Group Relative Policy Optimization for Speech Recognition

Prashanth Gurunath Shivakumar, Yile Gu, Ankur Gandhe, Ivan Bulyko
Amazon Science, Seattle, U.S.A
prashanth.g.s@ieee.org, {yilegu,aggandhe,ibbulyko}@amazon.com

***Abstract*—Speech Recognition has seen a dramatic shift towards adopting Large Language Models (LLMs). This shift is partly driven by good scalability properties demonstrated by LLMs, ability to leverage large amounts of labelled, unlabelled speech and text data, streaming capabilities with auto-regressive framework and multi-tasking with instruction following characteristics of LLMs. However, simple next-token prediction objective, typically employed with LLMs, have certain limitations in performance and challenges with hallucinations. In this paper, we propose application of Group Relative Policy Optimization (GRPO) to enable reinforcement learning from human feedback for automatic speech recognition (ASR). We design simple rule based reward functions to guide the policy updates. We demonstrate significant improvements in word error rate (upto 18.4% relative), reduction in hallucinations, increased robustness on out-of-domain datasets and effectiveness in domain adaptation.**

***Index Terms*—Speech Recognition, ASR, GRPO, LLM**

I. Introduction

Recently, significant strides have been made in Automatic Speech Recognition by adoption of large language models based on causal, decoder-only, transformer architectures. This is mainly driven by remarkable scaling properties of the LLMs with the ability to leverage large amounts of supervised, unsupervised speech and text data, streaming friendly properties due to causality imposed on transformers, and its simplicity in terms of less components which enables to treat the models as a black-box.

Typical auto-regressive LLM models the probability of the next token given a sequence of tokens. This paradigm can be extended to include speech modality by modeling the probability of next text token conditioned on a sequence of audio representations or audio tokens. [1] proposed to use acoustic tokens derived from K-means quantized HuBERT embeddings to model speech continuation with LLMs. VoxtLM [2], SpeechGPT [3] used such discrete audio units in application to ASR in a multi-task framework. Several studies [4]–[10] demonstrated strong ASR performance with directly feeding continuous speech embeddings to LLM. Conformer derived embeddings [4], [6], [10], HuBERT embeddings [11], Whisper encoder [5], [7]–[9], WavLM [5] are popular speech representations employed. While some studies have explored fixed conformer embeddings [6] with learnable speech projection layers, others choose to update conformer parameters during LLM training [7], [11]. [6], [8], [10] utilized low rank adapters (LoRA) for fine-tuning the LLM. Some prior studies [7] have adopted 2 stage training, which comprises freezing LLM while updating the audio encoder in the first stage. Second stage involves fixing the audio encoder and fine-tuning the LLM. [11] proposed a deep fusion mechanism based on gated cross attention modules operating on HuBERT features. The HuBERT model parameters were also updated during the LLM training. [9] conducted detailed study on three speech to LLM interface modules, including simple linear projection, multi-head cross-attention and Q-Former modules and found the latter to be better for ASR.

While the LLM based ASR has made significant strides in improving recognition rate, they often suffer with hallucinations [12], [13]. Hallucinations are characterized with high insertion rates and arise when the model deviate from input stimuli (speech signal), prioritizing distributional patterns resulting in semantically and phonetically unrelated outputs. Such hallucinations can have dangerous impacts, including deception, on applications in domain where high precision is critical. [12] conducted a detailed study characterizing the LLM hallucinations in the context of ASR. They find that low WER can often conceal big hallucinations, and the hallucinations can result from common audio signal perturbations including noise, pitch shift, temporal shifts. Further [13] finds that noise in human annotations, labeling, can lead to hallucinations.

One plausible way to further optimize speech recognition and increase its robustness to hallucinations is with re-inforcement learning (RL). Several studies have explored re-inforcement learning techniques in application to speech recognition. Early applications of re-inforcement learning attempted to correct errors in isolated word recognition [14]. A confidence based RL scheme based on incremental conditional entropy maximization was proposed in [15] demonstrating reduction in WER up-to 18%. Policy gradient based approach was explored in [16] which allows to directly optimize the edit-distance resulting in up-to 6% improvement in character error rate (CER). [17] applied re-inforcement learning in a hypothesis selection framework over the n-best model output. Authors in [18], proposed re-inforcement framework for fusing multiple modalities, i.e., speech and video for improving auto-regressive audio visual speech recognition.

