

Navigating the Reality Gap: Privacy-Preserving Adaptation of ASR for Challenging Low-Resource Domains

Darshil Chauhan¹, Adityasinh Solanki¹, Vansh Patel¹, Kanav Kapoor¹,
Ritvik Jain², Aditya Bansal², Pratik Narang¹, Dhruv Kumar¹

¹BITS Pilani, Pilani Campus, India

²Qure.ai, India

f20230609@pilani.bits-pilani.ac.in

Abstract

Automatic Speech Recognition (ASR) holds immense potential to assist in clinical documentation and patient report generation, particularly in resource-constrained regions. However, deployment is currently hindered by a technical deadlock: a severe “**Reality Gap**” between laboratory performance and noisy, real-world clinical audio, coupled with strict privacy and resource constraints. We quantify this gap, showing that a robust multilingual model (IndicWav2Vec) degrades to a **40.94% WER** on rural clinical data from India, rendering it unusable. To address this, we explore a **zero-data-exfiltration** framework enabling localized, continual adaptation via Low-Rank Adaptation (LoRA). We conduct a rigorous investigative study of continual learning strategies, characterizing the trade-offs between data-driven and parameter-driven stability. Our results demonstrate that multi-domain Experience Replay (ER) yields the primary performance gains, achieving a **17.1% relative improvement** in target WER and reducing catastrophic forgetting by **55%** compared to naive adaptation. Furthermore, we observed that standard Elastic Weight Consolidation (EWC) faced numerical stability challenges when applied to LoRA in noisy environments. Our experiments show that a stabilized, linearized formulation effectively controls gradient magnitudes and enables stable convergence. Finally, we verify via a domain-specific spot check that acoustic adaptation is a fundamental prerequisite for usability which cannot be bypassed 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 domains such as rural healthcare and telemedicine, this promise remains unfulfilled. The “reality gap,” which represents the disparity between clean training corpora and the chaotic, noisy, and privacy-constrained environments of real-world clinics, renders these state-of-the-art models practically unusable.

Our baseline analysis reveals that even robust multilingual models like IndicWav2Vec (Javed et al., 2022b) degrade to a prohibitively high **40.94% Word Error Rate (WER)** when exposed to real-world clinical audio from rural India (Bhanushali et al., 2022). This failure is compounded by a technical deadlock: patient data privacy laws preclude the use of cloud-based adaptation services (Nawaz et al., 2019), while rural infrastructure lacks the high-end compute required for traditional model retraining. This creates a scenario where models cannot improve because data cannot leave the local environment, and localized compute is too constrained for standard fine-tuning.

To break this deadlock, we propose a **privacy-preserving adaptation framework** for localized, stream-based learning. We define “privacy-preserving” strictly as **data residency**: ensuring that raw patient audio never leaves the local device. By leveraging Low-Rank Adaptation (LoRA) (Hu et al., 2022), we enable the model to fine-tune on incoming data streams using a fraction of the trainable parameters. To isolate the effects of acoustic adaptation and ensure feasibility on extreme edge devices (e.g., 35W mobile GPUs), we deliberately exclude external Language Models, allowing us to quantify the upper bound of purely acoustic plasticity in this domain. However, sequential adaptation introduces the risk of *catastrophic forgetting*, where the model loses its foundational linguistic capabilities (McCloskey and Cohen, 1989). To

counteract this, we integrate a **multi-domain experience replay** mechanism (Chaudhry et al., 2019b), interleaving small buffers of general-domain data with the incoming clinical stream to anchor the learning trajectory.

Furthermore, we document the challenges encountered when incorporating parameter-regularization (EWC) into this pipeline. We find that standard quadratic formulations can lead to numerical instability and gradient explosion in noisy clinical environments. We empirically demonstrate that a stabilized, linearized importance estimation (Absolute Fisher) (Hsu et al., 2023) aids robust convergence in LoRA-based continual adaptation. Our results demonstrate that these strategies yield a **17.1% relative reduction in WER** on the target domain, effectively stabilizing the model’s performance in a privacy-constrained setting. The primary research contributions of our work are:

- **Quantification of the Reality Gap:** We establish the 40.94% WER floor for state-of-the-art multilingual models in rural clinical environments, providing a rigorous benchmark for domain adaptation.
- **Privacy-Preserving Adaptation Framework:** We develop a LoRA-based on-device adaptation pipeline that enables continuous learning from data streams while ensuring 100% patient data residency and computational efficiency.
- **Investigation of Architectural Stability:** We diagnose the stability challenges of standard EWC when integrated with LoRA for noisy domain adaptation and empirically validate the benefits of linearized Fisher importance for long-term pipeline health.
- **Empirical Validation on Low-Resource Data:** We demonstrate a 17.1% relative improvement in WER on the Gram Vaani dataset, establishing a viable pathway for self-improving ASR systems in high-impact environments.

2 Related Work

Designing a privacy-preserving, 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., 2022b), scaling the architecture to cover 40 Indian languages.

The Gap: Despite these capabilities, a “usability gap” exists between laboratory benchmarks and real-world deployment. Radford et al. (2023) demonstrate that foundational models suffer severe degradation on low-resource languages, rendering raw transcripts unusable. This is exacerbated in rural healthcare by noisy telephonic audio and diverse regional dialects (Bhanushali et al., 2022). Standard pre-trained models struggle to generalize to these conditions without targeted adaptation. Furthermore, privacy constraints preclude cloud-based services (Nawaz et al., 2019), creating a deadlock where models cannot improve because data cannot leave the local environment.

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, particularly with scarce data. This presents an *efficiency gap*: the high-end GPUs required for full fine-tuning are unavailable on the resource-constrained edge devices of rural hospitals.

To address this, we adopt Low-Rank Adaptation (LoRA) (Hu et al., 2022). LoRA freezes pre-trained weights and injects trainable rank-decomposition matrices, reducing the parameter budget by up to 10,000 \times while maintaining performance parity. Recent multilingual ASR work, such as LoRA-Whisper (Xiao et al., 2024), validates PEFT’s efficacy in minimizing language interference. By constraining optimization to a low-dimensional subspace, LoRA enables local adaptation on modest

hardware, resolving the deployment challenges in privacy-sensitive clinics.

2.3 Continual Learning and the Stability Gap

While PEFT solves efficiency, self-improving systems face the *stability gap* of Continual Learning (CL), specifically *Catastrophic Forgetting*, where new training erodes previously learned capabilities (McCloskey and Cohen, 1989; Parisi et al., 2019). In ASR, naive updates on recent transcripts lead to overfitting on specific speakers or acoustic conditions.

To mitigate this, we employ regularization methods rooted in Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017). While Schwarz et al. (2018) proposed Online EWC, standard implementations require computing the Fisher Information Matrix (FIM) using squared gradients, which can lead to parameter “freezing” or divergence in noisy domains (Chaudhry et al., 2018). To align with our efficiency and stability constraints, we implement *LORA-EWC* (Wang et al., 2023), applying regularization specifically to the low-rank adapters. Departing from the standard quadratic FIM, we utilize a **Linearized Fisher importance** F_i based on the absolute value of the gradients ($|g|$), following the “Absolute Fisher” formulation (Hsu et al., 2023). This modification, combined with a quadratic parameter shift penalty, prevents the rigid locking associated with standard EWC, ensuring the model retains the plasticity needed to adapt to clinical domain shifts without losing foundational knowledge.

