded by strict data residency and compute constraints. We study this gap using Gram Vaani, a telephonic Hindi corpus spanning rural healthcare and agricultural helplines, as the closest publicly available proxy for clinical telephony speech, and show that a robust multilingual model (IndicWav2Vec) degrades from 11.60% WER on clean read Hindi to 41.72% WER on this data. We evaluate a progression of adaptation regimes, from full fine-tuning and offline Low-Rank Adaptation (LoRA) upper bounds to an on-device, stream-based continual adaptation framework in which raw audio never leaves the local device, and characterize the trade-offs between data-driven and parameter-driven stabilization strategies. Our evaluation covers both lexical accuracy (WER and CER) and semantic fidelity (BERTScore) on the target domain, alongside the retention of general-domain knowledge. Multi-domain Experience Replay (ER) yields the primary gains, improving target WER by 18.2% relative and reducing catastrophic forgetting by 54% compared to naive adaptation, with BERTScore reflecting consistent gains in semantic fidelity. Combining replay with Elastic Weight Consolidation based on a stabilized importance estimate (Absolute Fisher) yields the strongest retention at a small cost in plasticity. Finally, a language model spot check empirically verifies that the core mismatch lies at the acoustic level and cannot be resolved by language models alone.

1 Introduction

The recent surge in Self-Supervised Learning (SSL) has propelled Automatic Speech Recognition (ASR) to near-human performance on standardized benchmarks. Foundational models like Meta's Wav2Vec 2.0 (Baevski et al., 2020) and OpenAI's Whisper (Radford et al., 2023) promise a future where automated transcription can digitize patient reports, allowing clinics to improve throughput with reduced operational costs. However, for specialized, high-impact clinical domains such as rural health helplines and remote clinical services, this promise remains unfulfilled. The "reality gap", representing the disparity between clean training corpora and the chaotic, noisy, and privacy-constrained environments of real-world clinical telephony, renders these state-of-the-art models practically unusable. Consequently, ASR in these settings is best positioned as an assistive tool in clinician-in-the-loop workflows, where improvements reduce manual documentation effort rather than enable full automation.

This mismatch stems from several compounding acoustic factors. Telephonic audio in low-resource clinics is typically band-limited to 8 kHz, introducing severe channel attenuation compared to the high-fidelity speech used to pretrain foundation models. Furthermore, rural speech often features diverse regional dialects and non-standard pronunciations that are poorly represented in general-domain corpora, while ambient background noise and low-quality telephonic channels introduce non-linear distortions.

Our baseline analysis reveals that even robust multilingual models like IndicWav2Vec (Javed et al., 2022) degrade from 11.60% WER on clean read Hindi (Kathbath) (Javed et al., 2023) to a prohibitively high **41.72% Word Error Rate (WER)** on rural telephony speech from Gram Vaani (Bhanushali et al., 2022), a corpus of healthcare and agricultural helpline calls that we use as the closest publicly available proxy for clinical telephony. This failure is compounded by a technical deadlock: patient data residency requirements preclude cloud-based adaptation, while rural clinics lack the compute for conventional retraining. We

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 124–139
November 6-10, 2026 ©2026 Association for Computational Linguistics

therefore study an on-device adaptation framework for stream-based learning in which raw patient audio never leaves the local device. This ensures data residency but is not a formal privacy guarantee.

Using Low-Rank Adaptation (LoRA) (Hu et al., 2022), the model trains only a fraction of its parameters. We compare this setting against full fine-tuning and offline LoRA upper bounds that assume unrestricted access to the target data. However, sequential adaptation risks *catastrophic forgetting* of foundational linguistic capabilities (McCloskey and Cohen, 1989). To counteract this, we interleave small buffers of general-domain data with the incoming stream through multi-domain experience replay (Chaudhry et al., 2019).

We also evaluate parameter regularization through Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), alone and combined with replay, using an Absolute Fisher importance estimate (Benzing, 2022) that is more robust than the standard squared formulation to the high-variance gradients of noisy telephony speech. We measure forgetting on two general-domain Hindi benchmarks, Kathbath and FLEURS. Since Gram Vaani does not provide named-entity or disease-name annotations, clinically targeted evaluation is not possible; we therefore quantify performance with WER and CER, complemented by BERTScore (Zhang et al., 2020) to capture semantic fidelity that lexical metrics miss. Our contributions are:

- **On-Device Clinical Adaptation:** A continual adaptation framework for clinical telephony ASR that preserves data residency and runs on consumer-grade laptops.
- **Empirical Validation on Proxy Clinical Speech:** LoRA combined with multi-domain replay yields an 18.2% relative WER reduction on rural telephony speech, averaged over three seeds and evaluated on three datasets.
- **Investigation of Optimization Robustness:** The interplay between data-driven replay and parameter-driven regularization, validating that linearized Fisher importance promotes more reliable convergence in noisy, multi-speaker telephony.
- **Quantification of the Reality Gap:** A characterization of the gap on 8kHz clinical-proxy telephony, contextualized by full fine-tuning and offline LoRA upper bounds, showing that acoustic adaptation is necessary.

2 Related Work

Designing an on-device, data-residency constrained, adaptive ASR pipeline for low-resource medical settings requires synthesizing advancements in self-supervised acoustic modeling, parameter-efficient adaptation, and continual learning. We address critical gaps regarding adaptation efficiency, data scarcity, and optimization stability.

2.1 Self-Supervised Acoustic Foundations and the Domain Gap

Low-resource speech recognition increasingly relies on Self-Supervised Learning (SSL) to leverage unlabeled audio. We build upon Wav2Vec 2.0 (Baevski et al., 2020), which achieves high data efficiency through contrastive learning; Baevski et al. (2020) showed that fine-tuning on just ten minutes of data yields competitive performance on benchmarks like Librispeech (Panayotov et al., 2015). To support linguistic diversity, we utilize IndicWav2Vec (Javed et al., 2022), scaling the architecture to cover 40 Indian languages.

The Gap: Despite these capabilities, a "usability gap" exists between laboratory benchmarks and real-world deployment. Clinical ASR errors are particularly risky because they affect medically salient information such as dosages, negations, and symptom reporting (Chiu et al., 2018; Shafran et al., 2020). This is exacerbated in rural healthcare by noisy telephonic audio and diverse regional dialects. Recent large-scale efforts such as NIRANTAR (Javed et al., 2025) have begun characterizing these challenges across 22 Indian languages, while regularizers such as EWC have been used to mitigate regional disparities in ASR (Trinh et al., 2022). We utilize the Gram Vaani ASR dataset (Bhanushali et al., 2022) to represent these challenges; Bhanushali et al. (2022) highlight that standard models struggle to generalize to such 8kHz telephony audio without targeted adaptation.

While signal-level enhancement strategies, such as bandwidth expansion and speech super-resolution (Lin et al., 2023; Li et al., 2019), offer an alternative by reconstructing high-frequency components, they introduce significant computational overhead. Deploying a separate bandwidth expansion model (often GAN or diffusion-based) alongside the ASR system contradicts the strict low-latency requirements of edge devices. Furthermore, data residency constraints preclude cloud-based services (Leroy et al., 2019), creating a deadlock

where models cannot improve because data cannot leave the local environment. While n-gram language models can improve decoding accuracy, they do not address the underlying acoustic mismatch.

For evaluation, lexical metrics such as Word Error Rate (WER) and Character Error Rate (CER) are commonly used, though they may penalize clinically acceptable synonyms or minor morphological variations. Furthermore, no universal error threshold defines clinical usability, as acceptability depends on task and acoustic conditions (Yoon, 2026). To capture semantic fidelity, we additionally report BERTScore (Zhang et al., 2020), which correlates with human judgments of semantic similarity and has been reported alongside WER and CER in recent clinical ASR systems (Yang et al., 2026).

2.2 Parameter-Efficient Fine-Tuning (PEFT) and the Efficiency Gap

Bridging the domain gap typically requires fine-tuning on target data. However, for Large Audio Models (LAMs), full fine-tuning is computationally prohibitive and prone to overfitting. This presents an *efficiency gap*: high-end GPUs are unavailable on the resource-constrained edge devices of rural hospitals.

