Less is Enough: A Target Fine-tuning Strategy for Contextual Biasing Speech Recognition

Sichen Jin *
Samsung Research
chenehk@gmail.com

Shinji Watanabe
Carnegie Mellon University

Abstract

Fine-tuning large pre-trained models is a computationally expensive process that can lead to catastrophic forgetting, hindering their performance on original tasks. Parameter-Efficient Fine-Tuning (PEFT) techniques such as Low-Rank Adaptation (LoRA) have emerged to mitigate these issues. In parallel, the field of Mechanistic Interpretability (MI) is increasingly used to understand the internal workings and find circuits of these large models. This paper combines these two research areas by proposing a novel fine-tuning approach that leverages insights from MI. We demonstrate that for the task of contextual biasing for Automatic Speech Recognition (ASR), a few critical layers are responsible for biasing the model's output towards a given prompt. Our method restricts parameter updates to these specific layers, resulting in a significantly more efficient and simpler fine-tuning process. We show that this targeted approach yields strong performance comparable to more complex methods while preserving the model's original capabilities.

1 Introduction

Parameter-efficient fine-tuning (PEFT) have become an efficient and stable alternative to full-scale fine-tuning. In the realm of large language models (LLM), Low-Rank Adaptation (LoRA) (Hu et al., 2022) is a prominent method that significantly reduces the number of trainable parameters by injecting trainable rank deposition matrices on top of the model's layers. Beyond LLMs, LoRA has recently been adopted for automatic speech recognition (ASR) to facilitate language and task adaptation in larger ASR models (Song et al., 2024; Xu et al., 2024)

However, ASR features a distinct form of PEFT dedicated to integrating text-only data into audio models despite the modality mismatch. Typically, contextual biasing is a task to encourage the model to output infrequent phrases such as contact names or places of interest. Leveraging the modularized structure of ASR models (Chan et al., 2016; Graves, 2012; Gulati et al., 2020), researchers have explored fine-tuning various individual components with text-only data (Pylkkönen et al., 2021; Meng et al., 2022b; Lee et al., 2023) and proposed targeted structural modifications (Meng et al., 2023; Chen et al., 2022; Meng et al., 2024). More recently, with the advent of large-scale attention-based encoder-decoder (AED) models (Radford et al., 2023; Peng et al., 2024, 2025), prompt-based approaches (Li et al., 2024; Suh et al., 2024) have emerged to leverage the language model-like decoder.

In a parallel area of research, mechanistic interpretability (MI) seeks to understand the inner workings of large neural networks. Techniques such as sparse dictionary learning (Bricken et al., 2023; Templeton et al., 2023) and activation interventions (Meng et al., 2022a; Nanda, 2022) are used to project the meaning of internal representations and identify important activation pathways or *circuits* responsible for a specific model behavior.

Recently, these two fields have converged, with MI being applied to demystify the fine-tuning process. For instance, (Wang et al., 2025; Prakash et al., 2024) performs a circuit search on a model before and after fine-tuning to understand how the internal architecture changes. Taking a more proactive approach, (Li et al., 2025) proposes a circuit-aware fine-tuning method that iteratively repeats the graph-pruning and circuit-tuning steps. While this method can be more effective than standard LoRA, it significantly complicates the optimization process by requiring an extra forward pass to dynamically determine the relevant subgraph during fine-tuning.

In this paper, we propose a novel and straight-

*This work was done separately from the work at Samsung.

