Extended Abstract

Motivation Automatic Speech Recognition (ASR) systems traditionally rely on supervised learning using maximum likelihood training objectives, which often do not align directly with evaluation metrics like Character Error Rate (CER) or Word Error Rate (WER). Reinforcement Learning (RL) presents a promising direction to address this gap by enabling direct optimization of evaluation metrics. However, applying RL to ASR is complex due to the sequential nature of speech data, and reward sparsity. We investigate the use of Behavior Cloning (BC) and policy gradient-based RL methods to improve ASR performance, specifically focusing on REINFORCE, Actor-Critic, and a novel Group Relative Policy Optimization (GRPO) algorithm adapted to ASR with a custom reward function.

Method We begin with a BiLSTM-based ASR model trained with CTC loss on the LibriSpeech dataset. REINFORCE samples transcriptions using temperature-controlled multinomial sampling and receives a reward based on a weighted combination of CER and WER. Actor-Critic combines policy learning (actor) with a learned value function (critic) to reduce the variance of the policy gradient using an estimated advantage. Behavior Cloning imitates a teacher model (Whisper-tiny) to mimic the expert transcriptions and is evaluated using CER and WER. GRPO computes group-relative rewards within a batch of transcriptions and penalizes KL divergence from a frozen reference model. The reward is a weighted combination of CER, WER and a penalty for length deviation from ground truth.

Implementation All methods fine-tune a BiLSTM-based CTC ASR model trained on the LibriSpeech corpus. REINFORCE and Actor-Critic undergo Bayesian hyperparameter optimization. Behavior cloning and GRPO use Whisper preprocessing to ensure consistency in mel-spectrogram features. GRPO uses an 8-audio transcription group with KL regularization against the pretrained behavior cloning reference model. Both Behavior Cloning and GRPO required high algorithm runtime and computational costs. Initial GRPO experiments used 5K samples; full-dataset training was completed later.

Results Supervised baseline achieved 22.00% CER and 53.13% WER on LibriSpeech. REINFORCE achieved CER of 20.22% and WER of 49.10%. Actor-Critic also improved performance with 20.98% CER and 51.75% WER. Behavior Cloning achieved 21.55% CER and 55.49% WER. GRPO achieved the best results, reaching 17.87% CER and 48.62% WER after full-dataset training. Out-of-distribution evaluation on FLEURS showed performance degradation for all models, but both REINFORCE and Actor-Critic still outperformed the baseline model performance and showcased improved generalizability.

Discussion From our results we observe that direct metric optimization using reinforcement learning outperforms sequence likelihood optimization approaches in ASR. Temperature control is the most influential hyperparameter, highlighting the importance of sampling strategies in RL training. BC was limited by architectural differences between Whisper and BiLSTM, which hindered effective knowledge transfer. GRPO's group-relative rewards and KL regularization proved effective for improved performance. Both GRPO and Behavior Cloning experienced severe performance degradation on FLEURS potentially due to overfitting to Whisper's clean speech predictions in case of BC, while GRPO's degradation reflects its optimization being tuned specifically to LibriSpeech's acoustic distribution.

Conclusion Our findings suggest that RL-based fine-tuning, especially GRPO and REINFORCE, offers significant performance improvements over traditional supervised or imitation learning. Directly optimizing recognition metrics enhances both in-distribution and out-of-distribution robustness. Future work will extend these methods to transformer-based ASR, improved architectures and multilingual settings to validate their broader applicability.

Reinforcement Learning for Automatic Speech Recognition

Ali Sartaz Khan
Department of Computer Science
Stanford University
askhan1@stanford.edu

Prerana Rane
Center for Global and Online Education
Stanford University
prerana3@stanford.edu

Abstract

Automatic speech recognition (ASR) systems traditionally suffer from a fundamental misalignment between training objectives and evaluation metrics, optimizing sequence likelihood rather than the character and word error rates used for performance assessment. This work presents a comprehensive investigation of deep reinforcement learning approaches for ASR using Imitation Learning and ASR fine-tuning, implementing REINFORCE, Actor-Critic and Group Relative Policy Optimization (GRPO) algorithms to directly optimize recognition performance metrics. We fine-tune a CTC-based BiLSTM model on LibriSpeech using policy gradient methods with rewards computed from Character Error Rate (CER) and Word Error Rate (WER). We also implemented Behavior Cloning trained on OpenAI's Whisper model, creating a high-quality student model that serves as both a strong baseline and initialization point for fine-tuning using GRPO. GRPO achieves the strongest performance, reducing CER by 4.13% and WER by 4.51% compared to the supervised baseline. Comprehensive hyperparameter analysis reveals temperature control as the dominant factor in successful RL training, while systematic evaluation demonstrates that performance improvements generalize to out-of-domain data. Our results establish that direct metric optimization through reinforcement learning provides a viable solution to the training-evaluation mismatch in ASR, offering improvements over traditional supervised approaches and opening new directions for metric-aligned speech recognition training.