To address this, we adopt Low-Rank Adaptation (LoRA) (Hu et al., 2022), which enables local fine-tuning on modest clinical hardware. This approach is essential for meeting healthcare **Data Residency** requirements, as it allows for localized, high-performance adaptation without needing the centralized data storage often required by federated or cloud-based approaches (Rieke et al., 2020). Recent work by Song et al. (2024) on LoRA-Whisper validates this approach for multilingual ASR, demonstrating that PEFT can effectively adapt large Transformers without catastrophic interference, and LoRA has also been regularized with EWC to add new languages to ASR foundation models (Qian et al., 2024). By constraining optimization to a low-dimensional subspace, LoRA resolves deployment challenges in data-residency constrained clinics. To understand adaptation trade-offs, we evaluate both full fine-tuning and offline LoRA fine-tuning as upper bounds.

While PEFT solves efficiency, self-improving systems face the *stability gap* of Continual Learning (CL). In specialized clinical domains, annotated audio is scarce due to privacy constraints and labeling costs (Chiu et al., 2018), and is often collected incrementally in small batches; CL in ASR has accordingly focused on adapting models sequentially without storing historical data (Sadhu and Hermansky, 2020). Such systems face *Catastrophic Forgetting*, where new training erodes previously learned capabilities (McCloskey and Cohen, 1989). In healthcare, ensuring model stability over time is a critical safety requirement for clinical evaluation cycles (De Lange et al., 2022). In ASR, naive updates on recent telephony transcripts lead to overfitting on specific speakers or acoustic conditions. Benchmarks like NIRANTAR (Javed et al., 2025) further underscore this, demonstrating that no single CL strategy currently yields consistent performance across the diverse shifts encountered in real-world Indian speech.

To mitigate this, we employ a hybrid continual learning strategy that synergizes data-level and parameter-level consolidation. First, to address the distribution shift between segments, we implement **Experience Replay (ER)** (Chaudhry et al., 2019). Building on recent regularized CL studies for Indic ASR (T and Nirmala, 2025) and episodic memory frameworks (Yang et al., 2022), we utilize a *multi-domain experience replay* mechanism. While Yang et al. (2022) focused on gradient projection, we find that a direct data mixing strategy effectively mitigates catastrophic forgetting when combined with PEFT. Our buffer retains a mixture of "hard" examples (high-loss) from the target domain (Gram Vaani) and gender-balanced samples from a clean auxiliary domain (Kathbath).

Complementing this, we constrain parameter drift using **Online LoRA-EWC** (Kirkpatrick et al., 2017; Zheng et al., 2026), utilizing the **Absolute Fisher** importance estimation (Benzing, 2022); weight regularization has been shown to protect shared low-rank updates in continual learning (Zheng et al., 2026). We estimate importance using the *absolute value* of accumulated gradients ($|g|$) rather than the standard squared Fisher Information. While this linear scaling draws inspiration from the numerical stability of Memory Aware Synapses (MAS) (Aljundi et al., 2018), Absolute Fisher explicitly ties the importance to the likelihood of the adaptation task rather than unsupervised output sensitivity. By combining this regularized signal with multi-domain replay, our approach targets catastrophic forgetting through a dual mechanism: replay maintains the data distribution, while the regularization penalty preserves critical adapter parameters.

3 Methodology

We study an on-device, adaptive ASR framework designed to bridge the "reality gap" within the acoustic constraints of rural clinical telephony. Our approach focuses on Continual Learning (CL), enabling a pre-trained model to adapt sequentially to incoming data streams ($\mathcal{D}_{stream}$) while maintaining strict data residency and limiting catastrophic forgetting.

3.1 Base Acoustic Backbone

We utilize **IndicWav2Vec** as our acoustic backbone: a Wav2Vec 2.0 model pre-trained on diverse Indian languages, consisting of a convolutional feature encoder followed by a Transformer context network trained with the Connectionist Temporal Classification loss ($\mathcal{L}_{CTC}$). We select IndicWav2Vec as it achieves state-of-the-art performance across Indian languages unsupported by other architectures, while remaining compact enough for efficient inference and LoRA-based adaptation on consumer-grade laptops. We freeze the base model and inject trainable Low-Rank Adaptation (LoRA) matrices into the query and value projection layers, which also serves as a stability safeguard against large-scale corruption of pre-trained weights.

3.2 Evaluation Metrics

We evaluate performance using both lexical accuracy and semantic fidelity. Word Error Rate (WER) and Character Error Rate (CER) measure transcription accuracy at the word and character levels, respectively, as the number of substitutions, deletions, and insertions normalized by the reference length; CER provides a finer-grained analysis of phonetic accuracy, particularly useful for agglutinative languages and dialectal variations. However, these metrics rely on exact token matching and may penalize minor orthographic variations, such as *matra* errors in Devanagari, that do not significantly alter meaning. To better capture semantic fidelity, we also report BERTScore (Zhang et al., 2020), which matches each token of the predicted transcription to its most similar token in the reference (and vice versa) by cosine similarity of contextual embeddings, yielding precision P and recall R; we report their harmonic mean, $F_1 = 2PR/(P + R)$.

3.3 Progressive Adaptation Regimes

We define a progression of regimes with increasing deployment constraints. Full fine-tuning on the complete target dataset serves as an approximate upper bound when no constraints are imposed; offline parameter-efficient fine-tuning (LoRA) reduces computational overhead while still requiring full data access; and stream-based continual adaptation, the primary focus of our work, considers sequential data arrival under strict data residency constraints, reflecting real-world clinical deployment.

3.4 Data-Driven Stability: Multi-Domain Experience Replay

To mitigate catastrophic forgetting, we employ Experience Replay (ER), explicitly grounding the model's optimization trajectory with historical data. We implement a **Multi-Domain Replay** strategy that maintains a dual-source buffer $\mathcal{B}$:

- **General Domain Anchor ($\mathcal{B}_{gen}$):** A gender-balanced subset of standard Hindi samples (from Kathbath) that preserves foundational phonetic robustness across diverse speaker demographics.
- **Target Domain History ($\mathcal{B}_{spec}$):** A sliding window of "hard" examples (high loss) and random samples from previous segments, ensuring retention of recent domain-specific adaptations.

During training, the incoming stream $\mathcal{D}_{stream}$ is concatenated with samples from $\mathcal{B}$. The optimization objective for Experience Replay is:

$$\begin{aligned}\mathcal{L}_{ER} = \gamma \mathbb{E}_{x \sim \mathcal{D}_{stream}}[\mathcal{L}_{CTC}(x)] \\ +(1-\gamma)\mathbb{E}_{x \sim \mathcal{B}}[\mathcal{L}_{CTC}(x)]\end{aligned} \tag{1}$$

where γ represents the mixing ratio between new data and replayed data.

3.5 Parameter-Driven Stability: Elastic Weight Consolidation (EWC)

As a standalone strategy (V4.5) and a component of our hybrid approach, we implement Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) to prevent drift in critical parameters. Unlike ER, which requires data storage, EWC regularizes the model by penalizing changes to weights that are important for previous tasks. We compute the importance of each LoRA parameter θ_i using the diagonal of the Fisher Information Matrix (F). To ensure robust optimization against high-variance

gradients, we approximate importance using accumulated **absolute gradients** rather than squared gradients (Absolute Fisher; Benzing, 2022), which prevents transient acoustic artifacts from drowning out meaningful weights:

$$F_i = \frac{1}{N} \sum_{j=1}^{N} |\nabla_{\theta_i} \log P(y_j|x_j)| \quad (2)$$

The EWC regularization loss is then applied as a quadratic penalty:

$$\mathcal{L}_{EWC}(\theta) = \frac{\lambda}{2} \sum_i F_i(\theta_i - \theta_i^*)^2 \quad (3)$$

where θ_i^* represents the frozen optimal parameters from the previous segment and λ controls the regularization strength.

3.6 Hybrid Optimization Framework (V5.1)

To study the interplay between data-driven grounding and parameter-driven constraints, we evaluate a Hybrid ER + EWC framework (Strategy V5.1) that combines the multi-domain replay buffer with the EWC penalty:

$$\mathcal{L}_{Total} = \mathcal{L}_{ER}(\mathcal{D}_{stream}, \mathcal{B}) + \mathcal{L}_{EWC}(\theta) \quad (4)$$

where $\mathcal{L}_{ER}$ provides the gradients for acoustic adaptation and $\mathcal{L}_{EWC}$ defines a "safe" optimization region that limits catastrophic forgetting.