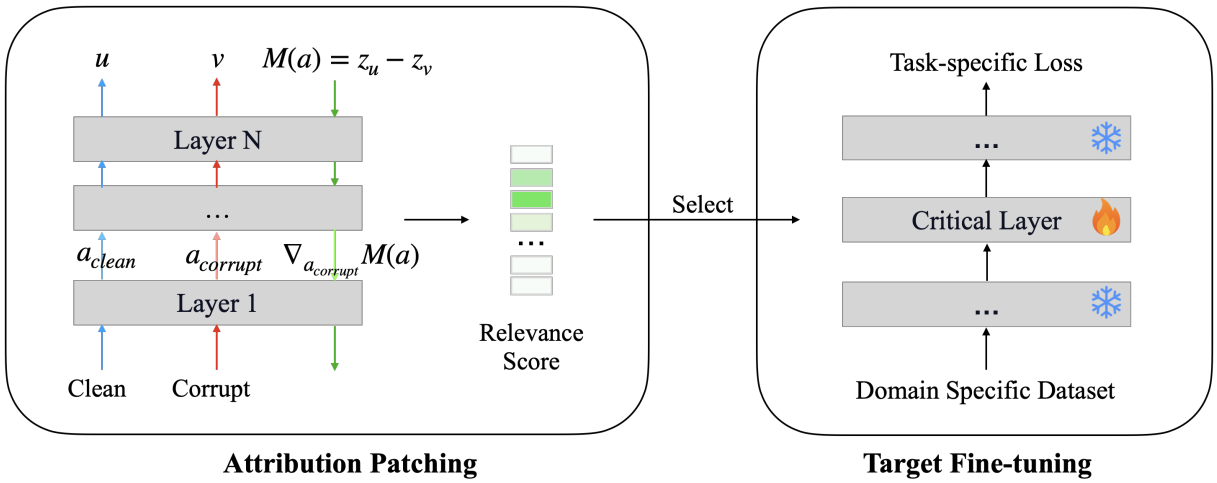

Figure 1: A simple depiction of Target Fine-tuning: (1) Identify the critical layers for a certain task with attribution patching. The red and blue arrows mean two forward passes with the *clean* and *corrupt* inputs, and the green arrows mean the backward pass for the corrupt input. (2) Execute Target Fine-tuning on the task only updating the parameters in the critical layers.

forward approach to effective fine-tuning for the task of contextual biasing. We show that for this task, a few specific decoder layers play a critical role in biasing the model's output toward a given prompt. Our work demonstrates that restricting parameter updates exclusively to these layers not only produces a simple and efficient fine-tuning process but also yields performances comparable to more complex methods. This approach differs from the previous MI based fine-tuning in that we force the model to update the existing circuit rather than exploring and forming new ones.

Our contribution comes as twofold: First, we identify the existence of the critical layers for contextual biasing for the first time. Second, we demonstrate that fine-tuning a few pre-defined layers is sufficient for simple tasks with obvious critical layers.

2 Related Work

2.1 contextual biasing

Attention-based encoder-decoder (AED) models have become one of the best-performing architectures for large-scale speech recognition and various downstream tasks. With speech features X over time as the input, a transformer encoder processes these features to generate a sequence of high-level acoustic representations H. A language model-like decoder then autoregressively generates the output transcript. At each decoding step u, the decoder takes the previous decoder state $\mathbf{s}_{u-1}$ produced from the previous outputs $y_{1:u-1}$, and a context vector $\mathbf{c}_{u-1}$ to calculate the output distribution for the next token y_u. The context vector is calculated through a cross-attention mechanism, $\mathrm{Context}$, where the query is the decoder state $\mathbf{s}_{u-1}$ and the keys and values are the acoustic representations H. The overall process is described as follows:

$$\begin{aligned} H &= \mathrm{Encoder}(X), \\ \mathbf{c}_u &= \mathrm{Context}(\mathbf{s}_{u-1}, H), \\ \mathbf{s}_u &= \mathrm{Decoder}(\mathbf{s}_{u-1}, \mathbf{c}_{u-1}), \\ P(y_u|X, y_{1:u-1}) &= \mathrm{Output}(\mathbf{s}_u), \end{aligned} \tag{1}$$

where Output is a projection layer with softmax outputs over tokens.

The architecture of the transformer decoder allows AED models to utilize prompting for dynamically adjusting the recognition output as a technique for contextual biasing. In this case, the tokens of the rare word (e.g. contact names and song names) are given to the decoder as a prompt P before the start-of-sentence token. This effectively primes the model's initial state with the desired context, which is defined as:

$$\mathbf{s}_1 = \mathrm{Decoder}(P) \tag{2}$$

2.2 Attribution Patching

Attribution patching (AP) (Meng et al., 2022a) is used to quantify the causal importances of the activations $\mathbf{a}^l$ in all intermediate layers l on the final

