A Domain Adaptation Framework for Speech Recognition Systems with Only Synthetic data

Minh Tran[1*], Yutong Pang[2], Debjyoti Paul[2], Laxmi Pandey[2], Kevin Jiang[2], Jinxi Guo[2],
Ke Li[2], Shun Zhang[2], Xuedong Zhang[2], Xin Lei[2]
[1] University of Southern California   [2] Meta AI

Abstract—We introduce DAS (Domain Adaptation with Synthetic data), a novel domain adaptation framework for pre-trained ASR model, designed to efficiently adapt to various language-defined domains without requiring any real data. In particular, DAS first prompts large language models (LLMs) to generate domain-specific texts before converting these texts to speech via text-to-speech technology. The synthetic data is used to fine-tune Whisper with Low-Rank Adapters (LoRAs) for targeted domains such as music, weather, and sports. We introduce a novel one-pass decoding strategy that merges predictions from multiple LoRA adapters efficiently during the auto-regressive text generation process. Experimental results show significant improvements, reducing the Word Error Rate (WER) by 10% to 17% across all target domains compared to the original model, with minimal performance regression in out-of-domain settings -1% on Librispeech test sets. We also demonstrate that DAS operates efficiently during inference, introducing an additional 9% increase in Real Time Factor (RTF) compared to the original model when inferring with three LoRA adapters.

Index Terms—Domain Adaptation, Speech Recognition

I. INTRODUCTION

Current state-of-the-art speech recognition systems are trained on extensive and varied datasets. For example, the Whisper model [1] underwent training with 680,000 hours of multilingual data, encompassing a wide array of audio conditions and environments. This extensive training strategy has enabled these models to express a high level of robustness in speech recognition across diverse conditions and multiple evaluation datasets [1]. Such robustness motivates us to question whether domain-specific data distribution, especially transcript distribution over language characterizing a given topic such as music or sports, are still relevant. Are state-of-the-art speech recognition systems universally effective, or is there still benefits in tailoring these robust models to specific domains?

In this paper, we explore this research question using Whisper, one of the most widely utilized and state-of-the-art automatic speech recognition (ASR) models. In particular, we introduce a domain adaptation framework, which we refer to as DAS . The framework is designed to efficiently adapt Whisper [1] to various language-defined domains, *i.e.,* speech with corresponding transcripts characterizing a given topic or genre. DAS is characterized by two key features: 1) it does not require any real data (audio or text) for adaptation to a target domain, thereby eliminating the substantial costs of human annotations and enhancing scalability, and 2) it

maintains effective adaptation across different domains without compromising its generalizability in out-of-domain conditions.

To achieve the first goal, we prompt large language models (LLMs) to generate texts that are relevant to a target domain. Subsequently, we employ text-to-speech to convert the generated text into corresponding audio data, and use this data to fine-tune the Whisper model specifically for that domain. To fulfill the second goal, we explore the use of a single encoder with multiple decoders, with each decoder dedicated to a specific domain. To mitigate the additional computational costs associated with multiple decoders, we employ LoRA (Low-Rank Adaptation) adapters. Additionally, we propose a novel one-pass decoding strategy that efficiently merge predictions from multiple LoRA adapters during each step of the auto-regressive text generation process. Experimental results demonstrate that our model consistently achieves a reduction ranging from 10% to 17% in the Word Error Rate metric (WER) over the original Whisper model on real speech across three target domains: music, weather/datetime, and sports. Notably, the model exhibit minimal performance regression in out-of-domain settings on multiple public benchmarks while introducing a modest latency increase of 9% in the Real Time Factor (RTF) metric compared to the original model. In summary, our contributions include:

- We provide a novel and thorough analysis on the effect of adapting ASR systems to language-defined domains using only synthetic audio/text data.
- We propose DAS , a novel domain adaptation framework for speech recognition systems, that effectively and efficiently adapt a base model to a set of target domains without sacrificing generalizabilty.
- We provide extensive experiments and ablation study to demonstrate the effectiveness of DAS .