4 Experiments and Results

We evaluate adaptation to the target domain (Gram Vaani) and retention of general linguistic knowledge (Kathbath and FLEURS). All results are reported as mean $\pm$ standard deviation over three random seeds.

4.1 Datasets

- **Gram Vaani (Clinical Helpline Proxy)** (Bhanushali et al., 2022): This dataset consists of rural telephonic speech from healthcare and agricultural helplines (originally 8kHz, upsampled to 16kHz as required by IndicWav2Vec; all reported results, including the baseline, are computed on the upsampled audio). Capturing rural, noisy, conversational telephony, it serves as the closest available proxy for clinical telephony audio. We evaluate on the official 3-hour Evaluation set (GV_Eval_3h). To simulate a continual learning scenario, the 103 hours of training data are partitioned into sequential segments processed one by one, mimicking a live data stream.

- **Kathbath (Source Domain)**: A high-quality, read speech dataset representing standard Hindi (Javed et al., 2023). A subset of the training set (approx. 25,800 samples) forms the pool from which general-domain replay samples are drawn, and the complete validation set (3,151 samples) is used exclusively to measure catastrophic forgetting.

- **FLEURS (General Domain)** (Conneau et al., 2023): We use the labeled Hindi (`hi_in`) test split as a second benchmark to evaluate generalization and measure catastrophic forgetting under domain shift.

4.2 Continual Adaptation Paradigms

We investigate four continual learning paradigms, contextualized by upper-bound regimes. All experiments start from `IndicWav2Vec`, with a baseline of 41.72% WER on Gram Vaani, 11.60% on Kathbath, and 18.16% on FLEURS. Every model is trained for 3 epochs: per segment for the streaming strategies, and on the full dataset for the upper bounds. LoRA hyperparameters were selected based on prior literature and preliminary validation experiments, and were kept consistent across the replay, EWC, and hybrid strategies to enable fair comparison (Appendix A).

4.2.1 Upper-Bound Reference: Full and Offline LoRA Fine-Tuning

Following Section 3.3, full fine-tuning (FT) updates the entire model on the complete target training set, while offline LoRA trains adapters on the complete dataset with (i) rank 16, $\alpha = 32$ and (ii) rank 24, $\alpha = 48$. **Outcome:** Full fine-tuning reaches **30.90%** target WER but nearly doubles general-domain error (22.88% on Kathbath). Offline LoRA reaches 34.90% ($r = 16$) and 34.62% ($r = 24$) target WER, with substantial forgetting (20.25% and 21.11% on Kathbath).

4.2.2 Paradigm 1: Naive Continual Fine-tuning (Baseline)

The model is fine-tuned sequentially on the incoming stream without explicit regularization, using LoRA rank 16 ($\alpha = 32$) and 19 segments of 2,000 samples. **Outcome (V1.1):** Target WER drops to

35.02%, but general-domain error increases by an absolute **5.81%** on Kathbath (5.01% on FLEURS), indicating severe catastrophic forgetting.

4.2.3 Paradigm 2: Experience Replay (ER) Strategies

We investigate two ER configurations, increasing LoRA capacity to rank 24 ($\alpha = 48$) to accommodate the dual-task nature of rehearsal.

- **Single-Domain ER (V2.1)**: Replays 400 samples from the target-domain history per segment using a "60% hard, 40% random" selection strategy, where hard samples have a loss above $1.2\times$ the segment average. Each segment combines 1,600 new samples with 400 replay samples (24 segments of 2,000 samples).
- **Multi-Domain ER (V3.1)**: Replays a balanced buffer of 300 target samples (same 60:40 split) and 300 general-domain (Kathbath) samples, so each segment contains 1,600 new and 600 replay samples. This strategy achieved our lowest continual target WER of **34.11%** and reduced forgetting by **54%** on Kathbath (56% on FLEURS) relative to the naive baseline.

4.2.4 Paradigm 3: Elastic Weight Consolidation (EWC)

We examine whether parameter regularization alone, using the Absolute Fisher penalty (Section 3.5) with $\lambda = 10$ (larger values were overly conservative or prevented adaptation; Section 5.6.1), can protect foundational knowledge without replaying data. V4.5 adds this penalty to the naive setup, using the same 19 segments of 2,000 samples as V1.1. **Outcome (V4.5):** V4.5 achieves a competitive **34.55% WER** on the target domain, but mitigates forgetting less efficiently than the ER strategies (+4.76% on Kathbath) and shows the highest variance across seeds.

4.2.5 Paradigm 4: Hybrid ER + EWC Optimization

V5.1 combines the V3.1 replay buffer with the EWC penalty ($\lambda = 10$; Section 3.6). **Outcome (V5.1):** This yields our most stable model, with the lowest forgetting: **14.15%** WER on Kathbath (+2.55%) and **20.27%** on FLEURS (+2.11%). This stability comes at a slight cost to plasticity, with a final target WER of **34.66%**.

5 Analysis and Discussion

5.1 Comparative Performance Summary

Table 1 summarizes the final metrics. All continual paradigms bridge the "Reality Gap", reducing target WER by 16.1–18.2% relative to the baseline and converging to 34–35% WER, and forgetting trends on FLEURS closely mirror those on Kathbath. Streaming strategies reach target WER comparable to the offline LoRA upper bounds (34.11–35.02% vs. 34.62–34.90%) while retaining substantially more general-domain knowledge. BERTScore closely follows WER across datasets, indicating consistent gains in semantic fidelity, although it is not sensitive to clinically critical distinctions such as negations or numerical magnitude errors.

5.2 Stability-Plasticity Pareto Efficiency

Figures 4 and 5 plot retention against adaptation, where movement toward the bottom-right corresponds to improved adaptation with minimal forgetting. The replay-based trajectories show a clear elbow near 35% target WER, beyond which further reductions incur disproportionately high stability costs. Multi-Domain ER (V3.1) navigates this trade-off most efficiently, reaching the lowest target WER at moderate forgetting, while the Naive (V1.1) and EWC-only (V4.5) trajectories reach comparable target WER at considerably higher forgetting. This behavioral ceiling suggests that for a given model architecture and domain mismatch, there exists an optimal frontier of acoustic adaptation. Crucially, the V3.1 trajectory has not plateaued, suggesting that further gains are possible but require significantly more data or linguistic context to unlock.

5.3 Mechanism Analysis: Data vs. Parameter Anchors

Our investigation reveals a fundamental difference between data-driven and parameter-driven regularization in lightweight models.

- **Data Anchors (ER)**: Replaying Kathbath samples provides an explicit gradient signal toward a generalized linguistic center, making replay the most efficient stabilizer.
- **Parameter Anchors (EWC)**: While EWC protects weights, it is "acoustically blind", penalizing drift without reconciling new acoustic features. This is evidenced by

Paradigm	Strategy	Dataset	WER (%)	CER (%)	BERTScore (%)
Upper Bound	Full Finetune	Gram Vaani	**30.90 ± 0.10**	**13.20 ± 0.09**	**89.60 ± 0.06**
		Kathbath	22.88 ± 0.03	6.73 ± 0.01	93.86 ± 0.01
		FLEURS	28.21 ± 0.04	10.17 ± 0.03	91.57 ± 0.04
	LoRA (r=16, α=32)	Gram Vaani	34.90 ± 0.05	14.66 ± 0.05	88.26 ± 0.02
		Kathbath	20.25 ± 0.05	5.61 ± 0.03	94.51 ± 0.03
		FLEURS	26.35 ± 0.12	9.13 ± 0.02	92.08 ± 0.03
	LoRA (r=24, α=48)	Gram Vaani	34.62 ± 0.09	14.58 ± 0.09	88.38 ± 0.04
		Kathbath	21.11 ± 0.05	5.99 ± 0.02	94.20 ± 0.03
		FLEURS	26.94 ± 0.12	9.53 ± 0.04	91.85 ± 0.04
Baseline	IndicWav2Vec	Gram Vaani	41.72 ± 0.00	19.05 ± 0.00	86.26 ± 0.00
		Kathbath	**11.60 ± 0.00**	**3.41 ± 0.00**	**97.28 ± 0.00**
		FLEURS	**18.16 ± 0.00**	**6.38 ± 0.00**	**94.74 ± 0.00**
Naive	V1.1	Gram Vaani	35.02 ± 0.08	14.38 ± 0.22	88.37 ± 0.08
		Kathbath	17.41 ± 0.17	4.65 ± 0.02	95.47 ± 0.03
		FLEURS	23.17 ± 0.17	7.99 ± 0.04	93.22 ± 0.05
ER	V2.1 (Single Domain)	Gram Vaani	34.24 ± 0.05	14.28 ± 0.11	88.77 ± 0.04
		Kathbath	15.06 ± 0.09	4.05 ± 0.03	96.61 ± 0.01
		FLEURS	21.09 ± 0.09	7.32 ± 0.06	94.45 ± 0.04
	V3.1 (Multi Domain)	Gram Vaani	**34.11 ± 0.02**	14.22 ± 0.05	**88.85 ± 0.04**
		Kathbath	14.27 ± 0.02	3.87 ± 0.01	96.78 ± 0.01
		FLEURS	20.36 ± 0.09	7.10 ± 0.02	**94.65 ± 0.04**
EWC	V4.5	Gram Vaani	34.55 ± 0.38	**14.21 ± 0.10**	88.65 ± 0.22
		Kathbath	16.36 ± 1.04	4.37 ± 0.31	95.97 ± 0.55
		FLEURS	22.20 ± 1.02	7.70 ± 0.36	93.75 ± 0.61
Hybrid	V5.1 (ER + EWC)	Gram Vaani	34.66 ± 0.07	14.46 ± 0.02	88.69 ± 0.04
		Kathbath	**14.15 ± 0.01**	**3.85 ± 0.01**	**96.80 ± 0.01**
		FLEURS	**20.27 ± 0.04**	**7.04 ± 0.01**	94.62 ± 0.01