In the context of LLMs, recent strides have been made in re-inforcement learning from human feedback (RLHF). OpenAI in [19] introduced a family of RL algorithms based on policy gradient methods named proximal policy optimization (PPO).

PPO paved a path towards a stabile, feasible and efficient way to enable RLHF for LLMs. Recently, [20] proposed Group Relative Policy Optimization (GRPO) for preference optimization of LLMs. GRPO simplifies PPO by dropping the critic model and instead employs average of rewards for advantage estimation. In its first application [20], the GRPO provided accuracy gains on math benchmarks. Subsequently, the algorithm provided a framework for building reasoning capabilities for LLMs [21]. Dynamic sampling policy optimization (DAPO) further tweaked the GRPO to enhance training effectiveness by introducing practical tricks including decoupled clip, token-level policy gradient loss and dynamic sampling [22]. [23] proposed a variant of GRPO named Dr. GRPO improving the token efficiency by making the optimization unbiased.

However, there have been few attempts at applying RLHF for LLMs operating on speech. [24] first proposed application of DPO to enhance speech continuation on spoken language models by introducing AI experts for generating automated preference data. Qwen2-Audio [25] applied DPO with human preferences to optimize speech understanding. Qwen2.5-Omni [26] applied DPO to enhance speech generation. Omni-R1 [27] further extended Qwen2.5-Omni model's capabilities by applying GRPO. [28] applied GRPO to improve on emotion recognition and reasoning capabilities that enables the model to better analyze visual and audio modalities. [29]–[31] fine-tuned Qwen2.5-Omni on audio QA datasets using GRPO to achieve state-of-the-art performance on MMAU and MMAR benchmarks. [32] applied GRPO for improving text-to-speech system by designing rewards to optimize the TTS-WER and speaker similarity of the synthesized speech.

In this study, we propose application of GRPO towards LLM based ASR as an additional stage of fine-tuning towards optimizing the overall performance and improve robustness of the system to hallucinations in application to out-of-domain, unseen acoustic conditions. To the best of our knowledge, this is the first attempt at application of RLHF to LLMs to improve speech recognition. We design and explore various reward functions and present our findings and strategies to improve speech recognition with re-inforcement learning. The rest of the paper is organized as follows: Section II presents the proposed LLM based ASR system. Section III provides the description of our experimental setup and datasets employed in our study. Section IV presents the experimental results and the discussions. Finally, the study and its findings are concluded in Section V.

II. Proposed Technique

A. Auto-regressive LLM based Speech Recognition

Auto-regressive, causal, decoder only LLMs model the probability of the next token given a sequence of tokens:

$$P_{LM}(X) = \prod_{t=1}^{T} P(x_t | x_{t-1}, \ldots, x_1) \quad (1)$$

where $x_t \in V_{txt}$ is a text token belonging to text vocabulary V_{txt}. Such a model can be adopted to the task of speech recognition using next token paradigm by modeling:

$$P_{ASR}(X|S) = \prod_{t=1}^{T} P(x_t | x_{t-1}, \ldots, x_1, s_N, \ldots, s_1) \quad (2)$$

where $s_1, \ldots, s_N$ are acoustic units or representations of length N frames corresponding to their transcriptions $x_1, \ldots, x_T$ of length T. In this work, s_t are continuous vector representations derived from a pre-trained acoustic encoder. The details of acoustic encoder are described under section III. A simple linear projection, feed-forward layer is used as interface in mapping the acoustic representation to LLM input. The acoustic encoder is frozen during the training, while both the linear projection and LLM parameters are updated. In case of ASR, we compute the next-token prediction loss only on the output transcriptions. The model is trained in 2 stages. During the first stage, LLM is pre-trained on large corpus of text using Equation 1. Next, the LLM is fine-tuned on parallel speech-text supervised data using Equation 2. The prompt format for ASR in our setup comprises `<User><BOS> Convert speech to text <S-BOS>` $s_1, s_2, \ldots, s_N$ `<System> <BOS>` $x_1, x_2, \ldots, x_T$ `<EOS>`. During inference, the LLM is prompted with `<User><BOS> Convert speech to text <S-BOS>` $s_1, s_2, \ldots, s_N$ `<System>`. In the above sequences, `<User>, <BOS>, <S-BOS>, <EOS>` are special tokens.