Complementing this regularization, we employ Experience Replay (ER), which interleaves a buffer of prior examples with current training data (Chaudhry et al., 2019a). Yang et al. (2022) demonstrated that replay-based methods perform robustly for ASR adaptation. We utilize a *multi-domain experience replay* mechanism, blending “clean” general-domain speech with “noisy” domain-specific samples via a prioritized buffer management strategy (Algorithm A.2). This combined strategy ensures the system specializes in the clinical environment while preserving the linguistic competence required for reliable long-term operation.

3 Methodology

We propose a privacy-preserving, adaptive ASR framework designed to bridge the “reality gap” between general-purpose models and the acoustic

constraints of rural healthcare. Our approach focuses on **Continual Learning (CL)**, enabling a pre-trained model to adapt sequentially to incoming clinical data streams (\mathcal{D}_{stream}) without data leakage or catastrophic forgetting.

3.1 Base Acoustic Backbone

We utilize **IndicWav2Vec** as our acoustic backbone. This model is built on the Wav2Vec 2.0 architecture and pre-trained on a massive corpus of diverse Indian languages. The network consists of a convolutional feature encoder $f(x)$ mapping raw audio to latent representations, followed by a Transformer context network optimized via Connectionist Temporal Classification (CTC) loss:

$$\mathcal{L}_{CTC} = -\log P(y|x) \quad (1)$$

To enable efficient adaptation on edge devices, we freeze the base model and inject trainable Low-Rank Adaptation (LoRA) matrices into the query and value projection layers.

3.2 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 high-resource, standard Hindi samples (sourced from Kathbath) to preserve foundational linguistic competence.
- **Target Domain History (\mathcal{B}_{spec}):** A sliding window of “hard” examples (high loss) and random samples from previous clinical segments, ensuring retention of recent domain-specific adaptations.

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

$$\begin{aligned} \mathcal{L}_{ER} = & \gamma \mathbb{E}_{x \sim \mathcal{D}_{stream}} [\mathcal{L}_{CTC}(x)] \\ & + (1 - \gamma) \mathbb{E}_{x \sim \mathcal{B}} [\mathcal{L}_{CTC}(x)] \end{aligned} \quad (2)$$

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

3.3 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) 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 estimation on small batches, we approximate the Fisher Information using the accumulated **absolute gradients** rather than **squared gradients**, inspired by Memory Aware Synapses (MAS) (Aljundi et al., 2018) and the Absolute Fisher metric (Hsu et al., 2023):

$$F_i = \frac{1}{N} \sum_{j=1}^N |\nabla_{\theta_i} \log P(y_j | x_j)| \quad (3)$$

This linear accumulation prevents outliers from dominating the importance metric. The EWC regularization loss is then applied as a quadratic penalty:

$$\mathcal{L}_{EWC}(\theta) = \frac{\lambda}{2} \sum_i F_i (\theta_i - \theta_i^*)^2 \quad (4)$$

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

3.4 Hybrid Optimization Framework (V5.1)

To leverage the synergistic effects of data-driven grounding and parameter-driven constraints, we propose a **Hybrid ER + EWC** optimization framework (Strategy V5.1). This approach combines the Multi-Domain replay buffer with the EWC regularization penalty.

The total optimization objective minimizes the transcription error over the mixed batch while simultaneously constraining parameter drift:

$$\mathcal{L}_{Total} = \mathcal{L}_{ER}(\mathcal{D}_{stream}, \mathcal{B}) + \mathcal{L}_{EWC}(\theta) \quad (5)$$

By using \mathcal{L}_{ER} to provide the necessary gradients for acoustic adaptation and \mathcal{L}_{EWC} to define a "safe" optimization region, this hybrid mechanism aims to maximize target domain accuracy while minimizing catastrophic forgetting.

4 Experiments and Results

We conducted a comprehensive evaluation to validate the effectiveness of our adaptive pipeline. The experimental design focuses on two key aspects: the ability to adapt to a specific clinical domain (Gram Vaani) and the ability to retain general linguistic knowledge (Kathbath) to prevent catastrophic forgetting.

4.1 Experimental Setup

4.1.1 Metrics

We evaluate performance using two standard metrics: Word Error Rate (WER) and Character Error Rate (CER). WER measures transcription accuracy at the word level, while CER provides a finer-grained analysis of phonetic accuracy, particularly useful for agglutinative languages and dialectal variations.

$$\text{WER} = \frac{S_w + D_w + I_w}{N_w}, \quad \text{CER} = \frac{S_c + D_c + I_c}{N_c}$$

where S , D , and I represent substitutions, deletions, and insertions, and N is the total count in the reference.

4.1.2 Datasets

- **Gram Vaani (Target Domain)** (Bhanushali et al., 2022): This dataset consists of rural telephonic speech (originally 8kHz, upsampled to 16kHz) and serves as a proxy for the challenging, domain-specific audio encountered in rural hospitals. The content includes medical and agricultural discussions, making it highly relevant for simulating real-world deployment in our target sectors. To strictly simulate a continual learning scenario, we partition the 103 hours of training data into sequential segments, processing them one by one to mimic a live data stream.
- **Kathbath (General Domain)**: A high-quality, read speech dataset representing standard Hindi (Javed et al., 2022a). We utilize a subset of the training set (approx. 25,800 samples) to populate the experience replay buffer, ensuring the model retains knowledge of standard Hindi. The complete validation set (3,151 samples) is used exclusively to measure catastrophic forgetting after adaptation.

4.2 Continual Adaptation Paradigms

To characterize the optimal pathway for localized adaptation, we investigate four distinct continual

learning paradigms. For all experiments, we utilize a base IndicWav2Vec model, which provides a pre-training baseline of **40.94% WER** on the target rural clinical domain and **11.57%** on the general Kathbath domain.

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

This represents the simplest form of adaptation, where the model is fine-tuned sequentially on incoming clinical data streams without explicit regularization. We utilize a LoRA rank of 16 and $\alpha = 32$.

- **Outcome (V1.1):** While the model achieves a significant reduction in target WER (down to 34.00%), it suffers from severe catastrophic forgetting, with general domain error increasing by an absolute **5.93%**.

4.2.2 Paradigm 2: Experience Replay (ER) Strategies

We investigate two ER configurations to stabilize the adaptation trajectory via data rehearsal. We increase the 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.
- **Multi-Domain ER (V3.1):** Replays a balanced buffer of 300 target samples and 300 general-domain (Kathbath) samples. This strategy achieved our lowest target WER of **33.94%** and reduced forgetting by **55%** relative to the naive baseline.

4.2.3 Paradigm 3: Elastic Weight Consolidation (EWC)

We examine whether parameter-regularization can implicitly protect the model’s foundational knowledge without replaying data. We utilize the **Absolute Fisher importance** (F_i) (Hsu et al., 2023) to penalize changes to weights deemed critical for the source domain.