Table 1: Performance of the investigated adaptation regimes on the target (Gram Vaani), source (Kathbath), and general (FLEURS) domains, without LM decoding. Results are mean ± std over 3 seeds. Bold marks the best continual strategy per dataset and metric; bold underlined values are reference results: the best target-domain performance (full fine-tuning) and the unadapted general-domain performance (baseline). Forgetting corresponds to the increase in Kathbath/FLEURS WER over the IndicWav2Vec baseline.

V2.1 (ER) providing superior stabilization (+3.46% forgetting on Kathbath, +2.93% on FLEURS) compared to V4.5 (EWC; +4.76% and +4.04%) despite similar target performance (34.24% vs. 34.55%).

- **The LoRA-Regularization Bottleneck**: We observe a conflict in the Hybrid model (V5.1), where combining LoRA's low-rank space with EWC's parameter-level penalties overly restricts optimization paths, leading to higher target WER (34.66% vs. 34.11% for V3.1).

5.4 Convergence Dynamics

While the Naive baseline (V1.1) exhibits erratic trajectories, the ER and Hybrid paradigms demonstrate controlled, largely monotonic trends (Figure 1; per-strategy progressions are in Appendix C). These trends are consistent across seeds: replay-based strategies show the lowest seed variance (e.g., ±0.02 target WER for V3.1), whereas EWC-only adaptation (V4.5) is the most seed-sensitive (±0.38 target WER, ±1.04 Kathbath WER). This stability is a critical requirement for clinical deployment, ensuring the system does not regress after extended use. However, we observe a "soft floor" around 34% WER, which we characterize as the **Acoustic Bottleneck of Clinical Telephony**. The mismatch between 8kHz telephony and the 16kHz model expectation creates a ceiling for acoustic-only adaptation.

5.5 The Necessity of Acoustic Adaptation

To verify that acoustic adaptation is a *prerequisite* for clinical deployment, we paired the baseline and adapted models with a domain-specific 4-gram LM (Table 2). The LM improves both, but its gains remain limited without acoustic adaptation: the baseline+LM reaches only 35.21% WER, compared to 30.32% for V3.1+LM. With the LM, full fine-tuning reaches 29.03% WER (at the cost of

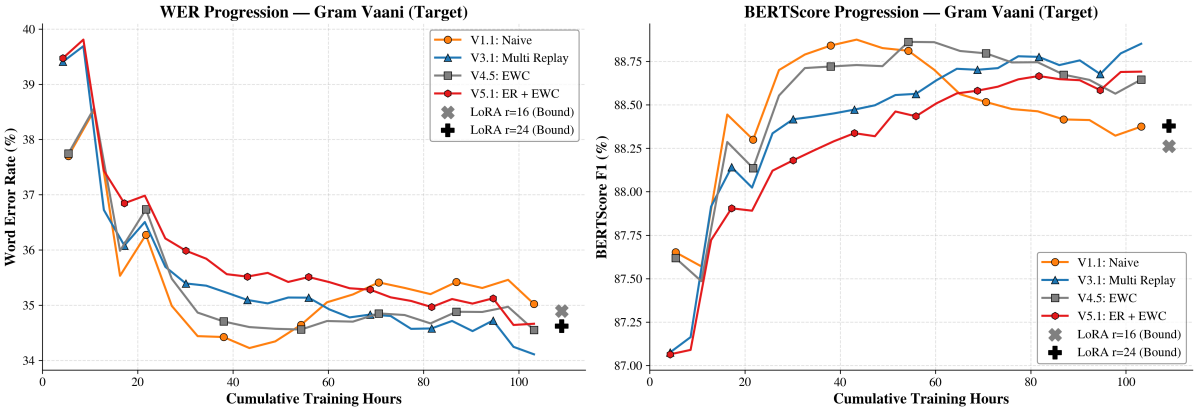


Figure 1: Progression of WER and BERTScore on the target domain (Gram Vaani) over cumulative training hours for the Naive (V1.1), Multi-Domain ER (V3.1), EWC (V4.5), and Hybrid (V5.1) strategies, with offline LoRA upper bounds for reference.

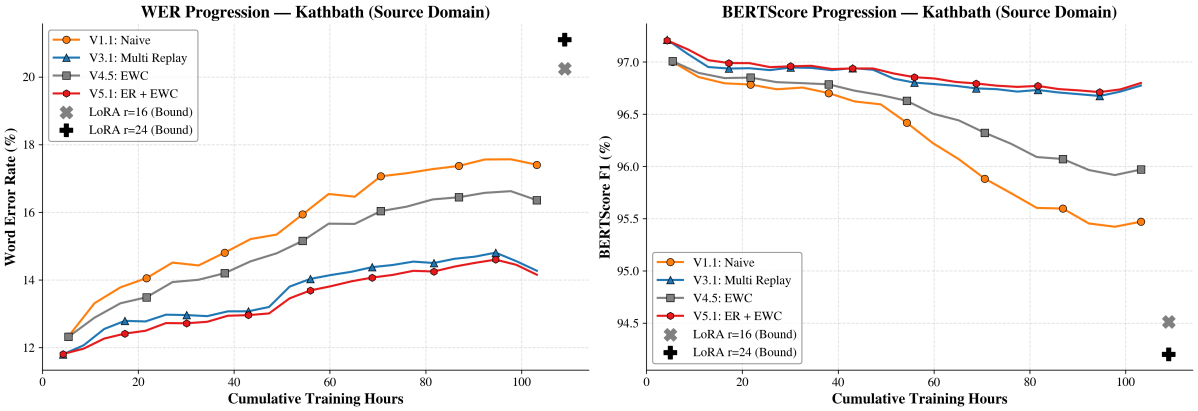


Figure 2: Catastrophic forgetting on the source domain (Kathbath): progression of WER and BERTScore as adaptation proceeds over Gram Vaani, with offline LoRA upper bounds for reference.