B. Group Relative Policy Optimization for Speech Recognition

In this study, we propose an additional fine-tuning stage based on RLHF for further performance optimization and robustness. RLHF algorithms provide an effective means of incorporating human preferences to direct LLM generations. PPO introduced in [19] is a popular policy gradient RL technique for LLMs. PPO is based on the Actor-Critic model, where the Actor interacts with the environment collecting rewards associated with each action which pertain to quantifying how good or bad was the action taken. A critic model is trained alongside to estimate the expected future reward from the current state. The PPO then computes advantages for each action taken which describes the quality of the action relative to the average expected return. More importantly, PPO paved a path towards a stable, feasible and efficient way to enable RLHF for LLMs. Group Relative Policy Optimization (GRPO) introduced in [20], is a variant of PPO, primarily designed for preference optimization of LLMs. GRPO simplifies PPO by dropping the critic model and instead employs average of rewards for advantage estimation.

In the context of ASR, such RL techniques can directly optimize the speech recognition output to human transcriptions. While, technically, both PPO and GRPO can be applicable, GRPO provides simpler framework to achieve our objectives. Further, in the case of ASR, the human feedback is derived directly from groundtruth transcripts, hence, objective, rule-based rewards can suffice. Thus, GRPO is a better fit without

necessitating the need for training separate reward models. The reward models themselves often suffer from reward hacking problem [33]. Moreover studies such as [20], have found GRPO to yield similar performance gains as PPO.

The GRPO optimizes and maximizes the following objective:

$$L_{GRPO} = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} min[\pi_\vartheta \hat{A}_{i,t}, clip(\pi_\vartheta, 1-\varepsilon, 1+\varepsilon])\hat{A}_{i,t}] - \beta \mathbb{D}_{KL}[\pi_\theta||\pi_{ref}] \tag{3}$$

where G is the number of generations, o_i is the i^{th} generated output hypothesis, $\hat{A}_{i,t}$ is the advantage, given by:

$$\hat{A}_{i,t} = \frac{\mathbb{R}(i) - \mathbb{E}[\mathbb{R}(i)]}{\sigma(R(i))} \tag{4}$$

where $\mathbb{R}(i)$ is the reward for generated output o_i, $\mathbb{E}[R(i)]$ is the expected reward across G generations and $\sigma(R(i))$ is the standard deviation. ε is the parameter introduced in PPO [19] for clipping and stabilizing the training, β is a hyper-parameter that controls the deviation of the policy from the reference seed model, $\mathbb{D}_{KL}[\pi_\theta||\pi_{ref}]$ is the Kullback-Leibler divergence between the reference model π_{ref} and current policy model π_θ given by:

$$\mathbb{D}_{KL}[\pi_\theta||\pi_{ref}] = \frac{\pi_{ref}(o_{i,t}|o_{i,<t}, s)}{\pi_\theta(o_{i,t}|o_{i,<t}, s)} - log\frac{\pi_{ref}(o_{i,t}|o_{i,<t}, s)}{\pi_\theta(o_{i,t}|o_{i,<t}, s)} - 1 \tag{5}$$

and π_ϑ is given by:

$$\pi_\vartheta = \frac{\pi_\theta(o_{i,t}|o_{i,<t}, s)}{\pi_{\theta_{old}}(o_{i,t}|o_{i,<t}, s)} \tag{6}$$

where $\pi_{\theta_{old}}$ is the old policy model as defined in PPO [19].

Additionally, we also take learnings from DAPO [22], Dr. GRPO [23] and explore its effect in this study. DAPO enforces three key modifications to the loss, (i) introduces an upper clipping threshold to increase probability of unlikely exploration, (ii) remove KL-divergence term ($\beta = 0$), and (iii) token-level policy gradient loss computation (versus sample-level) to account for biases introduced as a function of sample length. On the other hand, Dr. GRPO proposes two key modifications to loss towards unbiased optimization, (i) removal of standard deviation normalization in advantage computation, and (ii) length normalization of the GRPO loss. These modifications are geared towards preventing model's bias towards longer, incorrect responses.

C. Rule-based Rewards for Speech Recognition

Given that the human feedback for speech recognition is via ground-truth human transcriptions, we propose to use simple rule-based rewards. This simplifies the setup, reduces computational complexity and helps avoid reward hacking problem [33]. In this study, we explore the following rewards:

Word Error Rate (WER): Negated word-error-rate can serve as a potential reward and helps optimize directly to the target metric. WER is a normalized version of the edit-distance, hence can reinforce the model without any biases towards the length of the audio.

$$\mathbb{R}_i = -WER = -\frac{Sub + Del + Ins}{N} \tag{7}$$

where Sub, Del, Ins are substitutions, deletions and insertions respectively, derived from dynamic alignment, N is the total words in the reference.