- **Outcome (V4.5):** Initial experiments (V4.1–V4.3) with high λ values caused the model to “freeze,” preventing any adaptation to the noisy clinical domain. By relaxing the constraint ($\lambda = 10$) and utilizing linearized importance, V4.5 achieved a competitive **33.94%**

WER on the target domain, though it remained less efficient at mitigating forgetting than the ER strategies.

4.2.4 Paradigm 4: Hybrid ER + EWC Optimization

Finally, we evaluate the hybrid framework (V5.1) described in Section 3.4, which combines the data-driven guidance of Multi-Domain ER with the parameter-driven Absolute Fisher constraints.

- **Outcome (V5.1):** This paradigm yielded our most stable model, achieving the lowest absolute forgetting rate with a final general WER of **14.14%** (+2.57% increase). However, this stability came at a slight cost to plasticity, with a final target WER of **34.51%**.

5 Analysis and Discussion

5.1 Comparative Performance Summary

Table 1 summarizes the final metrics for the five primary investigative strategies. All proposed paradigms successfully bridge the “Reality Gap,” reducing error on the clinical domain by at least 15% relative.

5.2 Stability-Plasticity Pareto Efficiency

The efficiency of our adaptation strategies is best characterized through a Pareto analysis of retention versus adaptation (Figure 3). We observe that with the current amount of continual learning, multiple successful strategies converge to a range around **34–35% WER**. Below this threshold, further reductions in target WER begin to incur **disproportionately high stability costs**. For instance, while Multi-Domain ER (V3.1) successfully navigates this trade-off, attempts to achieve deeper adaptation in the Hybrid model (V5.1) encountered a stability bottleneck. 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 but has entered a logarithmic long-tail learning phase, 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.

Paradigm	Strategy	Final Target WER	Improvement (%)	Final General WER	Forgetting
Baseline	Pre-trained IndicWav2Vec	40.94%	-	11.57%	-
Naive	V1.1 Naive Fine-tuning	34.00%	+17.0%	17.50%	+5.93%
Experience Replay	V2.1 Single-Domain ER	33.98%	+17.0%	14.90%	+3.33%
Experience Replay	V3.1 Multi-Domain ER	33.94%	+17.1%	14.23%	+2.66%
EWC	V4.5 EWC ($\lambda = 1e1$)	33.94%	+17.1%	15.15%	+3.58%
Hybrid	V5.1 ER + EWC	34.51%	+15.7%	14.14%	+2.57%

Table 1: Comprehensive performance comparison of the investigated continual learning paradigms. Improvement is relative to the IndicWav2Vec baseline on rural clinical data. Forgetting is the absolute increase in WER on the general Kathbath domain.

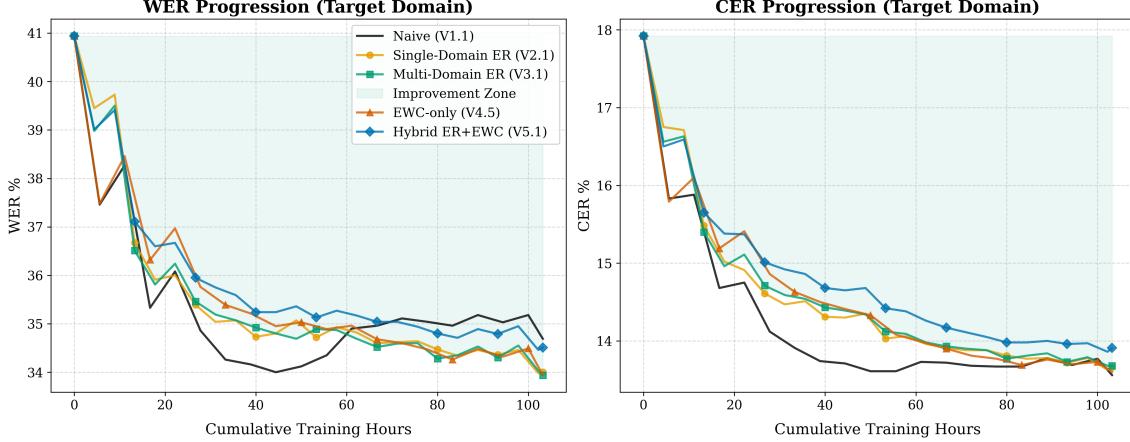


Figure 1: Progression of recognition performance on the target rural clinical data over 100 cumulative training hours. All investigated paradigms successfully bridge the “Reality Gap,” with Multi-Domain ER (V3.1) achieving the deepest adaptation.

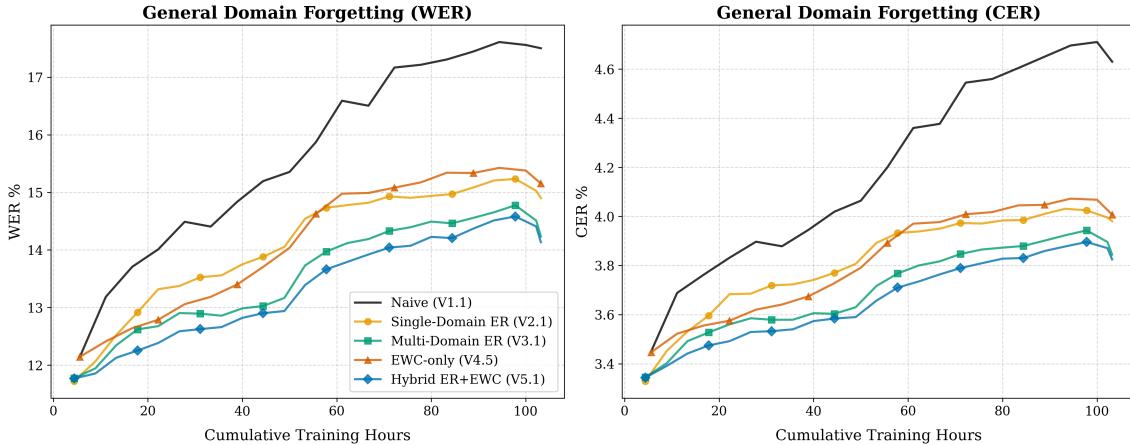


Figure 2: Catastrophic forgetting analysis on the general Kathbath domain. The plot illustrates how knowledge retention is preserved over time. Standard Naive adaptation (V1.1) and EWC-only (V4.5) show significant drift, whereas Multi-Domain ER and Hybrid models successfully flatten the error trajectory.

- **Data Anchors (ER):** Replaying clean source-domain samples (Kathbath) provides an explicit gradient signal that steers the optimization toward a generalized linguistic center. This mechanism proved the most efficient, as evidenced by the V3.1 trajectory on the Pareto

frontier.