II. RELATED WORK

Existing work on domain adaptation for ASR mainly focus on unsupervised domain adaptation settings, where both source and target domain speech are available but word transcripts are only available in the source domain. Sun *et al.* [3] proposed using domain adversarial learning with Gradient Reversal Layer [4] on deep neural networks to handle noise robustness or speech with different genres. Hsu *et al.* [5] propose a data augmentation method based on latent features extracted from variational autoencoders learned from both source and target domain speech. Meng *et al.* [6] extends
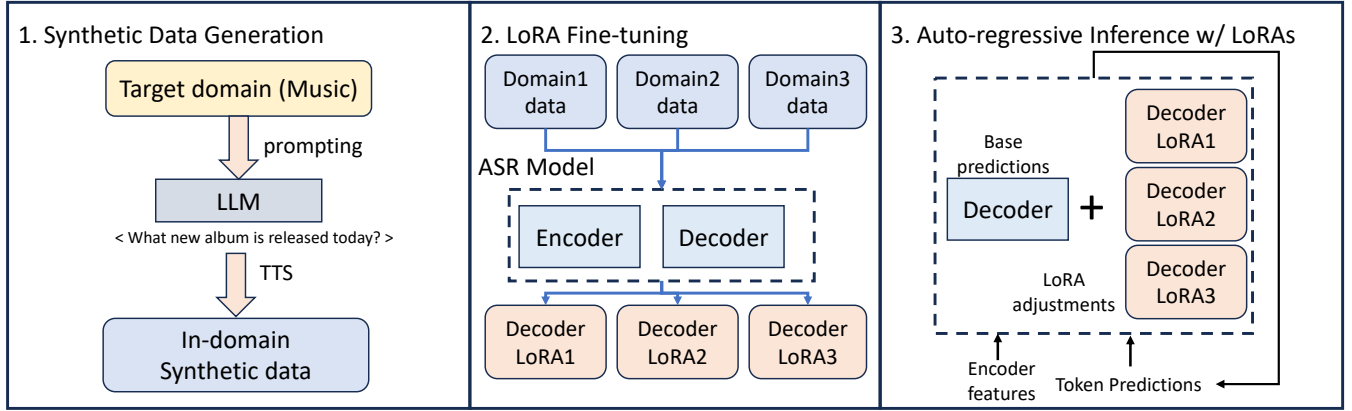
Fig. 1. **An overview of the proposed domain adaptation framework.** We first use LLM to generate texts related to a particular domain of interest (*e.g.,* music) and then use text-to-speech (TTS) to generate synthetic speech. We then fine-tune the decoder of *Whisper* with the generated synthetic data using Low-rank Adaptation (LoRA) [2]. Finally, we propose DAS to effectively generate predictions with multiple LoRAs to enhance in-domain speech recognition performance with minimal regression in out-of-domain settings.

Teacher/Student (T/S) learning to large-scale unsupervised domain adaptation of an end-to-end (E2E) model by using teacher's token posteriors as soft labels and one-best predictions for decoder guidance. Samarakoon *et al.* [7] explore domain adaptation for ASR in low-resource setting by adding domain labels to the label sequences, using factorized hidden layer adaptation and a novel domain-specific gating mechanism. Sim *et al.* [8] propose using factorized hidden layer for LSTM. Sim *et al.* [9] present an extensive study of different parameter-efficient fine-tuning methods for domain adaptation, and show the superior performance of Residual Adapters [10] and Low-rank Adapters (LoRA) [2]. Zheng *et al.* [11] explore the use of synthetic speech to enhance recognition performance on out-of-vocabulary words. In contrast, our work focuses on domain adaptation, aiming to increase the likelihood of specific vocabulary distributions being accurately predicted. Most relevant to our work, Huang et al. [12] fine-tune BART on a seed set of in-domain texts to generate domain-relevant texts. They use TTS-generated speech from these texts to augment the LibriSpeech dataset and train an RNN-T based model [13] from scratch, demonstrating significant improvement in in-domain ASR performance.

Compared to existing work, our main contributions are: 1) DAS does not assume any available in-domain text or real audio data, allowing scalability; 2) unlike [12], we focus on adapting a pre-trained ASR model and use only synthetic speech for adaptation rather than treating it as augmentation data alongside real speech data; and 3) we are the first to explore strategies to avoid performance regression for out-of-domain speech for ASR systems.