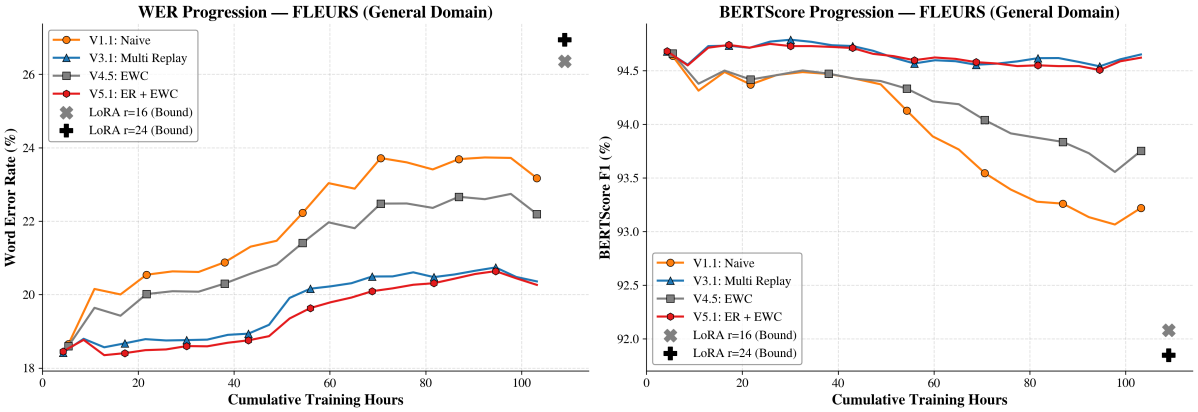


Figure 3: Catastrophic forgetting on the general domain (FLEURS): progression of WER and BERTScore as adaptation proceeds over Gram Vaani, with offline LoRA upper bounds for reference.

substantial forgetting), comparable to prior work on Gram Vaani (30.3% on the Development set with Conformer/TDNN models and external LMs; Patel and Scharenborg, 2022), and the continual strategies approach this performance under streaming constraints. We keep the LM out of the adaptation loop so that it does not mask the acoustic adaptation dynamics we study; our results without an LM are therefore a conservative estimate of achievable WER.

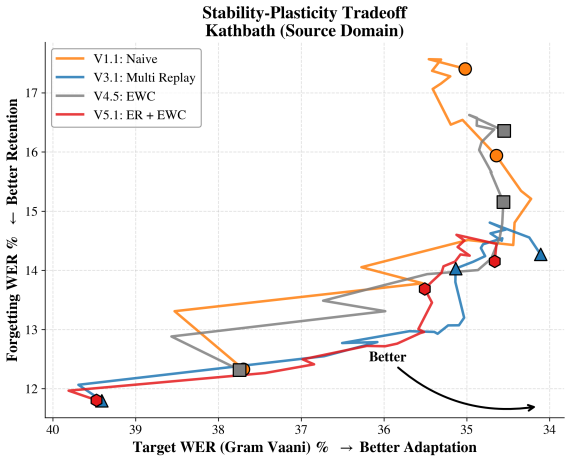


Figure 4: Stability-plasticity trajectories on Kathbath (source domain). Curves trace per-segment means over three seeds; markers denote the first, middle, and final segments.

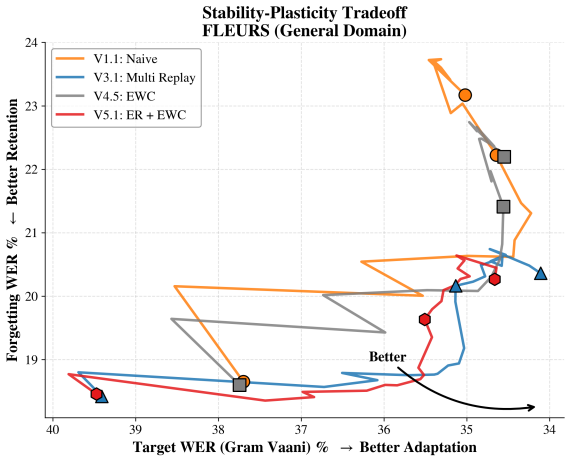


Figure 5: Stability-plasticity trajectories on FLEURS (general domain). Curves trace per-segment means over three seeds; markers denote the first, middle, and final segments.

Model (with LM)	WER (%)	BERTScore (%)
IndicWav2Vec (Baseline)	35.21 ± 0.00	87.94 ± 0.00
Full Finetune	**29.03 ± 0.12**	**90.14 ± 0.04**
LoRA Finetune (r=16)	31.06 ± 0.07	89.45 ± 0.03
V3.1 (Multi-Domain ER)	30.32 ± 0.03	89.91 ± 0.02
V5.1 (ER + EWC)	30.46 ± 0.03	89.83 ± 0.05

Table 2: LM spot check on Gram Vaani: effect of a domain-specific 4-gram LM on different adaptation regimes. Results are mean ± std over 3 seeds. Acoustic adaptation remains essential even with LM assistance.

5.6 Ablation Studies

5.6.1 Sensitivity to EWC Regularization (λ)

With high regularization ($\lambda \geq 10^3$), the EWC constraint dominated the gradients, effectively "freezing" the LoRA parameters and preventing meaningful adaptation to the noisy telephony domain. An intermediate value of $\lambda = 10^2$ allowed adaptation but remained overly conservative. A lower value of $\lambda = 10$ (V4.5) provided the best balance between adaptation and parameter protection, and linear (absolute) Fisher estimation yielded more stable optimization than higher-order approximations.

5.6.2 LR Warmup and Convergence Speed

Compared to a conservative 100-step learning rate warmup, an aggressive 10-step schedule significantly accelerated convergence early in each segment without increasing catastrophic forgetting ($< 0.05\%$ difference across all runs), suggesting that LoRA-based adaptation can safely use high initial learning rates to escape local minima from previous segments.

6 Conclusion

In this work, we investigated the "Reality Gap" that hinders the deployment of state-of-the-art ASR in rural clinical settings, quantifying it at 41.72% WER on a telephony proxy for clinical speech. Framed as an empirical study of adaptation dynamics under data residency constraints rather than a clinical validation, our comparison of four continual learning paradigms against full fine-tuning and offline LoRA upper bounds shows that Multi-Domain Experience Replay provides the best balance of plasticity and stability. While full fine-tuning achieves the strongest target performance, replay-based streaming adaptation reaches target WER comparable to offline LoRA while forgetting substantially less, improving target WER by 18.2% relative and reducing forgetting by 54% over naive adaptation, consistently across three seeds and two general-domain benchmarks. These findings provide a foundation for reliable, self-improving clinical ASR and its extension across Indic languages.

Privacy, Ethics, and Clinical Safety Considerations

The deployment of ASR in frontline healthcare settings introduces unique ethical and safety considerations which our framework explicitly addresses.

Data Governance and Residency: Our primary focus on on-device adaptation is a direct response

to healthcare data residency requirements. By ensuring that raw patient audio never leaves the local environment, we mitigate the risk of centralized data breaches and support compliance with data protection requirements in various jurisdictions (Leroy et al., 2019). We emphasize that data residency alone is not a formal privacy guarantee: we do not study differential privacy, secure aggregation, or defenses against model-based attacks such as membership inference or model inversion.

Clinical Safety and Error Risks: While our framework significantly reduces WER, automated transcription in medical contexts is never risk-free. ASR errors in medically salient keywords (e.g., negations or dosage numbers) can lead to serious diagnostic errors. We advocate for a **Clinician-in-the-Loop** model, where our adaptive system serves as a draft-generation tool that reduces manual workload while maintaining human oversight for final clinical validation. Improved accuracy reduces, but does not eliminate, clinical risk; moreover, our experiments use Gram Vaani as the closest available proxy for clinical telephony, and should not be read as clinical validation.

Algorithmic Bias and Equity: By prioritizing adaptation on rural dialects and low-bandwidth telephony, this work contributes to healthcare equity. Standard ASR models often fail on marginalized populations due to accent mismatch; our framework allows localized clinics to "correct" these biases by adapting specifically to their local community's speech patterns.

Controlled Adaptation and Model Drift: Continual adaptation introduces the risk of unintended model drift. Our use of experience replay and EWC-based regularization helps limit forgetting of general language knowledge while adapting to local conditions, which is particularly important in clinical settings where robustness and consistency are critical.

Limitations

While our framework demonstrates significant potential for clinical deployment, several limitations remain:

1. **Metrics as a Clinical Proxy:** Our evaluation relies on WER and CER, complemented by BERTScore. While BERTScore captures semantic similarity beyond exact token matching, none of these metrics is sensitive to clinically critical distinctions such as negations, numerical magnitude errors, or medical entity substitutions, and none explicitly weights the clinical significance of errors: an error in a patient's name may be less critical than an error in a medication dosage. Since Gram Vaani does not provide named-entity or disease-name annotations, clinically targeted evaluation (e.g., accuracy on disease names) was not possible. Developing medical-concept-aware metrics, and validating utility through deployment and clinician feedback, remain essential future directions.

2. **Proxy Data:** We use Gram Vaani, a corpus of rural healthcare and agricultural helpline calls, as the closest available proxy for clinical telephony. Although it reflects the channel and dialectal conditions of rural telephony, the content of real clinical conversations may differ. Collecting and evaluating on real clinical telephony data requires ethics approval and patient consent frameworks beyond the present scope; it remains a primary direction for future work.