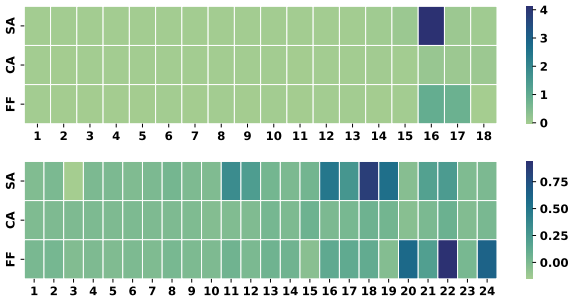

Figure 2: The *causal importance score* of each activataion of the layers in OWSM v3.1 (top) and Whisper medium (bottom) models. The horizontal axis shows different transformer blocks and the vertical axis shows three major layers in them: self-attention (SA), cross-attention (CA) and feed-forward (FF) layers.

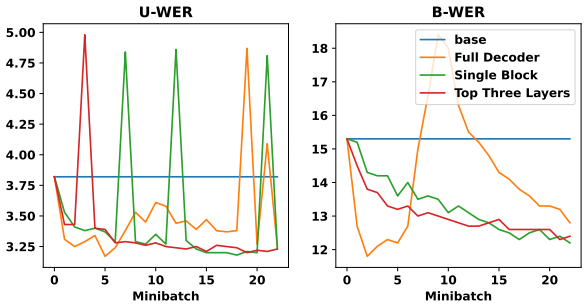

Figure 3: B-WER and U-WER on Librispeech test-other set measured along the fine-tuning process of OWSM v3.1 on three parameter update scopes. The performance glitches in U-WER are from the repetition of the last words.

model output y. The methodology establishes a contrast by using two inputs: a *clean* input X_{cln}, which yields the expected output $y_{cln} = u$ and clean activations $\mathbf{a}^l_{cln}$, and a *corrupt* input X_{cor}, which produces a different output $y_{cor} = v$ and corrupt activations $\mathbf{a}^l_{cor}$.

To identify the activations responsible for the change, a gradient is calculated for each activation vector based on a metric $M(y_{cor})$. The most common metric is the difference in logit scores between the desired output u and the corrupt output v, $M(y_{cor}) = z^u_{cor} - z^v_{cor}$, where $\mathbf{z}_{cor}$ is the pre-softmax logit vector and z^u is the logit score for token u.

The causal importance score $S(\mathbf{a}^l)$ for a specific layer's activation is defined as the dot product of the difference between the clean and corrupt activation vectors and the gradient of the metric M with respect to the activations:

$$S(\mathbf{a}^l) = (\mathbf{a}^l_{cln} - \mathbf{a}^l_{cor})^T \nabla_{\mathbf{a}_{cor}} M(y_{cor}), \quad (3)$$

3 Proposed Method

We propose Target Fine-tuning (TFT), a simple yet effective fine-tuning methodology that restricts the parameter update to a small group of task-critical parameters. The process is defined by two sequential steps as shown in Fig. 1: (1) Identifying the layers whose activations are most causally relevant to the target task's behavior and (2) executing fine-tuning only updating the parameters within the identified critical layer, freezing all other layers in the model.

For the task of contextual biasing, the clean and corrupt inputs for calculating the intermediate activations $\mathbf{a}^l$ from Eq. 3 were defined as follows: A specific example is selected as the clean input (X_{cln}) if the ASR model correctly transcribes a rare word when the biasing prompt is included, but fails to transcribe it without the prompt. The corresponding corrupt input (X_{cor}) is then constructed by removing the rare word from the prompt provided to the model.

We consider the three major layers within the Transformer decoder block (defined in Eq. 1) as the minimal units for our fine-tuning analysis, namely the self-attention (SA) layer, the cross-attention (CA) layer, and the feed-forward (FF) layer. As the causal importance scores $S(a^l)$ are measured for each layer, the layers within top K scores are used for fine-tuning.