1 Introduction

Automatic Speech Recognition (ASR) systems have achieved significant progress through deep learning, yet a key limitation remains: a misalignment between training objectives and evaluation metrics. Conventional ASR training relies on maximum likelihood estimation with losses like Connectionist Temporal Classification (CTC), which optimize alignment probabilities rather than directly minimizing Character Error Rate (CER) or Word Error Rate (WER).

Deep Reinforcement Learning (DRL) presents a promising, yet underexplored, alternative by enabling direct optimization of task-specific metrics using policy gradients. Unlike supervised approaches that maximize data likelihood, DRL frames sequence generation as a decision-making process where the model is rewarded based on output accuracy. Although DRL has been successful in tasks like machine translation and summarization, its application to ASR remains limited.

In this work, we conduct a comprehensive study of DRL for ASR with Behavior Cloning and ASR fine-tuning, evaluating three key algorithms: REINFORCE, Actor-Critic, and Group Relative Policy Optimization (GRPO). We treat the ASR model as a policy network generating character sequences, with rewards computed from CER and WER to align training and evaluation.

While REINFORCE has seen limited use in ASR, actor-critic methods remain largely unexplored, potentially due to the complexity of training additional value networks. Similarly, GRPO, originally developed for mathematical reasoning, has not been applied to ASR. We adapt GRPO for ASR by comparing multiple sampled transcriptions to ground truth and introducing a custom reward function to guide learning.

2 Related Work

Graves et al. (Graves et al. (2013)) demonstrated the effectiveness of deep recurrent neural networks for continuous speech recognition, combining Long Short-Term Memory (LSTM) networks with Connectionist Temporal Classification (CTC) for end-to-end training. The authors showed that deep bidirectional LSTM networks could learn complex temporal dependencies in speech signals without requiring pre-segmented training data or explicit pronunciation models. Their approach used CTC to handle the alignment problem between variable-length audio sequences and text transcriptions, enabling direct optimization of the mapping from acoustic features to character sequences. Prabhavalkar et al. (Prabhavalkar et al. (2017)) applied REINFORCE to directly optimize word error rate in attention-based models, demonstrating that policy gradient methods can effectively fine-tune sequence-to-sequence architectures beyond maximum likelihood training. Their approach involves using the attention-based sequence-to-sequence model as a policy network that generates transcription sequences, with WER as the reward signal for policy gradient updates.

Gao et al. (Gao et al. (2021)) addressed the challenge of improving end-to-end speech recognition performance through ensemble knowledge distillation for joint CTC-attention models. The authors proposed a framework where knowledge from multiple pre-trained acoustic teacher models is distilled into a single student model, leveraging the benefits of ensemble learning. Their approach uses the joint CTC-attention framework, which combines the alignment-free Connectionist Temporal Classification (CTC) objective with attention-based sequence-to-sequence learning to improve recognition accuracy. The authors demonstrated the feasibility of applying distillation techniques in speech recognition. Shao et al. (Shao et al. (2024)) introduced Group Relative Policy Optimization (GRPO) in their work as an alternative to Proximal Policy Optimization for fine-tuning large language models on mathematical reasoning tasks. GRPO eliminates the computationally expensive value network required by traditional PPO approaches, and uses group-relative advantage estimation where multiple responses are sampled for each input and advantages are computed relative to the group mean reward. The method uses KL divergence regularization to prevent the policy model from drifting from the reference model.

The existing work does not address a comprehensive study of different DRL algorithms to ASR. This work provides the first systematic comparison of multiple DRL algorithms for ASR and introduces the first application of GRPO to speech recognition with a novel multi-component reward function design.

