# DISCRIMINATIVE SPEECH RECOGNITION RESCORING WITH PRE-TRAINED LANGUAGE MODELS

*Prashanth Gurunath Shivakumar, Jari Kolehmainen, Yile Gu, Ankur Gandhe, Ariya Rastrow, Ivan Bulyko*

Amazon Alexa AI, USA

{psshvak,jkolehm,yilegu,aggandhe,arastrow,ibbulyko}@amazon.com

## ABSTRACT

Second pass rescoring is a critical component of competitive automatic speech recognition (ASR) systems. Large language models have demonstrated their ability in using pre-trained information for better rescoring of ASR hypothesis. Discriminative training, directly optimizing the minimum word-error-rate (MWER) criterion typically improves rescoring. In this study, we propose and explore several discriminative fine-tuning schemes for pre-trained LMs. We propose two architectures based on different pooling strategies of output embeddings and compare with probability based MWER. We conduct detailed comparisons between pre-trained causal and bidirectional LMs in discriminative settings. Experiments on LibriSpeech demonstrate that all MWER training schemes are beneficial, giving additional gains upto 8.5% WER. Proposed pooling variants achieve lower latency while retaining most improvements. Finally, our study concludes that bidirectionality is better utilized with discriminative training.

***Index Terms—*** Minimum word error rate, Discriminative training, GPT-2, BERT, ASR Rescoring

## 1. INTRODUCTION

End-to-end speech recognition systems have seen tremendous advancements in making speech recognition more accurate and closing the gap to human levels. However, large models with all neural end-to-end training requires large amount of transcribed speech data to achieve better performance. Availability of labelled speech data is always limited compared to the largely, freely available text data. Meanwhile, the recent advancements in computational efficiency has enabled scaling language models. The large language models (LLM) have demonstrated their ability to ingest vast amount of text data to match almost human level performance in many benchmark tasks. Exploiting knowledge from these LMs to improve second pass rescoring has become particularly attractive.

Several prior works have explored employing pre-trained models such as GPT, BERT for rescoring [1, 2, 3, 4, 5, 6, 7, 8, 9]. [1] conducted extensive study for rescoring with GPT models. [2] proposed using BERT for utterance level ASR rescoring and demonstrated the importance of bi-directional encoding. Pseudo-log-likelihood (PLL) for each n-best hypothesis is computed by summing the log-likelihoods obtained from the model by masking each token position in the input sequence, one-by-one. The PLL is interpolated with the first pass score to obtain the final n-best rank. Given that the PLL computation is expensive, [4] further extended BERT for mask-less scoring where the [CLS] token is fine-tuned using L2-loss to predict the PLL. This dramatically reduces the computation and makes BERT-based rescoring practically feasible. [4] conducted a detailed comparison of GPT-2 and BERT models for rescoring which demonstrates the effectiveness and advantages of bi-directional encoding of BERT models for rescoring. [3] conducted empirical study of GPT-2 and BERT models for rescoring on state-of-the-art first pass model and evaluates the effect of additional context. LLMs ranging from 70M to 540B parameters for second pass rescoring is investigated in [6] with findings of significant wins over strong first pass model.

On the other hand, several works have demonstrated the effectiveness of discriminative training with MWER criterion [10, 11, 12, 13, 14, 7, 8, 9]. [10, 11] proposed MWER loss for training end-to-end ASR with encoder-decoder architectures. [12] trained RNN-LMs with MWER for rescoring and achieved improvements over using log-likelihoods. [13] used MWER training with LSTM language models attending to audio embeddings from first-pass encoder and showed improvement over cross-entropy based training. In [14], a transformer based deliberation rescorer attending to first pass audio encoding and ASR hypothesis encoding is trained with MWER criteria and reported gains of 9% in WER.

Despite the advantages of MWER training, there are very few studies on adding MWER training to further improve rescoring performance of a pre-trained model. [5] proposed MWER training for BERT models, which is the first work that combines the importance of large pre-training with discriminative learning. The work builds upon the findings from [2] and [4] by proposing a technique that uses a pre-trained BERT model with a feed-forward layer on the [CLS] embedding similar to [4] and further extends by fine-tuning with MWER loss. Significant gains are reported with discriminative training over the n-best hypotheses [5, 7, 8, 9]. The scheme proposed in [5] is an approximation and deviates from

979-8-3503-0689-7/23/$31.00 

typical sequence probability based computation in favor of latency.

However, there lacks a work on combining MWER training with a pre-trained causal LM, and a comparison with its bidirectional counterpart. Although bi-directional encoding could be important, causal language models have certain advantages with latency and also naturally fits into streaming frameworks that make it attractive. For example, with causal models, re-scoring need not always wait until the completion of first pass. Moreover, promising performance of causal language models with zero-short learning [15] and scaling [16] attracts interest in its applicability to rescoring [17].