4 Experiment Results

We demonstrate the efficacy of the proposed method for the task of contextual biasing. After some investigation, we found that this task relies on very few layers to perform contextual biasing, and further proved that this phenomenon was common across different different training methods by employing two prominent open-source ASR models, Whisper medium(Radford et al., 2023) and OWSM v3.1 (Peng et al., 2024). All experiments were conducted using the read English based Librispeech corpus (Panayotov et al., 2015) as the fine-tuning and evaluation dataset. The ESPnet framework (Watanabe et al., 2018) was utilized for all training and decoding procedures. Following the setup in (Le et al., 2021), we defined the words that fall out of the top 5,000 most common words in the audio training set as rare words. 100 distractors were added along with the rare words and provided as the prompt P from Eq. 2 during inference. For performance evaluation, we measured the unbiased word error rate (U-WER) and the biased one, B-WER, where U-WER was measured over the ex-

	Target Component	#Params	Test-clean		Test-other	
			U-WER	B-WER	U-WER	B-WER
OWSM v3.1	Baseline	-	1.58	7.30	3.82	15.3
	Full Decoder (best)	315M	1.23	5.67	3.25	**11.8**
	Full Decoder (full epoch)	315M	1.26	6.17	3.25	12.8
	Single Block (layer 16)	16.8M	1.28	5.58	**3.23**	12.2
	TFT ($K = 3$)	20.9M	**1.19**	**5.36**	**3.23**	12.3

Table 1: The final performances of fine-tuning. Training the full decoder reached the best performance quickly but exhibited instability with rising WER, whereas the Target fine-tuning results show comparable performance with updates on much less parameters.

amples without a single rare word and B-WER was measured for the rest of the examples.

4.1 The Critical Layers

In order to identify the critical layers for contextual biasing, we executed attribution patching on the Librispeech dev-clean and dev-other subsets and used the average causal importance score of each activation over the examples.

The Attribution Patching analysis revealed that the critical function of contextual biasing was highly concentrated in a few specific layers located near the final output layer of the decoder stack. The scores for the Whisper model were notably more distributed compared to OWSM. After careful investigations on the error patterns, we found that this difference stemmed from the different text normalization procedures, and Whisper distributed the contextual biasing process across more layers to handle the prompt words with different alphabet casings as the model output. Consequently, we selected the OWSM v3.1 model for the subsequent Target Fine-Tuning experiments to ensure a focused evaluation free from secondary effects.

4.2 Target Fine-tuning

We fine-tuned the model on 1 epoch of the full 960-hour Librispeech training set on one V100 GPU for 23 hours. Similar to the causal analysis tests, we randomly chose and inserted 100 rare words into the transcripts to serve as the prompt for contextual biasing, which is exposed during fine-tuning as per the training strategy of OWSM v3.1.

We conducted a comparative study by freezing all parameters in the encoder and evaluating three distinct parameter update scopes and tracked the word error rates (WER):

- Full Decoder: Fine-tuning the entire decoder module.
- Single Block (layer 16): Updating a single decoder block containing the top causal importance score layer.
- TFT ($K = 3$): Updating K layers exhibiting the highest relevance scores.

As shown in Figure 3, training the Full Decoder quickly reached the best performance, but exhibited significant training instability with a sudden rise of B-WER before the entire epoch was completed. Updating an entire decoder block, a method often chosen heuristically, showed a continuous improvement in B-WER but demonstrated unstable U-WER by introducing hallucination (repeating the same token), likely due to mis-training in the CA layer. In contrast, the proposed method showed stable improvements on both U-WER and B-WER.

The final results detailed in Table 1 further suggests that the proposed method can achieve performance gains efficiently without the degradation of general ASR capability. TFT recorded the lowest U-WER on both test sets, confirming the targeted parameter update is beneficial for maintaining the model's original general purpose performance. For the contextual biasing task, TFT demonstrated superior performance by achieving the lowest Biased Word Error Rate (B-WER) on Test-clean and a comparable result on Test-other, proving that effective fine-tuning is achievable despite updating significantly fewer parameters. This confirms that restricting parameter updates only to the critical layers effectively mitigates catastrophic forgetting while successfully tuning the specific task circuit.

5 Conclusion

In this paper, we introduced Target Fine-tuning, a novel parameter-efficient fine-tuning approach that leverages attribution patching to identify the crucial components for a given task and restrict the updates to them. The experiments on the task of

contextual biasing demonstrates that the proposed method achieved comparable performance despite training on much less parameters.

Acknowledgments

AI Assistants were used purely with the language of the paper.

Limitations

We demonstrated the efficacy of Target Fine-tuning for a task with a concentrated causal importance in a single or a few layers. Further investigations need to be done on the following matters.