3 Dataset & Metrics

LibriSpeech. We used the LibriSpeech corpus (Panayotov et al., 2015), a large-scale dataset of read English speech with high-quality transcriptions. Audio was converted to 80-dimensional log-mel features (400-point FFT, 160-sample hop, 0–8kHz) and layer-normalized. Text was lowercased and filtered to a 29-token vocabulary (CTC blank, space, apostrophe, a–z). We removed samples exceeding 25 seconds, empty transcripts, and feature extraction errors. Our custom data loader handles per-batch padding and length tracking for accurate CTC and policy gradient computation.

FLEURS. To test out-of-distribution generalization, we used the English subset of the FLEURS benchmark (Conneau et al., 2022), which contains 2,000 utterances with diverse accents, recording setups, and prompt content. This makes FLEURS a challenging benchmark for evaluating robustness of models trained on LibriSpeech.

Metrics. We report Character Error Rate (CER) and Word Error Rate (WER), both based on Levenshtein distance:

$$\text{CER} = \frac{S_c + D_c + I_c}{N_c} \qquad \text{WER} = \frac{S_w + D_w + I_w}{N_w} \tag{1}$$

Here, S, D, and I are substitutions, deletions, and insertions at the character (c) or word (w) level, and N is the number of reference units. CER captures fine-grained errors, while WER reflects overall linguistic accuracy.

4 Method

4.1 Baseline CTC ASR Model

To provide a comparison for our reinforcement learning approach, we implemented a standard CTC-based ASR model using a bidirectional LSTM architecture. The model consists of three stages: convolutional feature extraction, recurrent sequence modeling, and CTC classification. The pipeline begins with two 1D convolutional layers (256 filters, kernel size 3), followed by batch normalization and ReLU activation to capture local temporal patterns and stabilize training. These features are passed to a multi-layer bidirectional LSTM, which models long-range dependencies in both directions for contextual understanding.

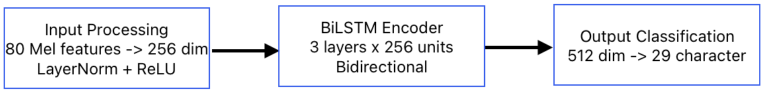

Figure 1: Character-level ASR pipeline using a Bidirectional LSTM

The LSTM output is passed through a two-layer fully connected network that maps hidden states to character-level logits. To aid CTC convergence, we initialize the blank token bias (index 0) to -2.0, promoting non-blank predictions early in training. A final log-softmax layer produces normalized log-probabilities for CTC loss. Variable-length inputs are handled using packed sequences, ensuring efficient batch processing. While more advanced architectures exist, we chose this model due to its strong performance in prior work and its feasibility with our compute constraints.

4.2 REINFORCE

To directly optimize recognition accuracy, we extended the baseline CTC model with a reinforcement learning approach using the REINFORCE algorithm. Unlike CTC, which maximizes alignment likelihood, REINFORCE treats transcription as a policy optimization task where the model learns to maximize rewards tied to evaluation metrics like CER and WER, which bridges the gap between training objectives and downstream performance.

During training, we apply temperature-controlled multinomial sampling to generate multiple transcription candidates per input: one greedy (argmax) and several stochastic samples. For instance, given the reference "HELLO WORLD," sampled hypotheses might include:

- **(1)** `HELLO WORLD`: greedy sample (correct).
- **(2)** `HELO WORLD`: character deletion.
- **(3)** `HELLO WORD`:character substitution.
- **(4)** `HELLO WORLDS`: word-level insertion.

This sampling strategy encourages exploration and allows the model to learn from diverse transcription outcomes. Rewards are computed as a weighted combination of CER and WER:

$$R = -\left(\alpha \cdot \text{CER} + (1 - \alpha) \cdot \text{WER}\right), \quad \hat{R} = R - \bar{R} \tag{2}$$

where α controls the tradeoff between CER and WER, and $\hat{R}$ is the mean-normalized reward used for variance reduction. The negative sign ensures lower error rates yield higher rewards. The policy gradient then updates the model to increase the likelihood of high-reward (accurate) transcriptions, directly aligning learning with evaluation performance.

4.3 Actor-Critic

To reduce the high variance in REINFORCE updates, we adopted an Actor-Critic framework that combines policy optimization with value function estimation. The actor, based on the pretrained BiLSTM ASR model, generates transcriptions via multinomial sampling. A separate multi-layer perceptron serves as the critic, which estimates the expected reward $V(s)$ from a summary of the actor's output logits (mean, max, and first timestep activations). The advantage is computed as:

$$A(s, a) = R(s, a) - V(s) \tag{3}$$

For example, if the actor outputs "HELO WORLD" (CER = 10%, WER = 0%) with reward $R = -5.0$ and the critic estimates $V(s) = -5.2$, the advantage becomes $A = 0.2$, indicating the action performed better than expected.