- **Parameter Anchors (EWC):** Implicitly constraining weight drift via EWC protects the model but is “acoustically blind.” It penalizes movement relative to the distribution

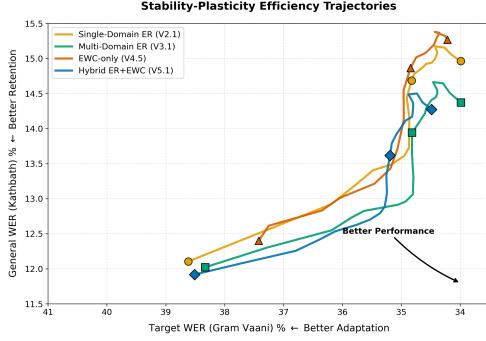


Figure 3: Stability-Plasticity Pareto Analysis: The trajectories illustrate the “cost” of adaptation. Notably, Multi-Domain ER (V3.1) achieves the best overall balance, outperforming other paradigms by maintaining the lowest error rates on both stability and plasticity metrics, whereas Hybrid and EWC approaches incur higher stability costs for similar plasticity gains.

of the source domain importance without providing the model with the actual acoustic features to reconcile. The superiority of data-driven anchors is further underscored by comparing Single-Domain ER (V2.1) with pure EWC (V4.5). Despite achieving similar target-domain performance, the data-driven rehearsal in V2.1 provided superior general-domain stabilization (+3.33% forgetting) compared to the weight-regularization in V4.5 (+3.58%), confirming that explicit data rehearsal is a more effective stabilizer for low-rank adaptation than implicit parameter constraints.

- **The LoRA-Regularization Bottleneck:** We observe a conflict when combining LoRA’s low-rank space with EWC’s parameter-level penalties. In the Hybrid model (V5.1), the cumulative constraints (low rank + EWC penalty) overly restrict the available optimization paths, leading to higher target WER.

5.4 Convergence Dynamics and The “Reality Gap” Floor

A consistent contrast is observed in the convergence behavior across paradigms. While the Naive baseline (V1.1) exhibits erratic, high-variance loss curves, the ER and Hybrid paradigms demonstrate a controlled, monotonic downward trend (Figure 1). This behavior is a critical indicator of long-term viability: it suggests that the system can safely continue to learn from an indefinite data stream without the risk of sudden divergence.

Finally, we observe a consistent convergence around **34% WER** given the current data volume. While this suggests that strong strategies like V3.1 could continue improving with additional streaming data, the progression (Figure 1) indicates diminishing returns. This points to a “soft floor” where further gains become increasingly expensive, likely due to the fundamental “**Reality Gap**”: the acoustic mismatch between 8kHz telephonic recordings and the pre-trained model’s expected 16kHz clarity. This underscores that while localized adaptation can recover significant linguistic performance, overcoming the final acoustic barrier may require architectural enhancements or speech-enhancement front-ends.

5.5 The Necessity of Acoustic Adaptation

To verify that acoustic adaptation is a *prerequisite* for usability, we paired both the baseline and adapted models with a domain-specific 4-gram LM trained on Gram Vaani transcripts (Table 2). The unadapted baseline+LM achieves only **34.96% WER**, while our V3.1+LM achieves **30.26% WER**, a **4.7% absolute improvement**. This confirms that LMs alone cannot bridge the reality gap; **acoustic adaptation is a fundamental prerequisite** for deployment.

Model	WER (%)	CER (%)
Baseline + LM	34.96	17.59
V3.1 Adapted + LM	30.26	15.08

Table 2: LM Spot Check: Acoustic adaptation remains essential even with LM assistance.

5.6 Ablation Studies

5.6.1 Sensitivity to EWC Regularization (λ)

A critical component of our investigation was identifying the optimal regularization strength for EWC. We explored a spectrum of λ values to characterize the trade-offs between model freezing and forgetting. Initial configurations with high regularization ($\lambda \geq 10^3$) resulted in extreme gradient dominance by the EWC constraint, effectively “freezing” the LoRA parameters and preventing any meaningful adaptation to the noisy clinical domain. Conversely, we found that a lower value of $\lambda = 10$ (V4.5) provided the optimal balance, allowing the model to bridge the reality gap while providing sufficient parameter protection.

5.6.2 LR Warmup and Convergence Speed

We evaluated the impact of learning rate warmup schedules on the stability of continual adaptation. By comparing a conservative 100-step warmup against an aggressive 10-step schedule, we observed that the latter significantly accelerated convergence in the early phases of each training segment. Crucially, this aggressive schedule did not result in an increase in catastrophic forgetting ($< 0.05\%$ difference across all runs). This suggests that in the context of LoRA-based ASR adaptation, the model can safely utilize high initial learning rates to rapidly escape local minima from previous segments without overwriting foundational linguistic features.

6 Conclusion

In this work, we investigated the “**Reality Gap**” that hinders the deployment of state-of-the-art ASR in rural clinical settings. By quantifying this disparity at 40.94% WER, we demonstrated the necessity for localized, privacy-preserving adaptation. Through a rigorous comparative study of four continual learning paradigms, we established that a Multi-Domain Experience Replay strategy provides the most efficient balance of plasticity and stability. Our results show a **17.1% relative improvement** in target accuracy and a **55% reduction** in catastrophic forgetting compared to naive baselines. Furthermore, our characterization of the trade-off between low-rank optimization and acoustic mismatch provides a technical roadmap for understanding the interplay between stability and plasticity. These findings establish a viable blueprint for building self-improving ASR systems that remain robust and reliable in high-impact, real-world environments.

7 Limitations and Future Directions

While our framework demonstrates significant potential for deploying ASR in resource-constrained environments, several limitations remain:

1. **Acoustic-Only Training:** Our primary training framework focuses exclusively on adapting the acoustic model (AM). While our validation “spot check” confirmed that even a domain-specific LM is insufficient without acoustic adaptation (Section 5), we did not integrate the LM into the continuous training loop (e.g., via shallow fusion during replay).

Capturing the synergy between acoustic and linguistic adaptation during training remains a key area for future work.

2. **Reliance on Supervision:** Our “continual learning” assumption relies on the availability of a stream of corrected transcripts (e.g., from medical professionals correcting dictations). In scenarios where such feedback is sparse, delayed, or noisy, the adaptation rate would likely degrade.
3. **Static vs. Dynamic Adaptation Scheduling:** Our analysis suggests that an optimal adaptation strategy would ideally employ Experience Replay in the initial phases to rapidly converge to the target domain, followed by a transition to aggressive regularization as the model approaches the “Adaptation Wall” to prevent stability collapse. However, because this wall is inherently dependent on the language, model, and domain, its onset is difficult to predict. Further research into the fundamental limits of adaptation in continual learning, particularly through the integration of language models that provide linguistic context will be critical for refining these strategies.
4. **Language Family Scope:** Our experiments were conducted on Hindi and its rural dialects. While we hypothesize the findings apply to other Indo-Aryan languages, the efficacy of this specific replay strategy on tonal languages or those with fundamentally different acoustic structures (e.g., Dravidian languages) remains to be validated.