In this study, we attempt to bridge this gap in understanding how to effectively combine MWER training with both causal and bidirectional pre-trained models. The contribution of this paper are in two folds. First, we propose three different ways to fine-tune LLM with MWER criterion. We explore three techniques: (i) using last-token embedding from GPT-2, (ii) attention pooling on all tokens, and (iii) sequence probability based MWER criteria [13]. Second, we perform a detailed comparison between pre-trained causal language model, i.e., GPT-2, and bidirectional model, i.e., BERT to assess the role of bidirectional encoding in ASR rescoring in MWER discriminative training setting. To the best of our knowledge this is the first work to explore pretrained, decoder-only transformer based causal language models with MWER training for ASR rescoring.

The rest of the paper is organized as follows: Section 2 describes the proposed techniques for discriminative training GPT-2 models with MWER loss. Section 4 presents the datasets and the experimental setup. Results are discussed in Section 5 and finally conclusions are drawn in Section 6.

## 2. NON-DISCRIMINATIVE ASR RESCORING

Likelihood scores from pre-trained language models can be natively used to rescore ASR n-best by interpolating with the first pass acoustic model scores:

$$s_i = logP_{LM}(y_i) + \lambda logP_{AM}(x|y_i) \tag{1}$$

where $P_{AM}(x|y_i)$ is the sequence probability of the 1st pass acoustic model, given an input audio sequence, $x$, for $i^{th}$ ASR hypothesis, $y_i$, $P_{LM}(y_i)$ is the likelihood of the rescoring language model and $\lambda$ is the interpolation weight.

### 2.1. GPT-2

In case of GPT-2, the likelihood for a sequence of length $L$ is computed in a auto-regressive fashion by modeling:

$$P_{LM}(y_i) = \prod_{t=1}^{L} P(y_{i,t}|y_{i,1}, \ldots, y_{i,t-1}) \tag{2}$$

using a soft-max over the entire vocabulary.

### 2.2. BERT

In contrast, bi-directional encoding models such as BERT use masked language modeling where log likelihood for sentences are not directly available. However, several works have demonstrated pseudo-log-likelihood (PLL) scores are a viable replacement [2] for rescoring. PLL is computed as

$$PLL(y) = -\sum_{t=1}^{L} \log P(y_t|y_{\setminus t}) \tag{3}$$

where $y_{\setminus t} = \{y_1 \ldots, y_{t-1}, [MASK], y_{t+1}, \ldots, y_L\}$. However, this comes with the expense of increased computation and memory since it requires copies and inference passes equal to the length of the sequence (see Fig.1b).

## 3. DISCRIMINATIVE ASR RESCORING

Authors in [10] propose to directly minimize expected word error rate while training sequence-to-sequence attention based models for ASR. Given an input audio sequence $x$ with ground-truth transcript $y^*$ and ASR hypothesis $y$, MWER loss can be computed as:

$$L_{mwer}(x, y^*) = E[\mathcal{E}(y, y^*)] = \sum_{y} P(y|x)\mathcal{E}(y, y^*) \tag{4}$$

where $\mathcal{E}(y, y^*)$ is the edit distance between the ASR hypothesis, $y$, and groundtruth transcript, $y^*$, $P(y|x)$ is the probability of ASR hypothesis $y$ given the input audio sequence $x$. To make the loss tractable, the expectation is approximated by restricting the sequence probability over the n-best hypothesis.

$$L_{mwer}(x, y^*) = \sum_{i=1}^{N} P(y_i|x)\mathcal{E}(y_i, y^*) \tag{5}$$

where $N$ refers to N-best ASR hypotheses.