The actor and critic losses are defined as:

$$L_{actor} = -A(s, a) \cdot \log \pi(a|s) \qquad L_{critic} = (V(s) - R(s, a))^2 \tag{4}$$

The total loss combines both objectives with an entropy regularization term:

$$L_{total} = L_{actor} + \beta \cdot L_{critic} + \gamma \cdot L_{entropy} \tag{5}$$

The entropy term encourages exploration by discouraging overly confident predictions. To further stabilize training, we normalize advantage values within each batch to have zero mean and unit variance, ensuring consistent gradients across varying reward scales.

4.4 Behavior Cloning

We implemented Behavior cloning using OpenAI's Whisper-tiny as the teacher model and a lightweight LSTM model as the student model. This approach is using a knowledge distillation framework to transfer knowledge from the Whisper model to the LSTM model, and addresses the computational challenges of deploying larger scale ASR models in resource limited environments.

Teacher model: Whisper-tiny is a transformer-based encoder-decoder architecture designed for speech recognition. Whisper-tiny contains approximately 39 million parameters and uses a multi-head self-attention mechanism within its encoder-decoder framework. The model processes 80-dimensional log-mel spectrograms computed over 25ms windows. The model is pre-trained on 680,000 hours of multilingual speech data making it suitable for teaching high-quality signals.

Student model: We have a compact LSTM-based model which transforms 80-dimensional mel features to 256-dimensional hidden representations, followed by layer normalization and ReLU activation with 0.1 dropout. The sequence modeling consists of 3-layer bidirectional LSTM with 256 hidden units per direction, producing 512-dimensional output representations (256 x 2 directions). The classification layer maps 512-dimensional LSTM outputs to 29-dimensional character probability distributions, followed by log-softmax for CTC compatibility. Our LSTM model contains approx. 4.2 million parameters which is 9x smaller than Whisper.

Whisper-tiny generates character-level predictions that serve as soft targets. These predictions encode the teacher's decision-making process and provide better supervision than simple hard labels. The algorithm uses CTC loss with teacher predictions as targets. CTC handles temporal alignment between variable-length audio sequences and character outputs without requiring explicit alignment information. The student model is trained using CTC loss to mimic the teacher's character-level predictions. Performance is evaluated using CER and WER.

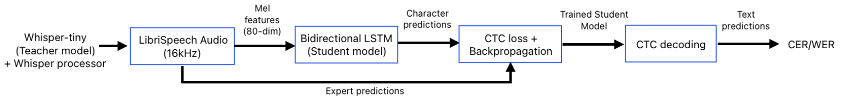

Figure 2: Behavior Cloning

4.5 Group Relative Policy Optimization

Our approach adapts Group Relative Policy Optimization (GRPO) from the DeepSeekMath framework to the automatic speech recognition domain. GRPO eliminates the need for value function estimation by computing rewards relative to group-level performance. GRPO uses groups of transcriptions generated from the same audio input. For each audio sample, we generate multiple diverse transcriptions and compute rewards based on CER and WER. Given an audio input x, we generate a group of G transcriptions $y_1, y_2, ..., y_G$ using the current policy π_θ with temperature sampling. For each transcription y_i in the group, we compute a reward $R_{\text{total}}(y_i, y*)$ where $y*$ is the ground truth transcription. Our reward function uses a weighted combination of CER and WER; and introduces a new reward term: penalty for generation incorrect length outputs.

$$R_{\text{total}} = R_{\text{CER}} + R_{\text{WER}} + R_{\text{length}} \tag{6}$$

where $R_{\text{CER}} = \alpha$ x $max(0, 1 - CER(y_i, y*))$, $R_{\text{WER}} = \beta$ x $max(0, 1 - WER(y_i, y*))$, $R_{\text{length}} = -\gamma|\text{length difference}|$. α, β, γ are different weights used to manipulate the impact of individual rewards.