Acknowledgement

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 Elhosseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework

- for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.
- 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, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–547.
- Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019a. Efficient lifelong learning with a-gem. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc’Aurelio Ranzato. 2019b. On tiny episodic memories in continual learning. In *Proceedings of the CVPR Workshop on Continual Learning*.
- Ming-Yu Hsu, Yian-Liang Liu, Tsung-Yi Lin, and Shao-Lun Chen. 2023. A case for the absolute fisher in continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- 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 *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Tahir Javed, Kaushal Santosh Bhogale, Abhigyan Raman, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022a. Indicsuperb: A speech processing universal performance benchmark for indian languages. *arXiv preprint*.
- Tahir Javed, Sumanth Ramaseswaran, Padmanabhan Abishek, Sunita Sarawagi, and Mitesh M Khapra. 2022b. Indicwav2vec: Multilingual self-supervised pre-training for indian languages. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.
- Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 24:109–165.
- A. Nawaz, T. Oliveira, and J. Levine. 2019. Privacy-preserving asr: Challenges and opportunities in healthcare. In *Proceedings of the Workshop on Privacy in Machine Learning*.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. In *Neural Networks*, volume 113, pages 54–71.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.
- Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Raia Hadsell, Razvan Pascanu, and Yee Whye Teh. 2018. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pages 4528–4537. PMLR.
- Jiannan Wang, Wangchunshu Zhou, Yuchen Jiang, and Weinan Zhang. 2023. Language models meet world models: Embodied experiences enhance language models. In *Advances in Neural Information Processing Systems (NeurIPS)*. Introduces the EWC-LoRA update rule for efficient continual learning.
- Qian Xiao, Jinyu Li, Yifan Zhao, and Yifan Gong. 2024. Lora-whisper: Parameter-efficient and extensible multilingual asr. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Fu-Cheng Yang, Hsuan-Jui Lo, Szu-Wei Fu, and Yu Tsao. 2022. Online continual learning of end-to-end speech recognition models. In *Proceedings of Interspeech*.

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.

A.1 Hyperparameter Configurations

Table 3 details the parameters used across our investigation. We utilize a split-table format to distinguish between universal settings and those tailored to specific continual learning strategies.

Parameter	Value
Optimizer	AdamW
Base Learning Rate	3×10^{-4}
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
Strategy-Specific Settings	
V1.1 Naive	$r = 16, \alpha = 32$
V2.1 Single-Domain ER	$r = 24, \alpha = 48$, Buffer: 400 Target
V3.1 Multi-Domain ER	$r = 24, \alpha = 48$, Buffer: 300 Target + 300 General
V4.5 EWC	$r = 24, \alpha = 48, \lambda = 10$
V5.1 Hybrid	$r = 24, \alpha = 48, \lambda = 100$, 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.

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 A.2 details the localized adaptation loop, and Algorithm A.2 describes our multi-domain buffer management strategy.

Algorithm 1: LoRA-based Hybrid Adaptation Loop

```

Require: Base model  $\theta_{base}$ , Replay buffer  $\mathcal{B}$ , Regularization  $\lambda$ 
1: Initialize LoRA adapters  $\theta_0 \subset \theta_{base}$  and importance  $F \leftarrow \mathbf{0}$ 
2: for segment  $k = 1, \dots, K$  do
    3: Receive clinical stream  $\mathcal{D}_k$ 
    4:  $\mathcal{D}_{train} \leftarrow \text{Sample}(\mathcal{D}_k) \cup \text{Sample}(\mathcal{B})$  Balanced Mix
    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 (Diagonal Approx.):
         $F_{new} \leftarrow \frac{1}{N} \sum_{x \in \mathcal{D}_{train}} |\nabla_{\theta} \mathcal{L}_{CTC}(x)|$  Absolute Fisher
    7: Asynchronous Consolidation:
         $F \leftarrow \frac{F \times (k-1) + F_{new}}{k}$  Segment-wise Average Checkpointing
    8:  $\theta_k^* \leftarrow \text{detach}(\theta_k)$ 
    9:  $\mathcal{B} \leftarrow \text{UpdateBuffer}(\mathcal{D}_k, \mathcal{L}_{inst})$ 
10: end for

```

Algorithm 2: Multi-Domain Buffer Management (Prioritization)

```

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

```

A.3 Computational Efficiency and Hardware Setup

A core goal of our work is to prove that high-performance, privacy-preserving adaptation is possible on standard mobile workstations rather than centralized server farms. We evaluated our pipeline on two laptop configurations:

1. Intel i7-13700H (14 cores) CPU + NVIDIA RTX 4050 (35W) GPU. Training time averaged **25–30 minutes** per data segment.
2. Intel i5-12500H (12 cores) CPU + NVIDIA RTX 3050 (70W) GPU. Training time was more variable but stayed within **50–60 minutes** per segment.

These benchmarks demonstrate the technical feasibility of localized, self-improving ASR systems in environments where high-end compute clusters are unavailable.

B Detailed Experimental Results

This appendix provides the detailed training progression for all experimental configurations. Each plot summarizes the recognition performance (WER/CER) on the target Gram Vaani domain alongside the CTC loss convergence trend.

B.1 Adaptation Dynamics: Naive and Single-Domain Replay

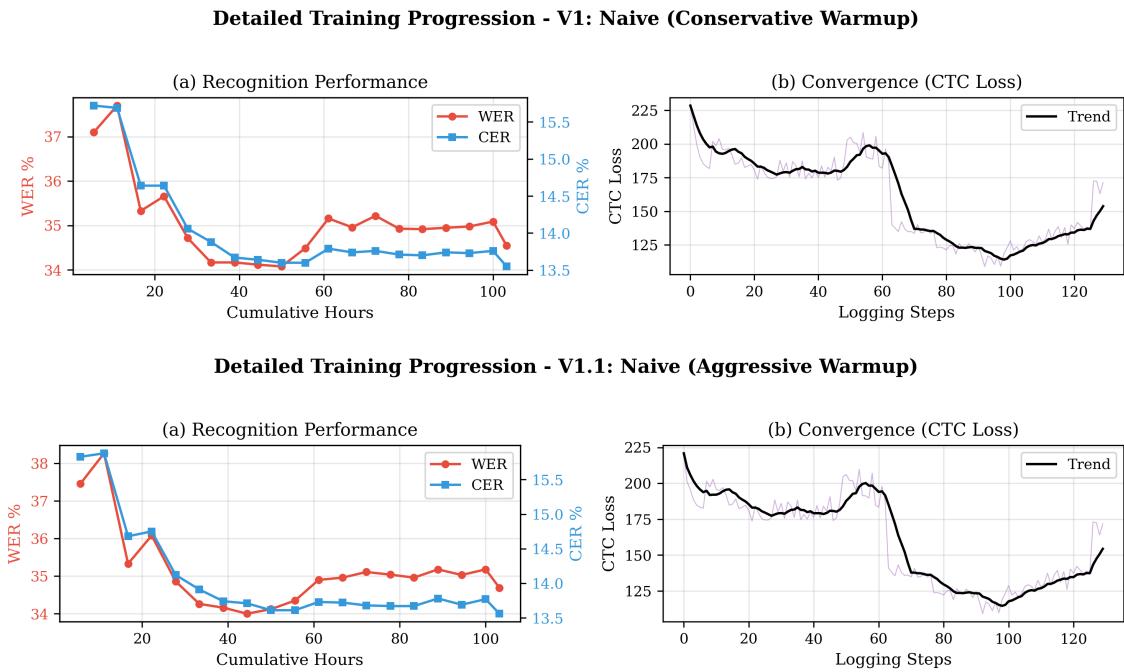
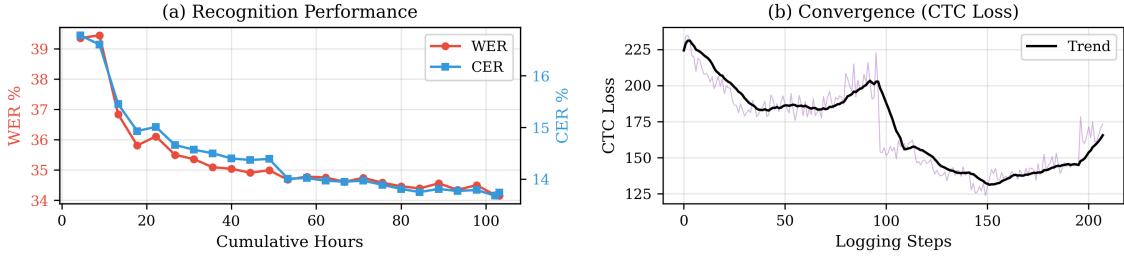


Figure 4: Detailed adaptation dynamics for Naive (V1, V1.1) paradigms.