III. METHOD

Problem Formulation. Given a pre-trained ASR model $\mathcal{M}$ and a set of language-defined domains $\{d_1, \ldots, d_k\}$ with corresponding real speech test sets $\{\mathcal{D}_1, \ldots, \mathcal{D}_k\}$, our goal is to produce a new model $\mathcal{M}'$ from $\mathcal{M}$ such that $\mathcal{M}'$ achieves better performance (lower Word-Error-Rate) than $\mathcal{M}$ across all

Music	What musical movement did the beatles define in the sixties
	What group was Stevie Nicks a part of before Fleetwood Mac
	What's the title of elvis presley's first studio album
Weather	What's the average temperature difference between winter and summer in this park?
	What's the current temperature outside right now?
	How many centimeters of snow have fallen in Paris so far this month?
Sports	Which boxer held the heavyweight title for the longest duration in the 20th century?
	Which team did Tom Brady lead to victory in Super Bowl XXXVI?
	Wow many consecutive games did Cal Ripken Jr. play in the MLB?

TABLE I
SAMPLES OF CHATBOT-STYLE SYNTHETIC TEXTS GENERATED FROM OUR IN-DOMAIN TEXT GENERATION PIPELINE WITH LLAMA3-70B [14].

in-domain test sets while maintaining comparable performance to $\mathcal{M}$ on any out-of-domain test sets. In this study, we focus on three real language-defined test sets pertaining to Music, Weather/Date-time, and Sports. Further details of the datasets are provided in Section IV.

A. Synthetic data generation

We use the Llama3-70b model [14] to generate synthetic text for a specific domain, eliminating the need for a seed text dataset as in [12]. To enhance the diversity of the generated texts, our data generation process consists of two phases. In the first phase, we utilize LLMs to create tailored instructions related to the domain of interest based on some input seed instructions. Our seed instructions include simple fields such as *usecase* (e.g., question generation in English), *skills* (e.g., English singer, English album name), *persona* (e.g., teenager), and *instruction* (e.g., generate chatbot questions related to music). We base our instruction generation pipeline on *CodecLM* [15], which initially encodes seed instructions into metadata (keywords generated on-the-fly) and then decodes the metadata to produce tailored instructions. These instructions

	music	weather	sports
original	27.94	14.97	15.59
ft enc-dec	25.89	16.72	16.23
ft dec (last 3 layers)	24.99	14.45	15.46
ft full dec (6 layers)	**23.20**	**12.88**	**15.30**

TABLE II
COMPARISON (WER METRIC) OF DIFFERENT FINE-TUNING COMPONENTS
USING TTS IN-DOMAIN SPEECH DATA WITH THE *Whisper-base* MODEL ON 3
DOMAIN-SPECIFIC DATASETS.

are then fed into the LLMs to generate in-domain texts. Table I show some samples of our text generation pipeline. We then use an internal text-to-speech system to convert the generated text into synthetic speech data. We only use a single target speaker for all synthetic data without any audio augmentation.

B. DAS

Decoder-only fine-tuning. Unlike previous approaches that treat synthetic data as an augmentation to a real speech dataset and train models from scratch [12], we face challenges when working with pre-trained ASR models, where it is costly to re-train the models and the pre-training dataset is sometimes unavailable as in the case of Whisper [1]. Furthermore, using a specific dataset such as LibriSpeech [16] as the real counterpart might cause the pre-trained model to overfit to the specific conditions of that dataset (*e.g.*, clean audio), potentially compromising its generalizability. To address these issues, we explore different strategies for *Whisper-base* when fine-tuned solely with TTS-generated data. The results, presented in Table II, indicate that fine-tuning with just the decoder proves most advantageous when no real data is accessible. This finding leads us to develop DAS to use a single version of the pre-trained encoder followed by multiple versions of the decoder, with each version tailored to a specific domain. More details are provided in Section IV.

Parameter-efficient fine-tuning. The decoder accounts for the majority of the parameters in the Whisper model (72.4% in the `Whisper-base` version). Maintaining multiple versions of the decoder for each specific domain could be prohibitively expensive. Hu *et al.* [2] show that adapting LLMs can be done in a storage and computation efficient manner by integrating low rank decomposition matrices into the Transformer architecture [17]. In particular, for a pre-trained weight matrix W_o in a pre-trained neural network, the fine-tuning update ΔW can be constrained by two light-weighted matrices A and B such that

$$h = W_0 x + \Delta W x = W_0 x + BAx \quad (1)$$

During training, the pre-trained weights W_0 remain frozen and only A and B are trainable, allowing for efficient learning and adaptation. In this paper, we employ one LoRA adapter with less than 2% of the total number of parameters in Whisper for each domain to adapt the decoder of the Whisper model.