The group-relative reward for transcription y_i is computed by subtracting the group mean reward from individual rewards. The policy network is initialized using the pretrained behavior cloning model. The policy parameters are updated using the policy gradient with computed group-relative rewards.

$$A_{\text{group}}(y_i) = r(y_i, y*) - (1/G)\Sigma_j r(y_j, y*) \qquad L_{\text{policy}} = -E[\Sigma_i A_{\text{group}}(y_i) log\pi_\theta(y_i|x)] \tag{7}$$

A frozen copy of this pretrained model serves as the reference policy π_{ref} for KL regularization, preventing the policy from deviating too far from the initial behavior during GRPO training. With KL regularization:

$$L_{\text{total}} = L_{\text{policy}} + \lambda_K LD_K L(\pi_\theta || \pi_{\text{ref}}) \tag{8}$$

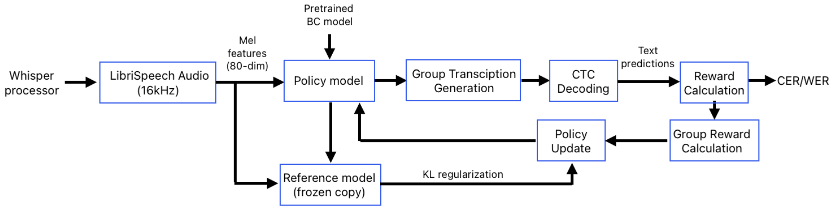

Figure 3: Group Relative Policy Optimization

Existing applications of GRPO require training a separate reward model, our GRPO implementation leverages established ASR evaluation metrics (CER, WER) as direct reward signals, eliminating the need for learned reward functions.

5 Experimental Setup

Data. LibriSpeech contains 960 hrs (28,539 samples) of training data and 5.4hrs of validation data (2,703 samples) from 251 speakers. FLEURS contains 2000 samples of English speech from speakers with various accents that we use to evaluate out-of-distribution performance.

Metrics. As mentioned in Section 3, CER and WER are used to evaluate model performance in speech recognition.

Baseline CTC ASR Model. We trained a 3-layer bidirectional LSTM (256 hidden units/layer, 0.5 dropout) on the LibriSpeech dataset using 80-dimensional log-mel filterbank features. The model was optimized with AdamW (learning rate 1e-3, weight decay 1e-5) and a batch size of 32. Mixed precision training and gradient clipping (max norm 5.0) were used for stability. Training ran for 40 epochs, with ReduceLROnPlateau halving the learning rate after 3 epochs of stagnant validation CER. The best model (lowest CER) was checkpointed. We enabled the zero-infinity option in the CTC loss to handle early-stage alignment instability.

REINFORCE. We fine-tuned the baseline model using REINFORCE with Bayesian sweeps over learning rate 1e-6–1e-3, number of samples 4–10, temperature 0.25–1.0, and WER weight 0.0–1.0 (CER weight as complement). Reward normalization (zero-mean) was applied per batch, with a fallback when variance was too low. Early stopping (patience=3) and ReduceLROnPlateau (factor=0.5, patience=2) helped regulate training. Runs were capped at 50 epochs. AdamW (1e-5 weight decay), mixed precision, and gradient clipping (max norm 1.0) were used.

Actor-Critic. We optimized actor and critic networks separately, tuning seven hyperparameters via Bayesian sweeps: actor LR (1e-6–5e-5), critic LR (1e-6–1e-3), dropout (0.3–0.7), sampling temperature (0.1–1.0), critic loss weight (0.1–1.0), entropy regularization (0.001–0.05), and WER weight. Training initialized from the baseline CTC model. The critic (3-layer MLP with GELU and dropout) was trained first, followed by joint training. The critic used a higher learning rate (2–10×) for faster convergence. AdamW, gradient clipping (1.0), and mixed precision were applied.

Behavior Cloning. To align with Whisper's format, we preprocessed audio using its mel spectrogram pipeline (16 kHz resampling, 25ms windows, 10ms hop, 80 mel bins). All clips were padded/truncated to 30s (480,000 samples). A singleton Whisper processor was used for efficiency. Due to high compute cost, we trained for 20 epochs on 5k samples, then 10 epochs on the full dataset. Models with 3M–16M parameters (varying hidden size and layers) were trained using AdamW (1e-3 learning rate, 1e-4 weight decay) with step decay (γ = 0.5 every 5 epochs). We also experimented with CTC blank bias values 2.0, 1.5, 0.5 to improve alignment.