3. **Reliance on Clinician Supervision:** Our continual learning setup assumes a stream of corrected transcripts (e.g., from medical professionals). In under-staffed rural clinics, this supervision may be sparse or delayed, necessitating future work into uncertainty-aware or semi-supervised adaptation.

4. **Acoustic-Only Training:** While we validated the necessity of acoustic adaptation, we did not integrate Language Models (LMs) into the training loop (e.g., via shallow fusion during replay), as doing so would conflate acoustic and linguistic adaptation signals. Capturing the synergy between acoustic and linguistic adaptation is critical for handling specialized medical terminology.

5. **Language, Dialect, and Backbone Scope:** Our experiments focus on rural Hindi dialects using a single acoustic backbone, IndicWav2Vec, chosen for its coverage of Indian languages. Evaluating additional ASR architectures would further establish robustness and remains important future work. Our im-

mediate aim, however, is to extend the study to other Indic languages, including tonal and Dravidian families, as annotated telephonic corpora become available, and for these languages IndicWav2Vec remains the most suitable backbone.

6. **Deployment Hardware:** In this work, "edge device" refers to consumer-grade laptops with entry-level GPUs, not mobile phones or embedded hardware. Reported runtimes (Appendix A.3) are approximate and vary across environments.

Acknowledgments

The authors wish to acknowledge the use of large language models in improving the presentation and grammar of this paper. The paper remains an accurate representation of the authors' underlying contributions.

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*, pages 144–161, Berlin, Heidelberg. Springer-Verlag.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Frederik Benzing. 2022. Unifying importance based regularisation methods for continual learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2372–2396. PMLR.

Anish Bhanushali, Grant Bridgman, Deekshitha G, Prasanta Ghosh, Pratik Kumar, Saurabh Kumar, Adithya Raj Kolladath, Nithya Ravi, Aaditeshwar Seth, Ashish Seth, Abhayjeet Singh, Vrunda Sukhadia, Umesh S, Sathvik Udupa, and Lodagala V. S. V. Durga Prasad. 2022. Gram Vaani ASR challenge on spontaneous telephone speech recordings in regional variations of Hindi. In *Interspeech 2022*, pages 3548–3552.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet Kumar Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019. Continual learning with tiny episodic memories. *CoRR*, abs/1902.10486.

Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, Justin Tansuwan, Nathan Wan, Yonghui Wu, and Xuedong Zhang. 2018. Speech recognition for medical conversations. In *Interspeech 2018*, pages 2972–2976.

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.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2022. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385.

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

Tahir Javed, Kaushal Bhogale, and Mitesh M. Khapra. 2025. NIRANTAR: Continual Learning with New Languages and Domains on Real-world Speech Data. In *Interspeech 2025*, pages 918–922.

Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M. Khapra. 2023. IndicSUPERB: A speech processing universal performance benchmark for Indian languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12942–12950.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Towards building ASR systems for the next billion users. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10813–10821.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau. 2019. Federated learning for keyword spotting. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6341–6345.

Xinyu Li, Venkata Chebiyyam, and Katrin Kirchhoff. 2019. Speech audio super-resolution for speech recognition. In *Interspeech 2019*, pages 3416–3420.

Yin-Tse Lin, Bo-Hao Su, Chi-Han Lin, Shih-Chan Kuo, Jyh-Shing Roger Jang, and Chi-Chun Lee. 2023. Noise-robust bandwidth expansion for 8K speech recordings. In *Interspeech 2023*, pages 5107–5111.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Tanvina Patel and Odette Scharenborg. 2022. Using cross-model learnings for the Gram Vaani ASR challenge 2022. In *Interspeech 2022*, pages 4880–4884.

Mengjie Qian, Siyuan Tang, Rao Ma, Kate M Knill, and Mark J.F. Gales. 2024. Learn and don't forget: Adding a new language to ASR foundation models. In *Proc. Interspeech 2024*, pages 2544–2548.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus H. Maier-Hein, Sébastien Ourselin, Micah J. Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. 2020. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):119.

Samik Sadhu and Hynek Hermansky. 2020. Continual learning in automatic speech recognition. In *Interspeech*, pages 1246–1250.

Izhak Shafran, Nan Du, Linh Tran, Amanda Perry, Lauren Keyes, Mark Knichel, Ashley Domin, Lei Huang, Yu-hui Chen, Gang Li, et al. 2020. The medical scribe: corpus development and model performance analyses. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2036–2044.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR. In *Interspeech 2024*, pages 3934–3938.

Gokul Adethya T and S. Jaya Nirmala. 2025. A study on regularization-based continual learning methods for Indic ASR. *Preprint*, arXiv:2508.06280.

Viet Anh Trinh, Pegah Ghahremani, Brian King, Jasha Droppo, Andreas Stolcke, and Roland Maas. 2022. Reducing Geographic Disparities in Automatic Speech Recognition via Elastic Weight Consolidation. In *Interspeech 2022*, pages 1298–1302.

Muqiao Yang, Ian Lane, and Shinji Watanabe. 2022. Online continual learning of end-to-end speech recognition models. In *Interspeech 2022*, pages 2668–2672.

Ruijie Yang, Yan Zhu, Peiyao Fu, Te Luo, Zhihua Wang, Xian Yang, Quanlin Li, Pinghong Zhou, and Shuo Wang. 2026. Development and multi-center evaluation of domain-adapted speech recognition for human-AI teaming in real-world gastrointestinal endoscopy. *arXiv preprint arXiv:2604.01705*.

Woongchang Yoon. 2026. Optimizing Whisper for Korean telemedicine: Fine-tuning domain-specific ASR for clinical telephone transcription. *DIGITAL HEALTH*, 12:20552076261444914.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net.

Yaoyue Zheng, Yin Zhang, Joost van de Weijer, Gido M van de Ven, Shaoyi Du, Xuetao Zhang, and Zhiqiang Tian. 2026. Revisiting weight regularization for low-rank continual learning. In *International Conference on Learning Representations*.

Parameter	Value
Optimizer	AdamW
Weight Decay	0.01
Warmup Steps	10 (Aggressive)
LoRA Target Modules	query, value (Attention blocks)
Batch Size (Effective)	64
Max Audio Duration	30.0 seconds
Training Epochs	3 per segment
Upper Bound Strategies	
Full Finetune	Base LR $= 1 \times 10^{-4}$, full model fine-tuned on combined target corpus
LoRA (r=16, α=32)	Base LR $= 3 \times 10^{-4}$, LoRA adapters only
LoRA (r=24, α=48)	Base LR $= 3 \times 10^{-4}$, LoRA adapters only
Continual Learning Strategies	
V1.1 Naive	Base LR $= 3 \times 10^{-4}$, $r = 16$, $\alpha = 32$
V2.1 Single-Domain ER	Base LR $= 3 \times 10^{-4}$, $r = 24$, $\alpha = 48$, Replay buffer: 400 target samples
V3.1 Multi-Domain ER	Base LR $= 3 \times 10^{-4}$, $r = 24$, $\alpha = 48$, Replay buffer: 300 target + 300 general
V4.5 EWC	Base LR $= 3 \times 10^{-4}$, $r = 24$, $\alpha = 48$, $\lambda = 10$
V5.1 Hybrid	Base LR $= 3 \times 10^{-4}$, $r = 24$, $\alpha = 48$, $\lambda = 10$, Replay buffer: 300 target + 300 general

Table 3: Comprehensive hyperparameter settings. r and α represent the LoRA rank and scaling factor, respectively. λ denotes the EWC regularization strength.