Auto-regressive decoding with multiple LoRAs. We introduce Auto-regressive Decoding with Multiple LoRA (ADML), detailed in Algorithm 1. ADML enables a single decoding pass that incorporates multiple LoRA adapters, efficiently leveraging batched computations with the base model weights W, as described in [18]. To optimize LoRA computations

---

Algorithm 1 Auto-regressive decoding with multiple LoRAs
___
Require: W, $\{(A_i, B_i)\}$ for $i \in [k]$, x: encoder features
1: $tokens \leftarrow []$
2: **while** $[eos] \notin tokens$ **do**
3:     h = $Softmax(W(x, tokens))$
4:     (next$_0$, c$_0$) = Argmax(h), Max(h) $\triangleright$ c denotes the confidence
5:     h_i = Softmax$((W + B_i A_i)(x, tokens))$ for $i \in [k]$
6:     (next$_i$, c$_i$) = Argmax(h_i), Max(h_i) for $i \in [k]$
7:     SELECT next from $\{$next$_0$, next$_1$,..., next$_k\}$
8:     INSERT next to $tokens$
9: **end while**
10: **return** $tokens$
___

	music	weather	sports
Synthetic dataset	44K	31K	46K
Evaluation dataset	2.1K	2.8K	5.1K

TABLE III
OF SAMPLES FOR EACH DATASET USED IN THIS STUDY.

$B_i A_i(x, \text{tokens})$, we restructured them from a sequential process into a batched computation, as follows:

$$\{B_i A_i x_i\}_{i=1}^N = B(AX) \quad (2)$$

where $A = concat(A_1, \ldots, A_k)$, $X = concat(x_1, \ldots, x_k)$, and $B = blk\_diag(B_1, \ldots, B_k)$[1]. Next, we select the optimal next token based on the confidence levels produced by different LoRA weights in Line 7. Through experiments with various next word selection strategies, we find out that it is best to select the next word from the $k + 1$ predictions according to

$$max\{c_i\}_{i=0}^k - c_0 \geq \tau \;\; OR \;\; min\{c_i\}_{i=0}^k - c_0 \leq -\tau \quad (3)$$

where τ is a threshold parameter. If both conditions are met, we priortize the word with the maximum confidence. If none of the conditions are met, we pick the word predicted by the original model.

IV. EXPERIMENTS

Datasets For evaluation, we utilize three real speech internal datasets collected using the Meta Ray Ban Glasses, with each sample being manually annotated. These datasets are divided into three language-defined domains: Music, Weather, and Sports. Given the nature of the data, the content predominantly features chatbot-style interactions, where a user requests Meta AI to perform tasks that involve verifying information via the internet. Statistics of the generated synthetic (for training) and evaluation datasets are available in Table III.

Implementation details We use `Whisper-base` as the pre-trained ASR model in this study. `Whisper-base` architecture resembles a vanilla transformer [17], with 6 layers of transformer encoder, 6 layers of transformer decoder and a feature dimension of 512 for a total of $74M$ parameters. For LoRA, we use the bottleneck dimension $r = 32$, $\alpha = 64$ with Rank-stable scaling [19], PiSSA initialization [20], and add LoRA adapters to the query and value attention blocks of Whisper, which are determined via hyper-parameter tuning. This results in around 1.2% additional parameters for each adapter added to the base model. During training, we fine-tune each LoRA