Exact Match (EM): Several studies [22] recommend simpler rewards such as exact match which is proven to be an effective approach for invoking reasoning capabilities.

$$\mathbb{R}(ref, hyp) = \begin{cases} 1 & \text{if } ref = hyp \\ 0 & otherwise \end{cases} \tag{8}$$

In context of ASR, exact match is invariant of the sample length as well as the number of errors.

Total Errors (ED): Number of incorrect recognition can be used as a reward. This is equivalent to un-normalized edit distance over word sequence between reference and generated candidates.

$$\mathbb{R}_i = -(Sub + Del + Ins) \tag{9}$$

This enables the optimization to concentrate towards samples that are drastically different to reference.

III. Data and Experimental Setup

A. Experimental Setup

In this work we employ two LLM models based on Llama3 architecture [40], (i) smaller, 2B parameter model, and (ii) larger, 8B parameter model. The 2B model comprises 24 layers, 16 attention heads, hidden dimension of 2048 and feed-forward dimension of 8192. The 8B model comprises of 32 layers, 32 attention heads, hidden dimension of 4096 and feed-forward dimension of 16384. The 8B model employs weight tying between the embedding and output layers. Both the LLMs use 8 query groups, SwiGLU activation and has a vocabulary size of 187178.

For consuming speech, our setup employs a speech encoder module based on [41]. The encoder is primarily based on the conformer architecture with 2B parameters. The encoder has a frame rate of 40ms and a hidden dimension of 2048, more details are available in [41]. A weighted combinations of multiple layers are used to encode speech signals onto 4096 dimensional embeddings. A linear projection layer is used to map the 4096 dimension embedding to the embedding dimension of the LLMs.

TABLE I
Speech Datasets Statistics

Datasets	Hours
Fleurs [34]	987 hrs
Multi-lingual Librispeech [35]	50k hrs
Voxpopuli [36]	1791 hrs
People Speech [37]	30k hrs
Common Voice [38]	2k hrs
Covost2 [39]	3.5k hrs

TABLE II
EXPERIMENTAL RESULTS: WORD-ERROR-RATE METRIC (ACRONYMS IN PARENTHESIS, WER, EM, ED, CORRESPOND TO REWARD FUNCTIONS)
SFT: SUPERVISED FINE-TUNING; GRPO: GROUP RELATIVE POLICY OPTIMIZATION; DAPO: DYNAMIC SAMPLING POLICY OPTIMIZATION

Datasets	**People-Speech**	**Multilingual Librispeech**						**Voxpopuli**					
Language	En	En	Fr	It	De	Es	Overall	En	Fr	It	De	Es	Overall
2B SFT	23.5	4.83	5.35	**10.5**	6.06	3.46	5.48	7.92	9.39	15.83	10.95	7.64	10.12
+ GRPO (WER)	21.53	4.85	5.22	10.7	6.15	3.43	5.49	7.82	8.5	14.93	9.89	7.24	9.47
+ $\beta = 0$	**21.3**	4.92	5.37	11.32	6.15	3.4	5.59	7.96	8.56	14.53	**8.96**	**7.14**	9.24
+ GRPO (ED)	**21.48**	5.19	5.39	10.7	**6.05**	3.66	5.64	7.87	8.55	14.91	9.86	7.38	9.51
+ GRPO (EM)	22.29	4.93	5.48	11.06	6.2	3.47	5.62	7.88	8.61	15.98	10.18	7.32	9.76
+ DAPO (EM)	21.71	**4.79**	5.24	11.54	6.13	**3.24**	5.51	7.98	8.53	15.14	9.81	7.43	9.57
+ DAPO (WER)	**21.30**	5.04	**5.2**	10.64	6.21	3.32	5.53	**7.8**	8.57	**14.42**	9.39	7.27	9.30
+ DR-GRPO (EM)	22.24	4.91	5.27	11.02	6.08	3.63	5.56	7.87	**8.48**	15.16	10.58	7.49	9.70
8B SFT	25.48	4.46	4.87	9.37	**5.07**	3.08	4.87	**7.5**	9.21	14.66	10.9	7.05	9.66
+ GRPO (WER)	**21.42**	4.64	**4.7**	9.45	5.33	**3.02**	4.95	7.66	8.6	**14.04**	**8.89**	**6.64**	8.98
+ GRPO (EM)	22.49	**4.41**	4.86	**9.16**	5.25	3.03	4.87	7.66	**8.42**	14.93	9.34	7.03	9.27

For speech recognition, the LLM is first pre-trained on text-only data with a constant learning rate schedule of 1e-4 using Adam optimizer. The global batch size is approximately 1M tokens with max-sequence length set to 2048. Motivation to pre-train on text is derived from [42]–[44], which have shown benefits for speech related tasks. Next, the LLM is fine-tuned using speech data with a cosine learning rate scheduler with peak learning rate of 5e-6 for 100k steps, with 1000 step warm-up. The global batch size is set to 128 and sequence length is capped at 2048 tokens.