Algorithm 1: LoRA-based Hybrid Adaptation Loop
Require: Base model θ_{base}, Replay buffer $\mathcal{B}$, Regularization λ
1: Inject LoRA adapters θ_0 into frozen θ_{base}; set importance $F \leftarrow \mathbf{0}$
2: for segment $k = 1, \ldots, K$ **do**
3: Receive clinical stream $\mathcal{D}_k$
4: $\mathcal{D}_{train} \leftarrow \text{Sample}(\mathcal{D}_k) \cup \text{Sample}(\mathcal{B})$
5: **Adaptation Step:**
$\theta_k \leftarrow \text{argmin}_\theta \mathcal{L}_{CTC}(\mathcal{D}_{train}) + \frac{\lambda}{2} \sum F_i(\theta_k - \theta^*_{k-1})^2$
6: **Importance Estimation:**
$F_{new} \leftarrow \frac{1}{|\mathcal{D}_{train}|} \sum_{x \in \mathcal{D}_{train}} |\nabla_\theta \mathcal{L}_{CTC}(x)|$
7: **Asynchronous Consolidation:**
$F \leftarrow \frac{F \times (k-1) + F_{new}}{k}$
8: $\theta^*_k \leftarrow \text{detach}(\theta_k)$ *Checkpointing*
9: $\mathcal{B} \leftarrow \text{UpdateBuffer}(\mathcal{D}_k, \mathcal{L}_{inst})$
10: end for

Algorithm 2: Buffer Management
Require: Segment data $\mathcal{D}_k$, Buffer $\mathcal{B}$, Threshold τ
1: Compute $\mathcal{L}_{inst}(x)$ for all instances $x \in \mathcal{D}_k$
2: Hard Example Mining:
Identify $\mathcal{S}_{hard} \leftarrow \{x \in \mathcal{D}_k \mid \mathcal{L}_{inst} > \tau \bar{\mathcal{L}}\}$
3: Buffer Update:
$\mathcal{B}_{gv} \leftarrow \text{Sample}(60\% \text{ from } \mathcal{S}_{hard} \cup 40\% \text{ Random})$
$\mathcal{B}_{gen} \leftarrow \text{Sample}(300 \text{ Balanced from Kathbath})$
4: $\mathcal{B} \leftarrow \mathcal{B}_{gv} \cup \mathcal{B}_{gen}$

A Experimental Setup and Reproducibility

To ensure the reproducibility of our findings, we provide the full configuration details for all experimental paradigms. All models were trained for **3 epochs per data segment** to ensure local convergence, and all results are averaged over three random seeds. All results, including the baseline, were obtained under a single, consistent set of software package versions.

A.1 Hyperparameter Configurations

Table 3 details the parameters used, utilizing a splittable format to distinguish between universal settings and those tailored to specific strategies. LoRA hyperparameters were selected based on prior literature and preliminary validation experiments; the replay, EWC, and hybrid strategies share an identical LoRA configuration ($r = 24$, $\alpha = 48$) to enable fair comparison.

A.2 Algorithm and Buffer Management

The technical contribution of our framework lies in the efficient integration of LoRA, linearized EWC, and prioritized experience replay. Algorithm 1 below details the localized adaptation loop, and Algorithm 2 describes our multi-domain buffer management strategy.

A.3 Computational Efficiency and Hardware Setup

A core goal of our work is to show that high-performance, on-device adaptation is feasible on consumer-grade laptops. We evaluated our pipeline on the configurations detailed in Table 4. These benchmarks demonstrate the technical feasibility of localized, self-improving ASR systems in environments where high-end compute clusters are unavailable. The reported runtimes are approximate, representative values and vary across software environments.

B Dataset Characteristics

The Gram Vaani dataset serves as a proxy for rural clinical environments due to its telephonic acquisition (originally 8kHz upsampled to 16kHz) and

	Config A (High-End)	**Config B (Mid-Range)**
CPU	Intel i7-13700H (14 cores)	Intel i5-12500H (12 cores)
GPU	NVIDIA RTX 4050 (35W)	NVIDIA RTX 3050 (70W)
Training Time (per segment)	25–30 minutes	50–60 minutes

Table 4: Hardware configurations (consumer-grade laptops) and approximate training time benchmarks for on-device adaptation.

Metric	**Value**
Total Duration	103.2 Hours
Number of Segments (k)	19 (V1.1, V4.5); 24 (V2.1, V3.1, V5.1)
New Samples per Segment	2,000 (V1.1, V4.5); 1,600 (V2.1, V3.1, V5.1)
Avg. Duration per Sample	9.4 s
Acoustic Condition	Noisy
Primary Dialect	Rural Hindi

Table 5: Characteristics of the partitioned Gram Vaani dataset used for continual adaptation.

focus on healthcare and agricultural discussions. Table 5 summarizes the key characteristics. Forgetting is evaluated on the complete Kathbath Hindi validation set (3,151 samples) and the FLEURS Hindi (`hi_in`) test split.

C Detailed Results by Paradigm

This section presents metric progressions for each continual strategy, grouped by paradigm: Naive (V1.1; Figure 6), Experience Replay (V2.1 and V3.1; Figure 7), EWC (V4.5; Figure 8), and Hybrid (V5.1; Figure 9). All metrics (WER, CER, BERTScore) are reported as percentages (%) with mean $\pm$ standard deviation across 3 independent seeds, each run for the full continual adaptation protocol; final-checkpoint results for all strategies are reported in Table 1. Figures show the full progression over training segments; the shaded band denotes ± 1 standard deviation. The EWC strategy (V4.5) shows visibly wider bands, indicating higher sensitivity to the random seed than the replay-based strategies.

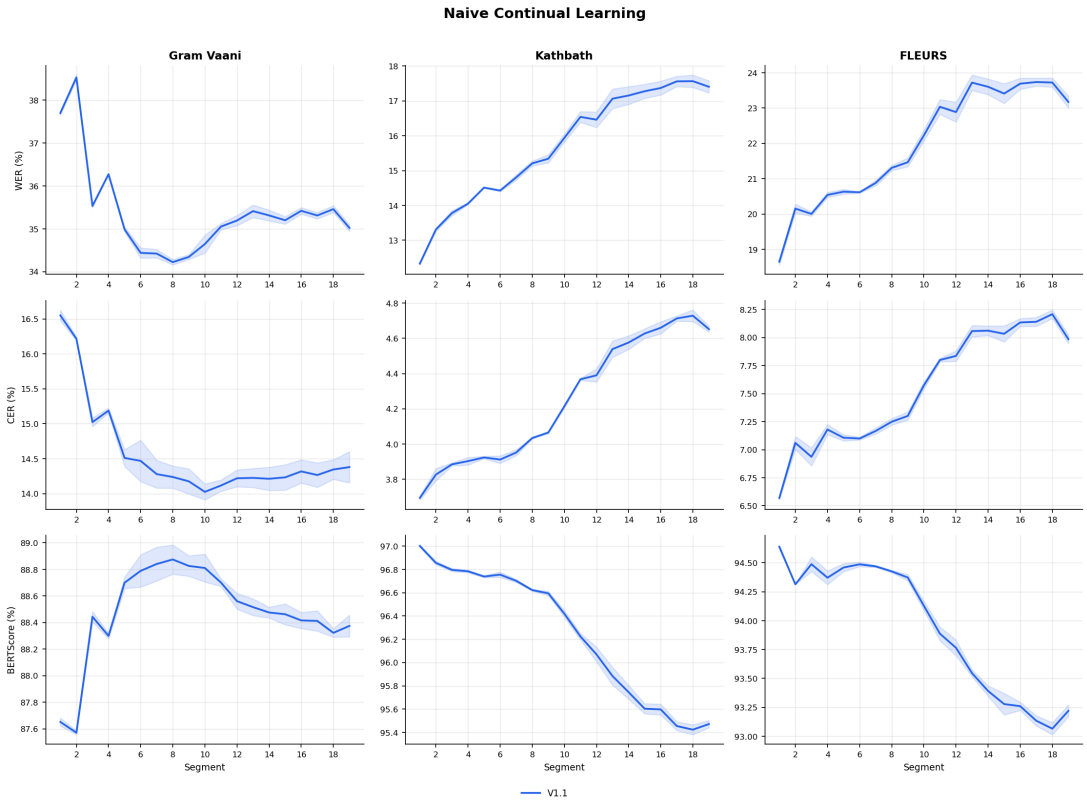


Figure 6: Metric progression (WER, CER, BERTScore) for the Naive strategy (V1.1) across all three evaluation datasets. Shaded bands denote ± 1 standard deviation across seeds.