---
[1]https://pytorch.org/docs/stable/generated/torch.block_diag.html

	Train set	music	weather	sports
Original	-	27.94	14.97	15.59
FT	TTS-Music	**23.20** (↑ 17.0%)	14.45 (↑ 3.5%)	20.1 (↓ 28.8%)
FT	TTS-Weather	33.05 (↓ 18.3%)	**12.10** (↑ 19.2%)	17.7 (↓ 13.5%)
FT	TTS-Sports	25.05 (↓ 10.3%)	15.96 (↓ 6.6%)	**15.3** (↑ 1.9%)
LoRA-ft	TTS-Music	**23.23** (↑ 16.8%)	13.27 (↑ 11.3%)	16.51 (↓ 5.9%)
LoRA-ft	TTS-Weather	26.65 (↑ 4.6%)	**11.70** (↑ 21.8%)	15.08 (↑ 3.3%)
LoRA-ft	TTS-Sports	27.14 (↑ 2.9%)	14.05 (↑ 6.1%)	**13.37** (↑ 14.2%)
FT-Multi	TTS(M+W+S)	**24.71** (↑ 11.6%)	24.53 (↓ 64.0%)	15.84 (↓ 1.6%)
LoRA-ft-Multi	TTS(M+W+S)	25.09 (↑ 10.2%)	13.70 (↑ 8.4%)	14.61 (↑ 6.3%)
DAS	TTS(M/W/S)	24.87 (↑ 11.0%)	**12.39** (↑ 17.2%)	**13.98** (↑ 10.3%)

TABLE IV

ASR PERFORMANCE COMPARISON BETWEEN DAS AND BASELINES ACROSS THREE REAL DOMAIN-SPECIFIC TEST SETS.

	OOD_1	OOD_2	OOD_3	OOD_4
Original	12.02	5.04	10.87	10.36
LoRA-ft Multi	13.79	5.78	11.48	10.82
DAS	12.25	5.1	11.06	10.29
% change	-1.02	-1.01	-1.02	+0.99

TABLE V

ASR PERFORMANCE COMPARISON BETWEEN DAS AND ORIGINAL (UNADAPTED) MODEL ACROSS FOUR OUT-OF-DOMAIN TEST SETS. OOD_1: LIBRISPEECH TEST-OTHER, OOD_2: LIBRISPEECH TEST-CLEAN, OOD_3: FLEURS-EN, OOD_4: VOXPOPULI-EN.

	RTF	$\%\Delta_P$	$\%\Delta_S$	Speedup
original	0.0244	-	-	-
1-domain	0.0269	0.102	0.102	1x
2-domain	0.0263	0.077	0.213	2.76x
3-domain	0.0267	0.094	0.381	4.05x
10-domain	0.0289	0.184	1.582	8.6x
25-domain	0.0451	0.848	3.862	4.6x

TABLE VI

LATENCY COMPARISON BETWEEN DAS AND ORIGINAL MODEL WITH DIFFERENT NUMBERS OF ADAPTERS (DOMAINS). WE ALSO PROVIDE SEQUENTIAL LORA COMPUTATIONS ($\%\Delta_S$) TO DEMONSTRATE THE BENEFITS OF PARALLELIZATION ($\%\Delta_P$) AS DESCRIBED IN EQ (2).

weights independently with an in-domain synthetic dataset. The LoRA adapters are trained with an AdamW optimizer with $lr = 3e^{-6}$ for 10 epochs with a batch size of 16. We use a linear warmup learning rate for 10% of the training process. For inference with multiple LoRAs, we use $\tau = 0.025$. For evaluation, we report the Word Error Rate (WER) metric without punctuations and wakewords.

Baselines We compare our model with the following baselines: **FT** - (Decoder) vanilla finetuning; **LoRA-ft** - (Decoder) finetuning with LoRA adapters; **FT-Multi** - Full finetuning Whisper's decoder parameters on all 3 domains by combining the generated TTS data, where each training batch is drawn equally from the 3 domains, and **DAS** - Our proposed framework.

V. RESULTS

Our evaluation results, presented with the mentioned baselines and DAS in Table IV, demonstrate that fine-tuning the decoder with LoRAs [2] can potentially surpass the performance of full-finetuning (FT). Notably, LoRA-ft in the Music domain shows comparable results to (FT) with a WER of around 23.2% while outperforming FT in the remaining two domains. This aligns with findings reported in [2].

It is also important to highlight that both FT and LoRA-FT exhibit performance regression in domains other than the one they were adapted to. For example, FT adapted to the music domain (TTS-M) achieves a WER of 23.3%, which is a substantial improvement of 16.9% for the Music domain. However, it also results in a significant regression in the Sport domain, with a relative WER increase of 28.9% compared to the original model.