GRPO training stage uses the fine-tuned model as the reference. It comprises generating G <System> responses when prompted with the <User> sequence to compute the loss in Equation (3). GRPO is conducted with a fixed learning rate of 1e-6 similar, global batch size of 64 for a maximum of 5000 steps. We experiment with different generation configurations including β, number of generations, generation strategies, reward scaling. For all of our experiments, the top_k and min_p is set to None, top_p = 1.0, repetition penalty = 1.0, β = 0.04 (when used).

B. Data

The pre-training text data is based on RedPajama [45]. The speech datasets used in this study is presented in Table I. We use open-sourced multi-lingual speech datasets comprising approximately 88,000 hours of training data for supervised fine-tuning as well as GRPO. The datasets are mixed with weights representative of their sizes, emphasizing English, French, Spanish, Italian and German. We maintain held-out development and evaluation partitions for each of the corpus. Best model checkpoints are picked based on development partition. We present evaluations on People's speech, multilingual librispeech and Voxpopuli datasets to capture diverse range of variability. We also provide a language level breakdown of WER for MLS and Voxpopuli evaluations. Additionally, we use AMI corpus [46], and TEDLium [47] for out-of-domain evaluations.

IV. RESULTS

Table II presents the experimental results for 2B and 8B models on people's speech, multi-lingual librispeech and voxpopuli. We present detailed results with different configurations on the 2B model. The 8B model is used to assess the effect of scaling on the proposed method. Firstly, comparing the 2B models, it is evident that the proposed method improves over the reference SFT model on most languages across datasets. We observe up-to 8.6% relative on people's speech, up-to 5.8% relative on multilingual librispeech (MLS) and up-to 18.2% relative improvements on Voxpopuli. Assessing the results with the 8B model, firstly we note the improvements with the bigger model in comparison to 2B SFT model (with the exception of people's speech). Comparisons of the 8B SFT model with the GRPO, paint a similar picture as 2B, with up-to 15.7% relative improvements on people's speech, up-to 3.5% relative improvement on MLS and 18.4% relative improvement on Voxpopuli. Degradations, if any, are small, demonstrating the robustness of the proposed technique across varying acoustic conditions.

Next, we assess different configurations of the proposed techniques:

Role of Rewards: We compare the 3 rewards as described under section II. We found that the WER and the total error rewards fare better when the absolute WER of the datasets are high. The exact-match performs relatively poor in high WER conditions likely due to lower probability of generating outputs that positively match the reference. However, the exact-match performs on-par with the WER reward on datasets where the absolute WER is low. On the other hand, the total errors (ED) exhibits an opposite trend. Overall, the WER based reward strikes a better balance across varying WER conditions.

Role of KL Divergence: β regulates the divergence of the policy model from the reference model. A β of 0 removes any regulation and essentially removes the KL Divergence term. In our experiments we do not observe significant divergence in WER when $\beta = 0$.

Role of RL Algorithms: We compare GRPO, DAPO and Dr.

TABLE III
EXPERIMENTAL RESULTS: OUT-OF-DOMAIN EVALUATIONS. WER AND ITS BREAKDOWN IN INSERTIONS, DELETIONS AND SUBSTITUTIONS

Datasets	**TEDLIUM**		**AMI-IHM**		**AMI-SDM**	
Models	Ins / Del / Sub	WER	Ins / Del / Sub	WER	Ins / Del / Sub	WER
2B SFT	0.5 / 1.7 / 1.6	3.9	38.4 / 10.3 / 7.3	56.0	56.7 / 16.6 / 14.6	87.88
+ GRPO (WER)	0.4 / 1.6 / 1.6	3.7	2.8 / 10.96 / 6.3	20.0	7.4 / 18.1 / 12.1	37.59
+ GRPO (ED)	0.5 / 1.7 / 1.6	3.74	5.6 / 11 / 6.3	22.84	8.9 / 17.6 / 12.1	38.52
+ GRPO (EM)	0.4 / 1.8 / 1.7	3.93	2.7 / 11.1 / 6.3	19.95	7.4 / 18.1 / 12.1	37.59
8B SFT	0.5 / 1.2 / 1.8	3.53	82 / 10.5 / 7.7	100.26	195.4 / 17.4 / 15.3	227.98
+ GRPO (WER)	0.4 / 1.28 / 1.67	3.36	4.6 / 11.6 / 6.0	22.16	9.0 / 20.2 / 10.5	39.69

TABLE IV
EXPERIMENTAL RESULTS IN WER: DOMAIN ADAPTATION.