Detailed Training Progression - V2: Single-Domain ER (Conservative)



Detailed Training Progression - V2.1: Single-Domain ER (Aggressive)

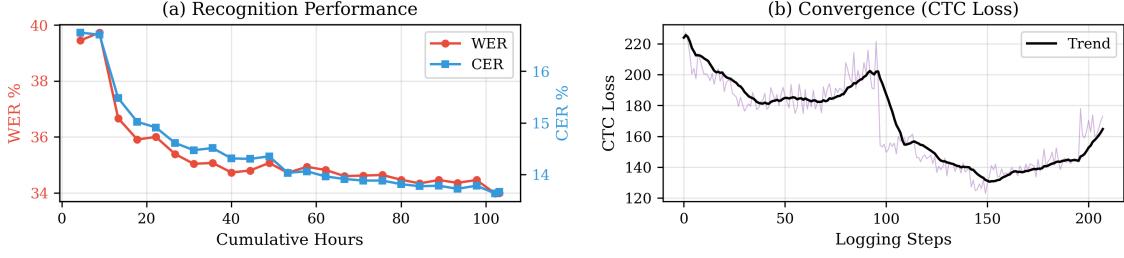
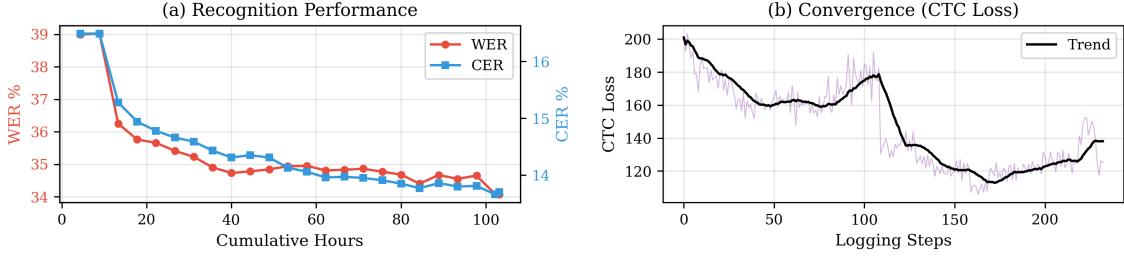


Figure 5: Detailed adaptation dynamics for Single-Domain Replay (V2, V2.1) paradigms.

B.2 Adaptation Dynamics: Multi-Domain Replay

Detailed Training Progression - V3: Multi-Domain ER (Conservative)



Detailed Training Progression - V3.1: Multi-Domain ER (Aggressive)

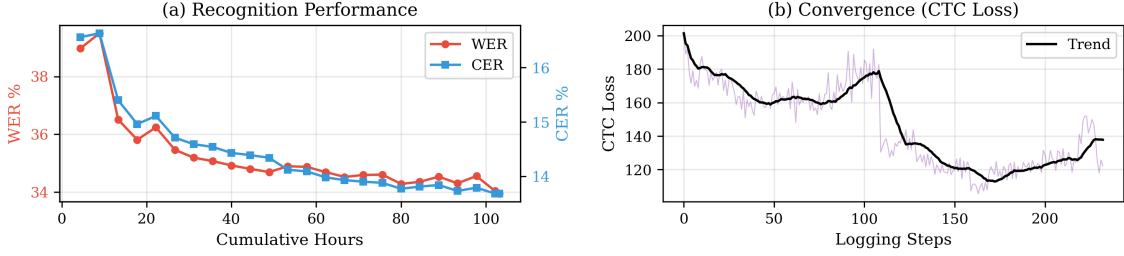
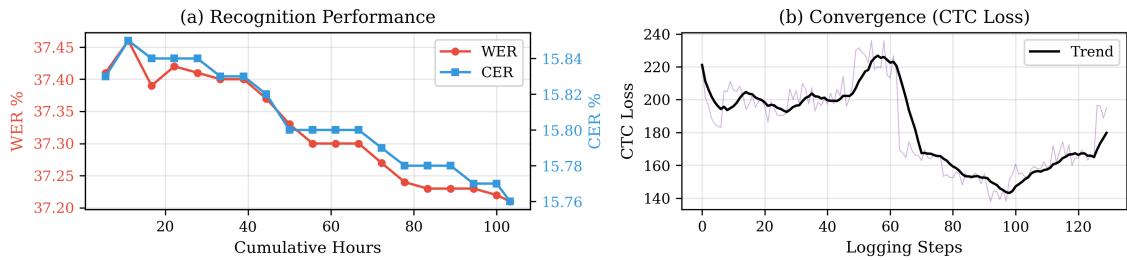


Figure 6: Detailed adaptation dynamics for Multi-Domain Replay (V3, V3.1) paradigms.

B.3 Adaptation Dynamics: EWC and Hybrid Strategies

Detailed Training Progression - V4.3: EWC ($\lambda=1e3$)



Detailed Training Progression - V4.4: EWC ($\lambda=1e2$)