Group Relative Policy Optimization (GRPO). The policy model is a 3-layer BiLSTM (256 hidden units) trained with 80D Whisper mel features and CTC decoding over a 29-token vocabulary. Training proceeds in two stages: behavior cloning followed by GRPO fine-tuning. GRPO uses group size G=8, sampling temperatures 0.3, 0.5, 1, KL penalties 0.02, 0.1, 0.15, learning rates 1e-5, 2e-5, 5e-5, and batch size 8 for 10 epochs. Only policy parameters are updated (reference model frozen) using AdamW and a StepLR scheduler (decay factor 0.7 every 3 epochs). The reward function combines CER (weight 1.0), WER (0.5), and length regularization (0.1). Advantages are computed relative to the group mean. KL penalties prevent policy drift; evaluation uses CER and WER.

6 Results

6.1 Quantitative Evaluation

LibriSpeech. Table 1 shows the performance of all training methods on the LibriSpeech validation set. All reinforcement learning (RL) methods outperformed the supervised CTC baseline, validating the benefit of optimizing directly for recognition accuracy.

GRPO achieved the best results (17.87% CER, 48.62% WER), likely due to its use of group-relative comparisons, which stabilize training and encourage diverse, high-quality outputs. REINFORCE followed (20.22% CER, 49.10% WER), benefiting from direct reward-based learning but limited by high variance in gradient estimation. Actor-Critic showed slightly worse performance (20.98% CER, 51.75% WER), possibly due to imperfect value estimates from the critic early in training.

Behavior Cloning, which imitates Whisper-tiny predictions, performed worst (21.55% CER, 55.49% WER). This underperformance may stem from architectural mismatches between the transformer-based Whisper and the BiLSTM model, leading to poor transferability of behavior patterns.

Out-of-Domain Evaluation on FLEURS. FLEURS introduces significant domain shift in terms of speaker diversity, recording conditions, and content. As expected, all models showed increased error rates compared to LibriSpeech.

REINFORCE generalized best (43.04% CER, 87.79% WER), likely because its sample-based learning enables exploration and adapts better to distribution shift. Actor-Critic followed closely (43.15% CER, 88.07% WER), benefitting from policy gradient updates but potentially hindered by critic overfitting to in-domain data.

GRPO and Behavior Cloning experienced the most degradation. Behavior Cloning suffered a large drop (84.33% CER, 98.16% WER), and demonstrated poor generalization due to overfitting to

Whisper's predictions without learning speech patterns. GRPO's degradation likely stems from the same domain mismatch issues - the policy learned to optimize rewards on LibriSpeech's clean speech distribution, which doesn't transfer to FLEURS' more challenging acoustic conditions.

These results suggest that reinforcement learning approaches, particularly REINFORCE, offer better generalisability to domain shifts than imitation or supervised learning alone.

Table 1: DRL ASR performance comparison across datasets

	LibriSpeech		FLEURS	
Method	**CER (%)**	**WER (%)**	**CER (%)**	**WER (%)**
CTC Baseline	22.00	53.13	43.65	88.02
REINFORCE	20.22	49.10	**43.04**	**87.79**
Actor-Critic	20.98	51.75	43.15	88.07
Behavior Cloning	21.55	55.49	84.33	98.16
GRPO*	**17.87**	**48.62**	81.58	98.07

*GRPO results on LibriSpeech are from a full-dataset run (10 epochs, 50 hours), as opposed to the 5k dataset run mentioned in the poster.

6.2 Qualitative Analysis

6.2.1 LibriSpeech

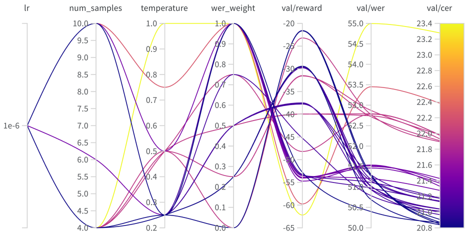

Figure 4: REINFORCE hyperparameter sweep

Table 2: REINFORCE: Importance & Correlation

Param	**CER**	**WER**
Temp	High (0.90)	High (0.94)
CER Wt.	Low (0.28)	Med (0.10)
WER Wt.	Low (-0.28)	Med (-0.10)
Samples	Low (0.09)	Low (0.02)

REINFORCE. Figure 4 and Table 2 show that REINFORCE is most sensitive to temperature, which strongly correlates with CER and WER. Lower temperatures improve performance by focusing sampling on high-probability outputs. Reward weights show expected inverse correlations, and sample count has minimal effect, indicating efficiency with modest sampling.