Datasets	**AMI-IHM**	**AMI-SDM**	**Multilingual Librispeech**					**Voxpopuli**				
Language	En	En	En	Fr	It	De	Es	En	Fr	It	De	Es
2B Baseline	56.0	87.88	**4.83**	**5.35**	**10.5**	**6.06**	**3.46**	**7.92**	**9.39**	**15.83**	**10.95**	**7.64**
AMI-SFT	19.17	44.34	6.91	12.10	14.88	8.39	7.57	9.80	12.48	19.33	12.92	12.21
AMI-GRPO (WER)	**15.55**	**31.98**	6.59	10.34	15.80	9.57	5.90	9.10	12.15	19.40	14.51	10.62

GRPO with exact-match reward. The results suggest that both DAPO and Dr. GRPO outperform traditional GRPO in most cases.

Role of Generation Strategies: Beam search decoding and multi-nomial sampling decoding strategies were explored. We found that beam search often leads to better improvements on noisy datasets like people speech and the multinomial sampling offers better improvements on cleaner datasets with lower WER.

Number of Generations: In our experiments we tested $G = \{6, 10\}$. However, we found the impact to be insignificant and hence skip the results.

A. Out-of-Domain Performance Evaluations

Translation of reliable performance to unseen acoustic environments is critical for any robust ASR system. Particularly, in case of auto-regressive LLMs, this often leads to hallucinations especially in noisier acoustic environments, characterized with reverberations, overlapping speech and background noises. To assess the performance of the proposed technique, we conduct evaluations on unseen datasets including TEDLIUM and AMI meeting corpus. TEDLIUM comprises of TED-talks with diverse speakers that help us probe on speaker related challenges including diverse range of accents, fluency. AMI meeting corpus poses challenges with far-field speech, overlapped speech and noise. Note, both TEDLIUM and AMI meeting corpus are not incorporated during SFT and subsequent GRPO training.

Table III presents the results on out-of-domain evaluations. The 2B SFT model, gives a WER of 3.85% on TEDLIUM, however, performs poorly on AMI, i.e., 55.96% on IHM and 87.88% on SDM. A deeper inspection of the errors on AMI corpus in terms of insertions, deletions and substitutions reveal that insertions dominate the errors hinting towards hallucinations. After GRPO, we see a dramatic reduction in insertions which drives significant improvements in WER. It is also important to note that there is significant reduction in substitutions after GRPO. In case of the 8B model, we observe the 8B SFT model scales well on TEDLIUM with reduction in baseline WER over 2B SFT model. However, the WER explodes to greater than 100% which is suggestive of increased hallucinations. It is likely that bigger models adapt well to in-domain data and as a consequence worsens hallucinations on unseen acoustical environments. After GRPO, we see drastic reduction in insertions and substitutions similar to the 2B models. The results highlight that GRPO can increase the robustness of the LLM and reduce hallucinations. More importantly the learning extends to unseen datasets and acoustic conditions.

B. Domain Adaptation

Speech signals are characterized by high variability in multiple domains including speaker environment, noise, room characteristics, reverberation, recording conditions, speaker variability spanning linguistics, accents, fluency, age, and gender. It is typical to adapt ASR models to unseen domains to optimize performance. In case of ASR-LLMs, one straightforward option is to fine-tune on new domain. We evaluate the proposed GRPO training as an alternative and assess the overall robustness. We start from the 2B SFT model trained on data presented under Table I as the baseline (corresponding to row 1 in Table II). We train 2 candidate model on AMI speech corpus as a new domain: (i) SFT adaptation, (ii) GRPO adaptation to assess effectiveness of SFT versus proposed method for domain adaptation. The choice of AMI speech corpus is due to its distinct characteristics in terms of acoustic environment (supported by results in Table III). The results are presented under Table IV. Along with results on AMI, we also provide the results on Multi-lingual librispeech and Voxpopuli to assess the performance trade-off after adaptation. From the results, it is clear that the baseline model performs poorly on the out-of-domain AMI data. After adapting the