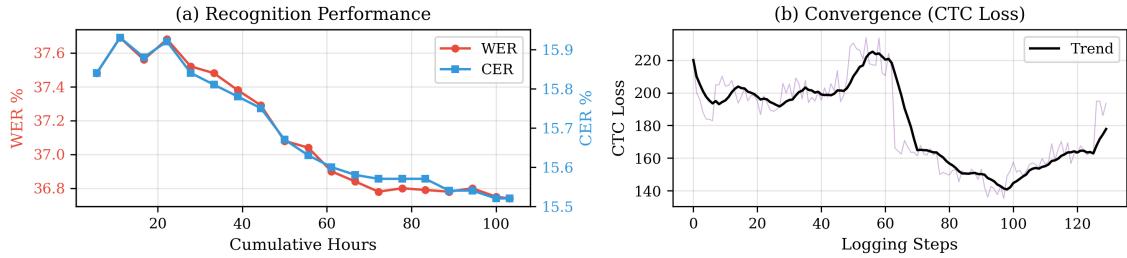
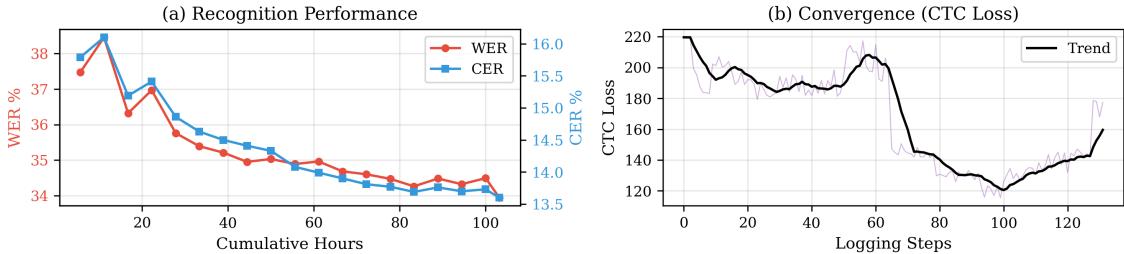
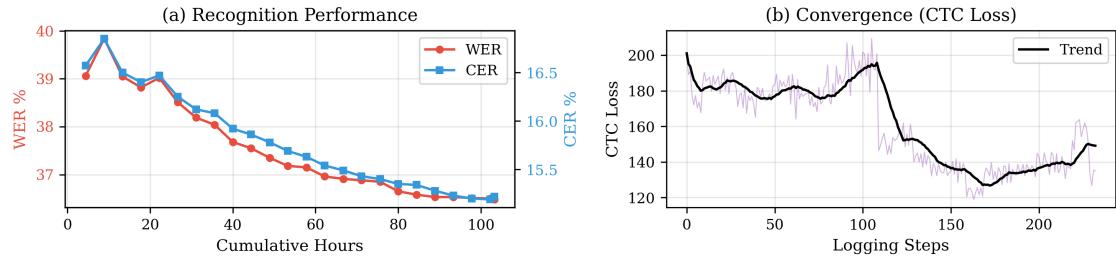


Figure 7: Detailed adaptation dynamics for EWC strategies with high/medium regularization (V4.3, V4.4).

Detailed Training Progression - V4.5: EWC ($\lambda=1e1$)



Detailed Training Progression - V5: Hybrid ER+EWC ($\lambda=1e2$)



Detailed Training Progression - V5.1: Hybrid ER+EWC ($\lambda=1e1$)

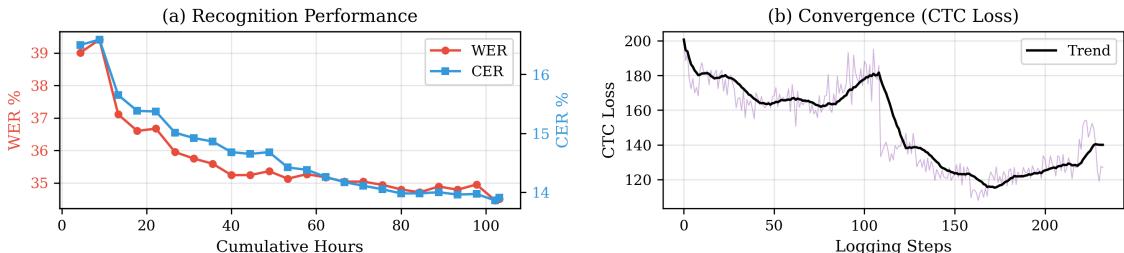


Figure 8: Detailed adaptation dynamics for optimal EWC (V4.5) and Hybrid ER+EWC strategies (V5, V5.1).

B.4 Catastrophic Forgetting Analysis

The preservation of foundational general-domain knowledge (Kathbath) is evaluated across all experimental paradigms. Figures 9, 10, 11, and 12 illustrate the performance stability on the source domain during continuous adaptation.

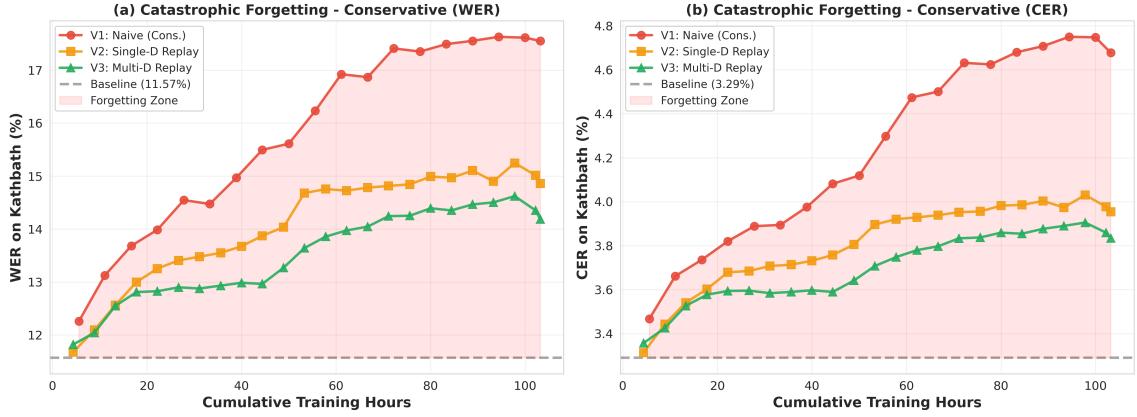


Figure 9: Catastrophic forgetting analysis for Conservative strategies (V1, V2, V3).

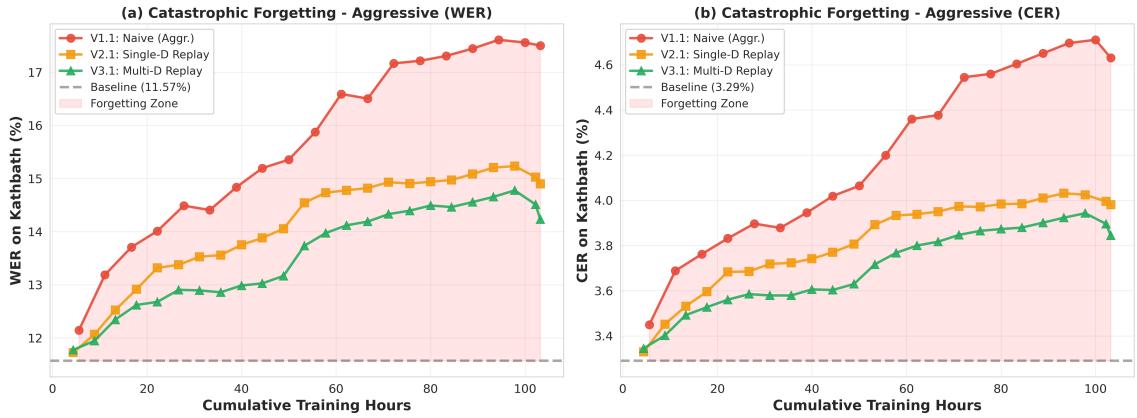


Figure 10: Catastrophic forgetting analysis for Aggressive Warmup strategies (V1.1, V2.1, V3.1).