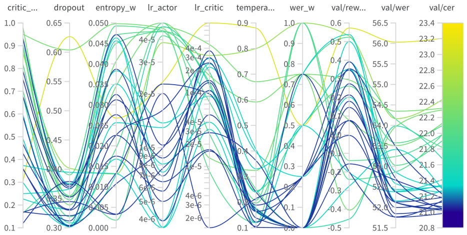

Figure 5: Actor-Critic hyperparameter sweep

Table 3: Actor-Critic: Importance & Correlation

Param	**CER**	**WER**
Actor LR	High (0.72)	High (0.59)
Dropout	Low (0.65)	High (0.64)
Temp	Med (0.74)	Med (0.73)
Critic LR	Low (0.52)	Low (0.54)
CER Wt.	Low (-0.62)	Low (0.24)
WER Wt.	Low (0.62)	Low (-0.24)
Critic Coef	Low (0.35)	Low (0.09)
Entropy Wt.	Low (0.21)	Low (0.01)

Actor-Critic. Figure 5 and Table 3 highlight Actor-Critic's sensitivity to actor learning rate and temperature, both strongly correlated with performance. Dropout also plays a key role in avoiding overfitting. Critic-related parameters show moderate influence, and reward weights again show asymmetric effects, with CER weight improving CER but slightly hurting WER. Entropy and critic coefficients had minimal impact.

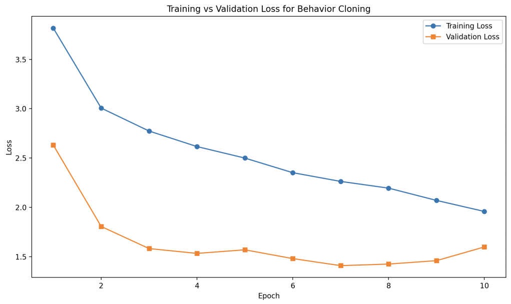

Figure 6: Training and validation loss for BC model

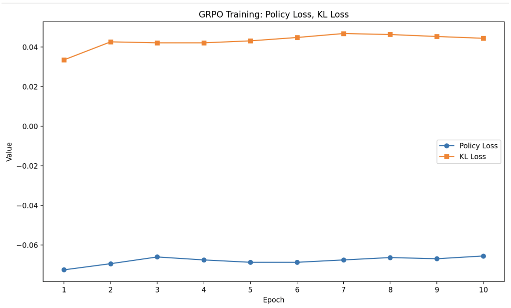

Figure 7: Policy and KL loss for GRPO model

Behavior Cloning. Figure 6 shows that the training loss is steadily decreasing which indicates that the model is learning from the training data effectively. Validation loss also decreases initially showing us that the model is generalizing well to unseen data. However, at higher epochs, the validation loss starts increasing which could be an indicator of overfitting.

Group Relative Policy Optimization (GRPO). Figure 7 shows that the policy loss is consistently negative but trending towards zero. This is the expected output for GRPO as the model optimizes, the group relative rewards become smaller. The KL loss is slowly increasing and is fairly stable indicating the policy is diverging from the pretrained behavior cloning model, enabling more exploration.

6.2.2 Out-of-Distribution Qualitative Analysis on FLEURS

Table 4: Sample transcriptions on FLEURS

Model	Output
Reference	however due to the slow communication...
CTC Baseline	however do do tesol coninicaten...
Actor-Critic	however dod de tesfol cominicate...
REINFORCE	however do do tefol coninicatein...
Behavior Cloning	oe theoeniy sthod...
GRPO	ooer ed o theol minniay...

Table 4 shows representative transcriptions on the FLEURS dataset. All models struggle with domain shift, exhibiting phonetic confusion and poor handling of complex or technical terms.

The CTC baseline suffers from fragmented words and frequent character-level errors (e.g., “coninicaten” for “communication”). Actor-Critic improves slightly with more coherent word boundaries and partial preservation of structure (e.g., “cominicate”).

REINFORCE shows the best phonetic alignment and syntactic coherence among RL approaches, with outputs like “coninicatein” retaining greater similarity to the target word. It also maintains better grammatical structure overall, likely due to direct reward optimization.