baseline on AMI data using typical next-token prediction SFT, we observe substantial improvements, 66% relative WER reduction on AMI-IHM and 49% WER reduction on AMI-SDM subsets. Meanwhile, we also observe degradations on MLS and Voxpopuli across all languages. Looking at the results with the proposed GRPO adaptation, we see significant improvement relative to both the baseline (72% on AMI-IHM and 63.61% on AMI-SDM) as well as SFT adapted model, i.e., 18% relative WER reduction on AMI-IHM and 27.9% on AMI-SDM. Notably, we observe degradations on MLS and Voxpopuli, however, the degradations are relatively lower compared to the SFT adapted model. This suggests that the proposed GRPO is a better tool to use for adapting the model to a new unseen data or domain.

A highlight of the above results on AMI is that the proposed GRPO training even without inclusion of AMI datasets, i.e., out-of-domain results in Table III, row 2, gives better results compared to the SFT model adapted on AMI. This concretely establishes the robustness benefits obtained using proposed method on out-of-domain datasets.

V. Conclusion

In this work, we propose an additional RLHF training stage for LLM based ASR models using GRPO. We propose 3 simple rule-based rewards for GRPO to facilitate performance improvements and robustness. We carefully design experiments to evaluate the performance benefits, assess the robustness of the model to hallucinations, out-of-domain datasets. Further, we demonstrate the proposed method as an effective tool for domain adaptation purposes. The experiments demonstrate significant WER reductions obtained using the proposed method. We also show that the resultant model performs drastically better on out-of-domain datasets that are otherwise prone to hallucinations. Additional experiments support the viability of the proposed method as an effective model adaptation tool. We provide detailed discussions on the role of different hyperparameter settings and present strategies and recommendation for effective usage.

In future, interesting rewards can be designed for specific applications, for example, improve slot-error-rate in spoken language understanding applications, or semantic measures to facilitate better, semantically aligned speech recognition outputs. This also opens up possibilities in responsible AI domain in censoring certain ASR outputs. Overall, the proposed method opens up possibilities in aligning and controlling certain aspects of ASR system.

References

[1] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[2] S. Maiti, Y. Peng, S. Choi, J.-w. Jung, X. Chang, and S. Watanabe, "Voxtlm: unified decoder-only models for consolidating speech recognition/synthesis and speech/text continuation tasks," *arXiv preprint arXiv:2309.07937*, 2023.

[3] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," *arXiv preprint arXiv:2305.11000*, 2023.

[4] Z. Du, J. Wang, Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma *et al.*, "Lauragpt: Listen, attend, understand, and regenerate audio with gpt," *arXiv preprint arXiv:2310.04673*, 2023.

[5] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang, Z. Du, F. Yu, Q. Chen, S. Zheng, S. Zhang *et al.*, "An embarrassingly simple approach for llm with strong asr capacity," *arXiv preprint arXiv:2402.08846*, 2024.

[6] J. Wu, Y. Gaur, Z. Chen, L. Zhou, Y. Zhu, T. Wang, J. Li, S. Liu, B. Ren, L. Liu *et al.*, "On decoder-only architecture for speech-to-text and large language model integration," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[7] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[8] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[9] W. Yu, C. Tang, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Connecting speech encoder and large language model for asr," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 637–12 641.

[10] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli *et al.*, "Prompting large language models with speech recognition abilities," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13 351–13 355.

[11] Y. Li, Y. Wu, J. Li, and S. Liu, "Prompting large language models for zero-shot domain adaptation in speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[12] H. Atwany, A. Waheed, R. Singh, M. Choudhury, and B. Raj, "Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models," *arXiv preprint arXiv:2502.12414*, 2025.

[13] R. Frieske and B. E. Shi, "Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models," *arXiv preprint arXiv:2401.01572*, 2024.

[14] K.-F. Lee and S. Mahajan, "Corrective and reinforcement learning for speaker-independent continuous speech recognition," *Computer Speech & Language*, vol. 4, no. 3, pp. 231–245, 1990.

[15] C. Molina, N. B. Yoma, F. Huenupán, C. Garretón, and J. Wuth, "Maximum entropy-based reinforcement learning using a confidence measure in speech recognition for telephone speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1041–1052, 2009.