DAS addresses this issue with the ADML component, which efficiently encodes transcripts in a single decoding pass using multiple domain-specific adapters, leading to consistent improvements of 11.0%, 17.2%, and 10.3% across the music, weather, and sports domains, respectively. Further comparisons of DAS with FT-Multi and LoRA-ft-Multi—two baselines finetuned on a combination of all three TTS datasets—demonstrate

that DAS still maintains superior performance. We further evaluate DAS in a completely out-of-domain setting using the *LibriSpeech test-clean* and *test-other* test sets, which feature vocabulary unrelated to that chatbot language that DAS was adapted to. As demonstrated in Table V, DAS maintains performance comparable to the original Whisper model, with only about a 1% regression in performance. This demonstrates its robustness and resistance to out-of-domain regression.

With the additional LoRA adapters, DAS introduces some extra computations with respect to the orginal auto-regressive decoding process. We benchmark the additional inference latency in Table VI. In particular, we run inference on a V100-32GB GPU on $3K$ randomly sampled utterances drawn from the three evaluation datasets and measure the Real-time-factor (RTF), which is the ratio between the processing time and the audio length. We also evaluate the scalability of DAS for more than three domains by replicating the initial three domains until reaching the desired number of domains. DAS introduces a modest inference latency increase of approximately 9% compared to the original model and demonstrates strong scalability as the number of inference domains grows. However, the computational demands rise with additional domains (e.g., 25 domains), potentially resulting in diminishing performance gains on less powerful hardware. We anticipate that leveraging more advanced GPUs, such as A100 or H100, could alleviate these limitations.

VI. CONCLUSION

In conclusion, we introduce DAS , a novel domain adaptation framework for the pretrained ASR systems with only synthetic data. By employing a single encoder with multiple domain-specific decoders finetuned with Low-Rank Adapters (LoRAs), the framework enables better domain-specific performance while maintaining high efficiency and robustness to out-of-domain regression.

REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[2] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[3] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[4] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[5] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *2017 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2017, pp. 16–23.

[6] Z. Meng, J. Li, Y. Gaur, and Y. Gong, "Domain adaptation via teacher-student learning for end-to-end speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 268–275.

[7] L. Samarakoon, B. Mak, and A. Y. Lam, "Domain adaptation of end-to-end speech recognition in low-resource settings," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 382–388.

[8] K. C. Sim, A. Narayanan, A. Misra, A. Tripathi, G. Pundak, T. N. Sainath, P. Haghani, B. Li, and M. Bacchiani, "Domain adaptation using factorized hidden layer for robust automatic speech recognition." in *Interspeech*, 2018, pp. 892–896.

[9] K. C. Sim, Z. Huo, T. Munkhdalai, N. Siddhartha, A. Stooke, Z. Meng, B. Li, and T. Sainath, "A comparison of parameter-efficient asr domain adaptation methods for universal speech and language models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6900–6904.

[10] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," *Advances in neural information processing systems*, vol. 30, 2017.

[11] X. Zheng, Y. Liu, D. Gunceler, and D. Willett, "Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5674–5678.

[12] Z. Huang, G. Keren, Z. Jiang, S. Jain, D. Goss-Grubbs, N. Cheng, F. Abtahi, D. Le, D. Zhang, A. D'Avirro *et al.*, "Text generation with speech synthesis for asr data augmentation," *arXiv preprint arXiv:2305.16333*, 2023.

[13] D. Le, M. Jain, G. Keren, S. Kim, Y. Shi, J. Mahadeokar, J. Chan, Y. Shangguan, C. Fuegen, O. Kalinli *et al.*, "Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion," *arXiv preprint arXiv:2104.02194*, 2021.

[14] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[15] Z. Wang, C.-L. Li, V. Perot, L. T. Le, J. Miao, Z. Zhang, C.-Y. Lee, and T. Pfister, "Codeclm: Aligning language models with tailored synthetic data," *arXiv preprint arXiv:2404.05875*, 2024.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[17] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[18] Y. Sheng, S. Cao, D. Li, C. Hooper, N. Lee, S. Yang, C. Chou, B. Zhu, L. Zheng, K. Keutzer *et al.*, "S-lora: Serving thousands of concurrent lora adapters," *arXiv preprint arXiv:2311.03285*, 2023.

[19] D. Kalajdzievski, "A rank stabilization scaling factor for fine-tuning with lora," *arXiv preprint arXiv:2312.03732*, 2023.

[20] F. Meng, Z. Wang, and M. Zhang, "Pissa: Principal singular values and singular vectors adaptation of large language models," *arXiv preprint arXiv:2404.02948*, 2024.