### 3.1. Sequence probability based MWER

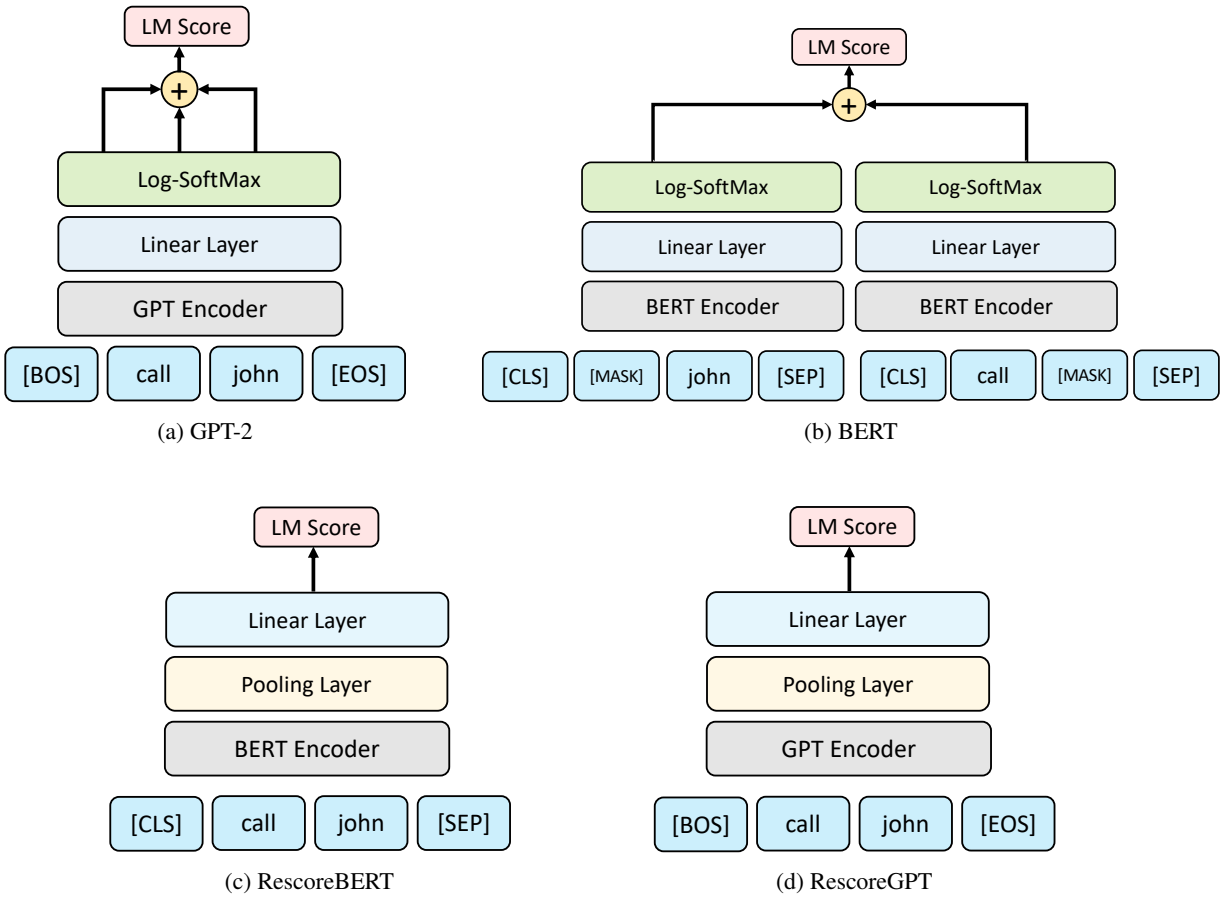
The above MWER loss for discriminatively training can be applied for 2nd pass rescoring model [13]. For rescoring, the n-best posterior probability $P(y_i|x)$ is computed by considering a linear combination of first pass acoustic model score and the LM, $s_i$:

$$P(y_i|x) = \frac{e^{s_i}}{\sum_{j=1}^{N} e^{s_j}} \tag{6}$$

#### *3.1.1. GPT + MWER*

Similarly, we can use GPT-2 to compute MWER loss as follows:

$$L_{mwer}(x, y^*) = \sum_{i=1}^{N} \frac{e^{s_i}}{\sum_{j=1}^{N} e^{s_j}} \mathcal{E}(y_i, y^*) \tag{7}$$




(a) GPT-2 (b) BERT

(c) RescoreBERT (d) RescoreGPT

**Fig. 1**: Illustration of model architectures (a): GPT-2 based rescoring where token probabilities are predicted for tokens after beginning of sentence by Eq. (2), (b): BERT rescoring where each token is predicted by the masked sequence according to Eq. (3), (c) RescoreBERT that uses pooling layer on the BERT encoder and predicts a score using a linear layer, (d): RescoreGPT model that uses a pooling layer on the GPT-2 encoder and a linear layer to predict score. Note, for architectures (c) and (d), the pooling layer is either CLS embedding or attention pooling over all token embeddings.

and

$$s_i = log(\prod_{t=1}^{L} P(y_{i,t}|y_{i,1}, \ldots, y_{i,t-1})) + \lambda log P_{AM}(x|y_i) \quad (8)$$

where $s_i$ and $s_j$ are computed by plugging Eq. (2) into Eq. (1). The model architecture is depicted in Fig. 1a.

#### *3.1.2. BERT + MWER*

In case of BERT, we propose to use Eq. (3) as LM-score, i.e., $s_i$ is given by:

$$s_i = -\sum_{t=1}^{L} \log P(y_t|y_{\setminus t}) + \lambda \log P_{AM}(x|y_i) \quad (9)$$

which is plugged into Eq. (7) for MWER loss computation. The model architecture is depicted in Fig. 1b.