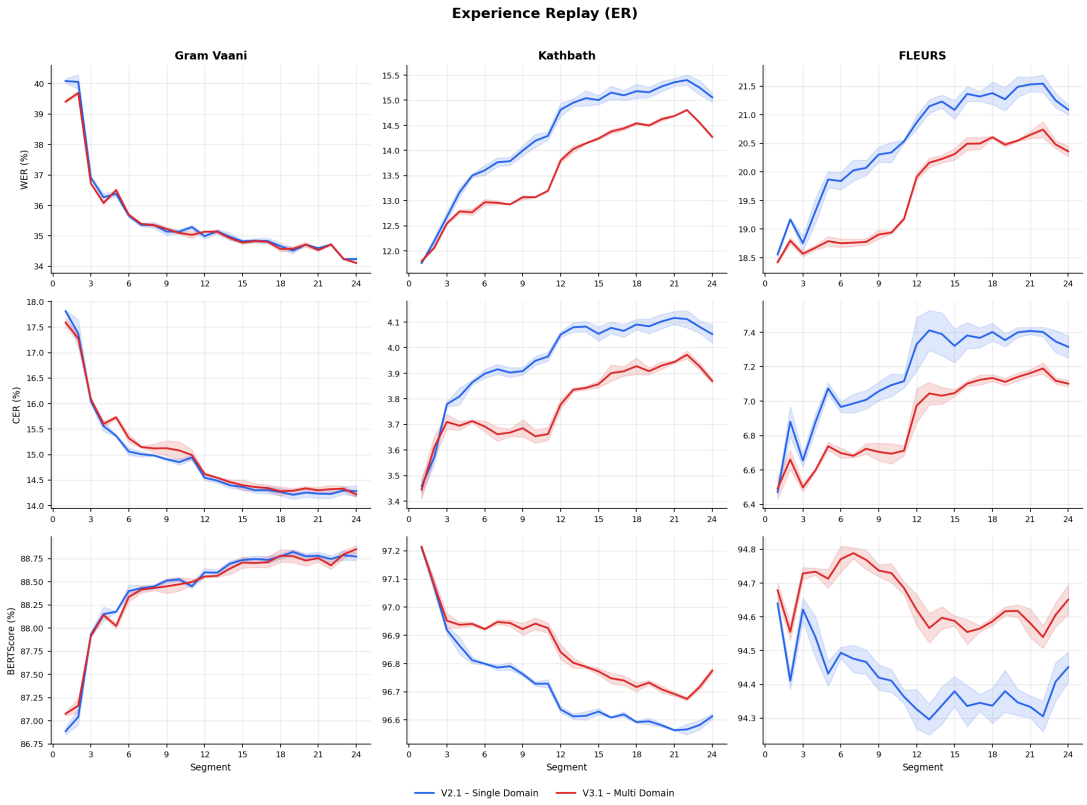


Figure 7: Metric progression for Experience Replay strategies (V2.1 and V3.1) across all evaluation datasets. Shaded bands denote ± 1 standard deviation across seeds.

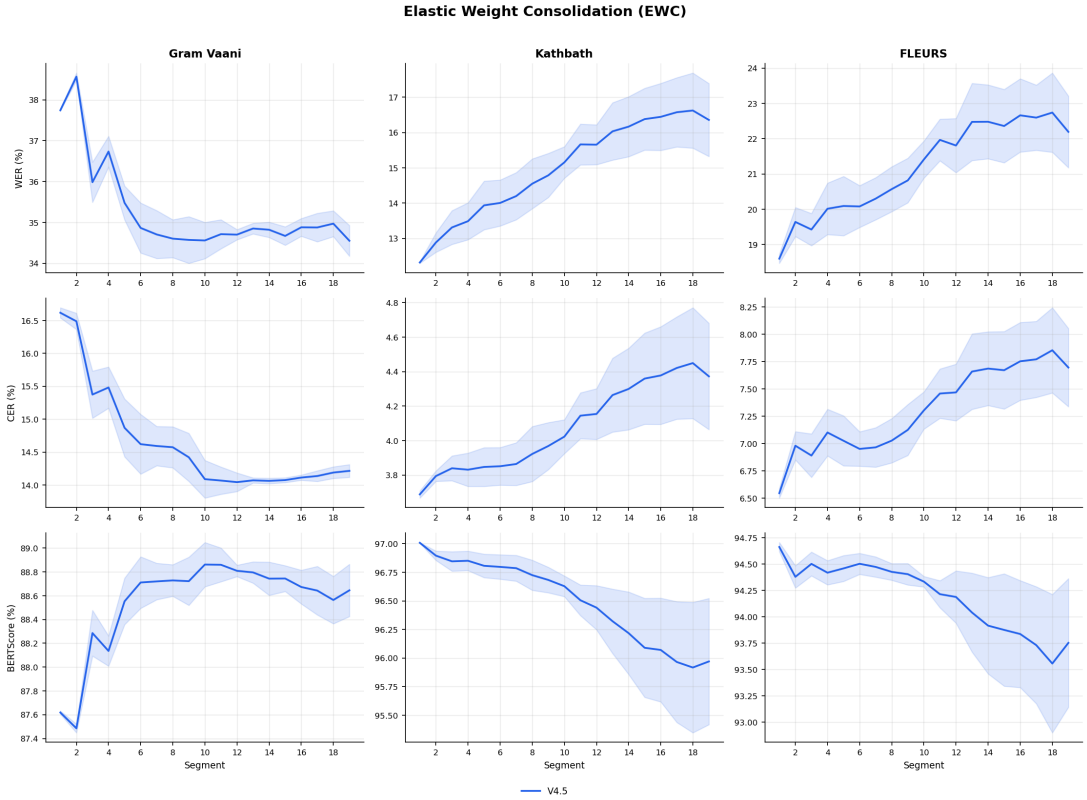


Figure 8: Metric progression for the EWC strategy (V4.5). The larger standard deviation bands indicate higher sensitivity to the random seed than the replay-based strategies.

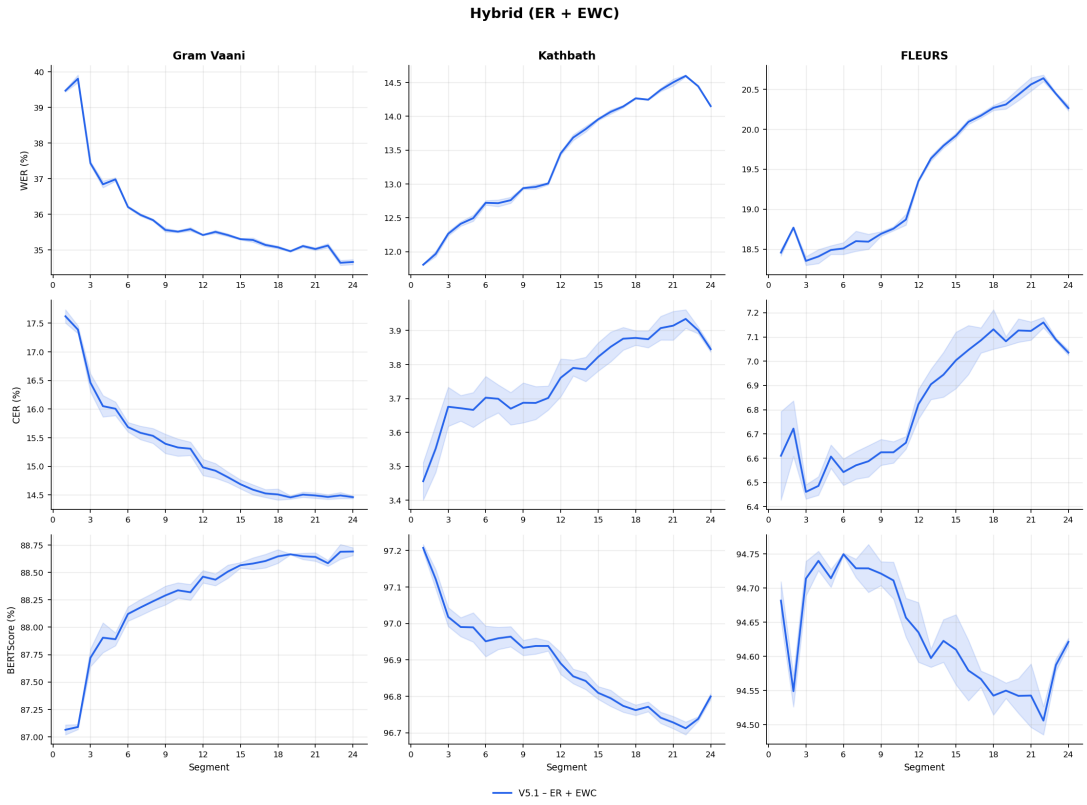


Figure 9: Metric progression for the Hybrid strategy (V5.1) across all evaluation datasets. Shaded bands denote ± 1 standard deviation across seeds.