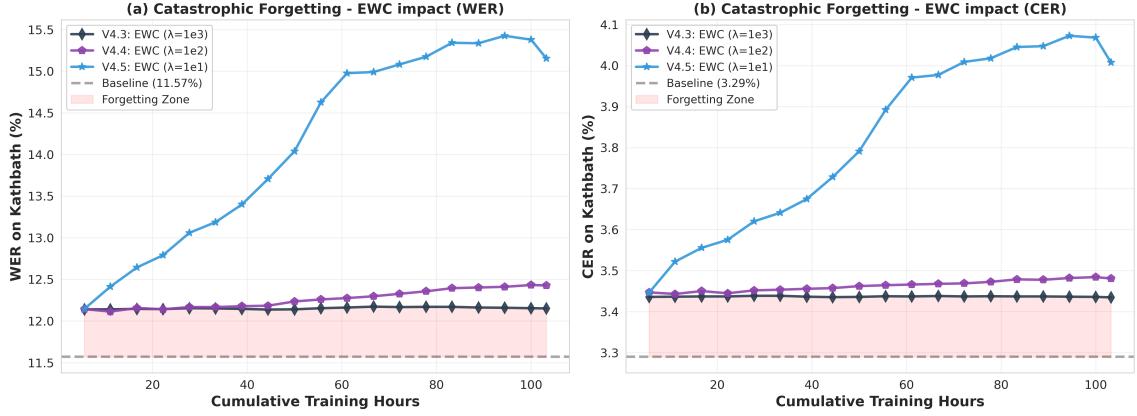


Figure 11: Impact of EWC regularization (λ) on Catastrophic Forgetting (V4.3, V4.4, V4.5).

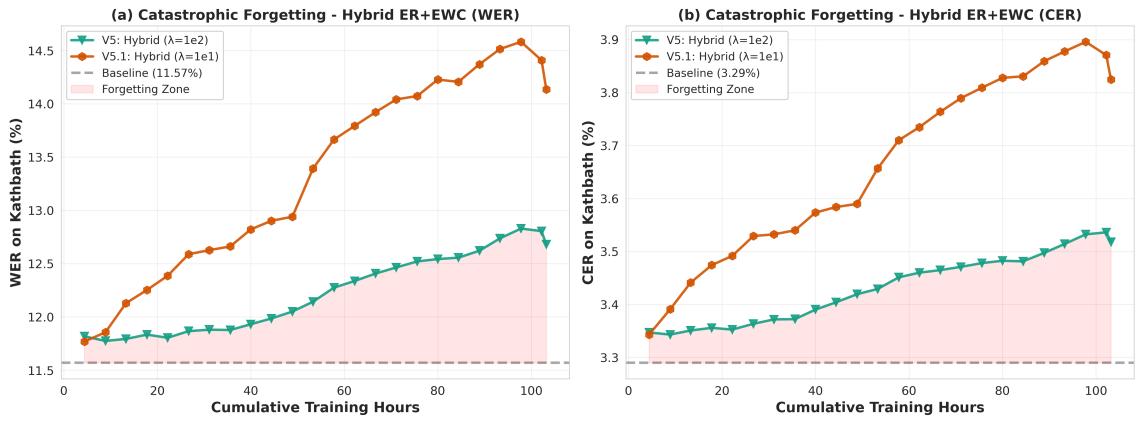


Figure 12: Catastrophic forgetting analysis for Hybrid ER+EWC strategies (V5, V5.1).

B.5 Numerical Stability and Gradient Analysis

A critical challenge encountered during the development of the EWC-based paradigms (V4 series) was the phenomenon of *gradient explosion*. In early iterations of the V4.1 paradigm, gradients used for Fisher Information consolidation were not properly detached from the computation graph. This caused the regularization term to recursively accumulate gradient histories across segments, leading to a computational overhead and gradient norms exceeding 1000. Even after fixing the detaching logic, the standard quadratic Fisher importance in V4.1 remained highly sensitive to noisy data, frequently dominating the optimization objective. This numerical instability caused the model parameters to “freeze,” preventing the system from learning the target domain features.

As shown in Figure 13, our proposed Linearized EWC (L-EWC) strategy successfully stabilizes the gradient norm throughout the 24 adaptation segments. By utilizing absolute gradient values for importance estimation, we ensure that the optimization trajectory remains healthy, with gradient magnitudes comparable to the naive baseline while still enforcing the necessary stability constraints to prevent forgetting.

C Dataset Characteristics

The Gram Vaani dataset serves as a rigorous proxy for rural clinical environments due to its telephonic acquisition (originally 8kHz upsampled to 16kHz) and focus on medical/agricultural discussions. Table 4 provides a breakdown of the dataset characteristics.

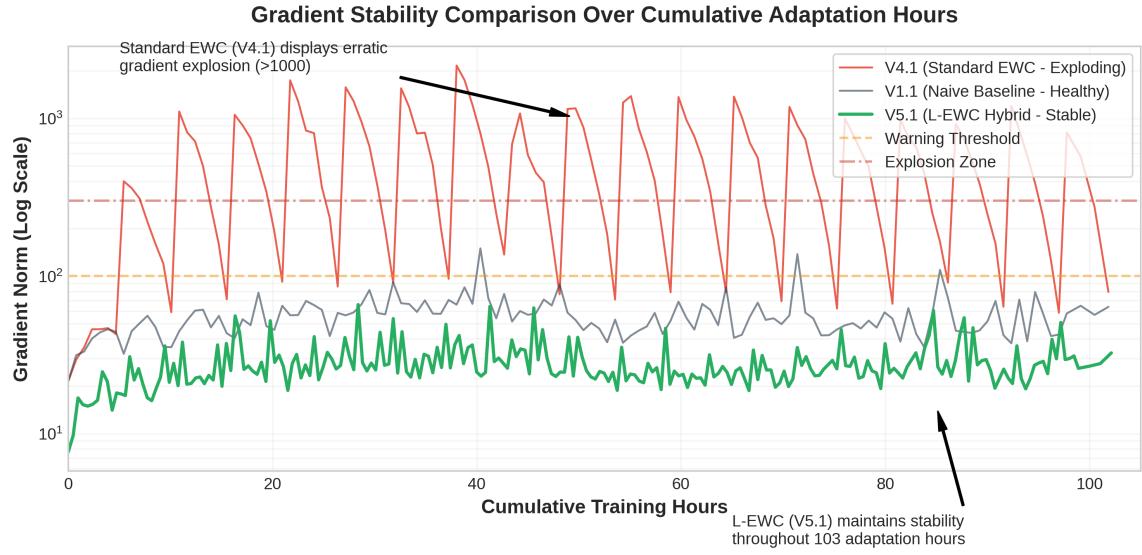


Figure 13: Gradient stability analysis comparing Standard EWC (V4.1), Naive Baseline (V1.1), and our proposed Hybrid L-EWC (V5.1). The log-scale plot demonstrates how the linearized importance estimation in L-EWC prevents the gradient explosion seen in standard quadratic formulations.

Table 4: Characteristics of the partitioned Gram Vaani dataset used for continual adaptation.

Metric	Value
Total Duration	103.2 Hours
Number of Segments (k)	24
Samples per Segment	$\approx 1,600$
Avg. Duration per Sample	9.4 Seconds
Acoustic Condition	Noisy Telephonic (8kHz up)
Primary Dialect	Rural Hindi (various)