- The scope of this work is limited to tasks with simple circuits. It is worth exploring how to separate the primary circuit related to the target task from a distributed functionality like the one in Whisper mentioned in Section 4.1. Another subsequent task can be comparing models with concentrated and distributed causal importances and find training strategies that encourage or suppress this phenomenon.
- This work lacks mathematical analysis on the capacity and bounds of performance improvement achievable by the target method. This theoretical grounding is necessary to mathematically determine the inherent complexity of the task-critical circuit, allowing us to precisely adjust the trade-off between simplicity and performance.

References

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, and 5 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964.

Xie Chen, Zhong Meng, Sarangarajan Parthasarathy, and Jinyu Li. 2022. Factorized neural transducer for efficient language model adaptation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8132–8136. IEEE.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, Yatharth Saraf, and Michael L. Seltzer. 2021. Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion. In *Interspeech 2021*, pages 1772–1776.

Kyungmin Lee, Haeri Kim, Sichen Jin, Jinhwan Park, and Youngho Han. 2023. A more accurate internal language model score estimation for the hybrid autoregressive transducer. In *Interspeech 2023*, pages 869–873.

Yuang Li, Yinglu Li, Min Zhang, Chang Su, Jiawei Yu, Mengyao Piao, Xiaosong Qiao, Miaomiao Ma, Yanqing Zhao, and Hao Yang. 2024. CB-whisper: Contextual biasing whisper using open-vocabulary keyword-spotting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2941–2946, Torino, Italia. ELRA and ICCL.

Yueyan Li, Wenhao Gao, Caixia Yuan, and Xiaojie Wang. 2025. Fine-tuning is subgraph search: A new lens on learning dynamics. *Preprint*, arXiv:2502.06106.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Zhong Meng, Tongzhou Chen, Rohit Prabhavalkar, Yu Zhang, Gary Wang, Kartik Audhkhasi, Jesse Emond, Trevor Strohman, Bhuvana Ramabhadran, W Ronny Huang, and 1 others. 2023. Modular hybrid autoregressive transducer. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 197–204. IEEE.

Zhong Meng, Yashesh Gaur, Naoyuki Kanda, Jinyu Li, Xie Chen, Yu Wu, and Yifan Gong. 2022b. Internal language model adaptation with text-only data for end-to-end speech recognition. In *Interspeech 2022*, pages 2608–2612.

Zhong Meng, Zelin Wu, Rohit Prabhavalkar, Cal Peyser, Weiran Wang, Nanxin Chen, Tara N. Sainath, and

Bhuvana Ramabhadran. 2024. Text Injection for Neural Contextual Biasing. In *Interspeech 2024*, pages 2985–2989.

Neel Nanda. 2022. Attribution patching: Activation patching at industrial scale.

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

Yifan Peng, Muhammad Shakeel, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. 2025. OWSM v4: Improving Open Whisper-Style Speech Models via Data Scaling and Cleaning. In *Interspeech 2025*, pages 2225–2229.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee weon Jung, and Shinji Watanabe. 2024. OWSM v3.1: Better and Faster Open Whisper-Style Speech Models based on E-Branchformer. In *Interspeech 2024*, pages 352–356.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *Proceedings of the 2024 International Conference on Learning Representations*. ArXiv:2402.14811.

Janne Pylkkönen, Antti Ukkonen, Juho Kilpikoski, Samu Tamminen, and Hannes Heikinheimo. 2021. Fast text-only domain adaptation of rnn-transducer prediction network. In *Interspeech 2021*, pages 1882–1886.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

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

Jiwon Suh, Injae Na, and Woohwan Jung. 2024. Improving Domain-Specific ASR with LLM-Generated Contextual Descriptions. In *Interspeech 2024*, pages 1255–1259.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, and 7 others. 2023. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. 2025. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. In *Interspeech 2018*, pages 2207–2211.

Tianyi Xu, Kaixun Huang, Pengcheng Guo, Yu Zhou, Longtao Huang, Hui Xue, and Lei Xie. 2024. Towards Rehearsal-Free Multilingual ASR: A LoRA-based Case Study on Whisper . In *Interspeech 2024*, pages 2534–2538.