Behavior Cloning and GRPO perform worst under domain shift, producing highly corrupted outputs. These methods appear more sensitive to acoustic mismatch, indicating weaker generalization in out-of-domain settings.

7 Discussion

Our evaluation of reinforcement learning (RL) approaches for ASR fine-tuning yields several key findings. Despite its simplicity, REINFORCE outperformed Actor-Critic across both in-domain and out-of-domain settings. This suggests that direct policy optimization may be more effective for ASR than value-based variance reduction, likely due to the difficulty of learning accurate value estimates from speech data.

Hyperparameter analysis highlighted temperature as the most influential factor in both REINFORCE and Actor-Critic, with lower values improving transcription accuracy by concentrating sampling near high-probability outputs. In contrast, sample count showed low importance, indicating that effective learning is possible with modest computational cost.

Reward weighting revealed asymmetric behavior: increasing CER weight generally improved both CER and WER, while high WER weighting degraded performance. This suggests that prioritizing character-level accuracy leads to better generalization, particularly under noisy or mismatched conditions.

Behavior Cloning underperformed across all evaluations, especially on FLEURS. Its reliance on outputs from a structurally different model (Whisper-tiny) likely contributed to poor alignment, reinforcing that direct reward optimization is more effective than cross-architecture imitation.

Training BC and GRPO presented significant computational challenges due to high runtime and memory requirements from large dataset processing and Whisper model integration, requiring extensive memory optimization techniques including chunked data loading and singleton processor caching to achieve practical training times.

Out-of-domain evaluation on FLEURS showed performance degradation for all models, but REINFORCE and Actor-Critic retained their relative gains, indicating stronger robustness compared to the baseline. GRPO and Behavior Cloning degraded most, likely due to overfitting or architectural fragility. These results emphasize RL's potential to enhance generalization in ASR systems, though domain adaptation remains a critical challenge.

8 Conclusion

We present the first comprehensive comparison of REINFORCE, Actor-Critic, and Group Relative Policy Optimization (GRPO) for fine-tuning CTC-based ASR models. All RL methods outperformed the supervised baseline on LibriSpeech, with GRPO achieving the lowest CER (17.87%) and WER (48.62%) and REINFORCE showing the best out-of-domain performance on FLEURS.

Our findings underscore the importance of temperature control for stable training and show that small sample sizes suffice for effective learning. We also demonstrate that reward shaping, favoring character-level accuracy, is key to optimizing transcription quality.

Behavior Cloning underperformed, highlighting the limitations of knowledge distillation between divergent model architectures. In contrast, metric-driven RL fine-tuning improved generalization, making it a promising direction for future ASR research.

These results lay the groundwork for scaling RL-based ASR to transformer architectures and multilingual settings, offering a path toward more reliable and generalizable speech systems.

9 Team Contributions

- **Ali Sartaz Khan:** Worked on CTC baseline, REINFORCE, Actor-Critic. and FLEURS inference.
- **Prerana Rane:** Worked on modified CTC baseline with Whisper integration, behavior cloning and GRPO.

Changes from Proposal Due to the incomplete functionality of REBORN's codebase (Tseng et al., 2024), we shifted focus to applying DRL techniques to supervised ASR. We split the work as two DRL algorithms per team member.

References

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. *arXiv preprint arXiv:2205.12446* (2022). https://arxiv.org/abs/2205.12446

Yan Gao, Titouan Parcollet, and Nicholas Lane. 2021. Distilling Knowledge from Ensembles of Acoustic Models for Joint CTC-Attention End-to-End Speech Recognition. arXiv:2005.09310 [cs.LG] https://arxiv.org/abs/2005.09310

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

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

Rohit Prabhavalkar, Tara N. Sainath, Yonghui Wu, Patrick Nguyen, Zhifeng Chen, Chung-Cheng Chiu, and Anjuli Kannan. 2017. Minimum Word Error Rate Training for Attention-based Sequence-to-Sequence Models. arXiv:1712.01818 [cs.CL] https://arxiv.org/abs/1712.01818

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] `https://arxiv.org/abs/2402.03300`

Liang-Hsuan Tseng, En-Pei Hu, Cheng-Han Chiang, Yuan Tseng, Hung-yi Lee, Lin-shan Lee, and Shao-Hua Sun. 2024. REBORN: Reinforcement-Learned Boundary Segmentation with Iterative Training for Unsupervised ASR. *arXiv preprint arXiv:2402.03988* (2024).