[16] A. Tjandra, S. Sakti, and S. Nakamura, "End-to-end speech recognition sequence training with reinforcement learning," *IEEE Access*, vol. 7, pp. 79 758–79 769, 2019.

[17] T. Kala and T. Shinozaki, "Reinforcement learning of speech recognition system based on policy gradient and hypothesis selection," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5759–5763.

[18] C. Chen, Y. Hu, Q. Zhang, H. Zou, B. Zhu, and E. S. Chng, "Leveraging modality-specific representations for audio-visual speech recognition via reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 607–12 615.

[19] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[20] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[21] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[22] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, T. Fan, G. Liu, L. Liu, X. Liu *et al.*, "Dapo: An open-source llm reinforcement learning system at scale," *arXiv preprint arXiv:2503.14476*, 2025.

[23] Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin, "Understanding r1-zero-like training: A critical perspective," *arXiv preprint arXiv:2503.20783*, 2025.

[24] G.-T. Lin, P. G. Shivakumar, A. Gourav, Y. Gu, A. Gandhe, H.-y. Lee, and I. Bulyko, "Align-slm: Textless spoken language models with reinforcement learning from ai feedback," *arXiv preprint arXiv:2411.01834*, 2024.

[25] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[26] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang *et al.*, "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[27] H. Zhong, M. Zhu, Z. Du, Z. Huang, C. Zhao, M. Liu, W. Wang, H. Chen, and C. Shen, "Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration," *arXiv preprint arXiv:2505.20256*, 2025.

[28] J. Zhao, X. Wei, and L. Bo, "R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning," *arXiv preprint arXiv:2503.05379*, 2025.

[29] A. Rouditchenko, S. Bhati, E. Araujo, S. Thomas, H. Kuehne, R. Feris, and J. Glass, "Omni-r1: Do you really need audio to fine-tune your audio llm?" *arXiv preprint arXiv:2505.09439*, 2025.

[30] C. Wen, T. Guo, S. Zhao, W. Zou, and X. Li, "Sari: Structured audio reasoning via curriculum-guided reinforcement learning," *arXiv preprint arXiv:2504.15900*, 2025.

[31] G. Li, J. Liu, H. Dinkel, Y. Niu, J. Zhang, and J. Luan, "Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering," *arXiv preprint arXiv:2503.11197*, 2025.

[32] X. Sun, R. Xiao, J. Mo, B. Wu, Q. Yu, and B. Wang, "F5r-tts: Improving flow-matching based text-to-speech with group relative policy optimization," *arXiv preprint arXiv:2504.02407*, 2025.

[33] T. Everitt, V. Krakovna, L. Orseau, M. Hutter, and S. Legg, "Reinforcement learning with a corrupted reward channel," *arXiv preprint arXiv:1705.08417*, 2017.

[34] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.

[35] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[36] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[37] D. Galvez, G. Diamos, J. Ciro, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," *arXiv preprint arXiv:2111.09344*, 2021.

[38] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[39] C. Wang, A. Wu, and J. Pino, "Covost 2 and massively multilingual speech-to-text translation," *arXiv preprint arXiv:2007.10310*, 2020.

[40] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[41] P. R. Male, S. N. Ray, H. Arsikere, A. Jaiswal, P. Swarup, P. Sen, D. Chakrabarty, K. V. V. Girish, N. Bhave, F. Weber, S. Bhattacharya, and S. Garimella, "Durep: Dual-mode speech representation learning via asr-aware distillation," 2025. [Online]. Available: https://arxiv.org/abs/2505.19774

[42] M. Hassid, T. Remez, T. A. Nguyen, I. Gat, A. Conneau, F. Kreuk, J. Copet, A. Defossez, G. Synnaeve, E. Dupoux *et al.*, "Textually pretrained speech language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[43] T. A. Nguyen, B. Muller, B. Yu, M. R. Costa-Jussa, M. Elbayad, S. Popuri, C. Ropers, P.-A. Duquenne, R. Algayres, R. Mavlyutov *et al.*, "Spirit-lm: Interleaved spoken and written language model," *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 30–52, 2025.

[44] S. Maiti, Y. Peng, S. Choi, J.-w. Jung, X. Chang, and S. Watanabe, "Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13 326–13 330.

[45] T. Computer, "RedPajama: an open dataset for training large language models," 2023. [Online]. Available: https://github.com/togethercomputer/RedPajama-Data

[46] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The AMI meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[47] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Esteve, "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*. Springer, 2018, pp. 198–208.