### 3.2. Embedding based MWER Rescoring

As demonstrated in [4] and [5] for BERT, an alternative way to compute a score for rescoring is to extract [CLS] embedding for the sentence and add a linear layer. We name these model architectures RescoreBERT and RescoreGPT.

#### *3.2.1. RescoreBERT*

[4] showed that the PLLs can be approximated by fine-tuning the classification ([CLS]) token via L2 regression. Building upon this, [5] proposed MWER training over the [CLS] token, i.e.,

$$\log P_{LM}(y_i) = W_F(BERT_{CLS}(y_i)) \quad (10)$$

where $BERT_{CLS}$ refers to the BERT [CLS] embedding and $W_F$ is a learnable feed-forward layer. The model architecture is illustrated in Fig. 1c.

#### 3.2.2. RescoreGPT

In case of GPT-2 models, we propose to use the probability computation on the last token for MWER training. The motivation here is that by considering the last token embedding, the model has seen all the tokens and has similar information as BERT model. Thus, we propose to use the following for MWER computation:

$$\log P_{LM}(y_i) = W_F(GPT_{last}(y_i)) \tag{11}$$

Note, this has advantages in reducing the memory and computational footprint during training and inference, drastically, since soft-max is computed over n-best and not over the vocabulary, and only once per utterance. The architecture is illustrated in Fig. 1d, where the pooling layer is the last token embedding.

#### 3.2.3. Attention Pooling

A natural extension of the above approach is to make use of all the output embeddings from GPT instead of only the embedding corresponding to the last token. One way to achieve this is by attention pooling over the GPT output embeddings, i.e.,

$$\log P_{LM}(y_i) = W_F(\text{softmax}(\hat{Q}K^T/\sqrt{d_k})V) \tag{12}$$

and

$$\hat{Q} = \hat{q}W_Q \qquad K = HW_K \qquad V = HW_V$$

where $H = \{h_1, \ldots, h_L\}$ represent hidden output embeddings, $W_Q$, $W_K$, $W_V$ are learnable weights associated with query, key, values respectively, and $\hat{q}$ is a learnable vector. This can be adopted similarly in case of both GPT and BERT encoders as shown in Fig. 1d and 1c.

## 4. DATASETS AND EXPERIMENTAL SETUP

### 4.1. Data

Publicly available LibriSpeech corpus is employed for experimentation. The dataset comprises approximately 1000 hours of read English speech derived from audiobooks from LibriVox [18]. We also make use of Librispeech LM corpus based on Project Gutenbert books for domain adaptation [18].

### 4.2. Experimental Setup

#### 4.2.1. First-pass ASR

For all our experiments, we use Whisper tiny model [19]. The architecture of the Whisper first-pass is based on Transformer sequence-to-sequence models [20]. Whisper-tiny comprises 39M parameters and is trained in a supervised manner on 680k hours of audio collected from internet. We generate top-10 hypothesis for second-pass rescoring purposes. The hypothesis is post-processed by removing punctuation and converting case according to target LibriSpeech transcripts. The oracle WER of 10-best on test-clean and test-other is 3.63% and 9.22% respectively.

#### 4.2.2. Second-pass Rescoring

We use the publicly available GPT-2 model [21] and bert-base-cased as the BERT [22] model for all the experiments. Since we plan to compare the role of bidirectionality, special care is taken to ensure that the two models are similarly sized (110M parameters in BERT; 117M parameters in GPT-2), pre-trained on comparable data-sets (BERT on BooksCorpus and Wikipedia; GPT-2 on WebText) and the experimental setting is as similar as possible, following [4].

We domain adapt both BERT and GPT-2 models on the Librispeech LM corpus for all our experiments. Next, MWER training is performed as described in Section 2. Finally, for rescoring, we tune the optimal interpolation weight $\lambda$ in Eq. (1) on the corresponding development sets.

Batch size of 16 is used during domain adaptation and batch size of 4 is used during MWER training. Learning rate schedule is used during training with an initial learning rate of $1e-5$. Optimal checkpoint is picked on the corresponding dev-sets and the results on unseen test-sets are reported.

### 4.3. Baseline

In this study, we employ two baselines: (i) the first-pass, tiny-Whisper model, (ii) Log-likelihood based rescoring for GPT-2, PLL based rescoring for BERT. [2]. For vanilla log-likelihood and PLL rescoring baselines, the BERT and GPT-2 models are domain adapted first on librispeech LM corpus and then on audio corpus for fair comparison.

## 5. RESULTS

Table 1 presents the experimental results of the baselines and proposed MWER discriminative training schemes on GPT-2 and BERT models. The Whisper first pass gives 5.67% and 12.88% WER on the test-clean and test-other sets of LibriSpeech. Typical log-likelihood based rescoring provides upto 15.9% improvement over the baseline for GPT-2, after domain adaptation on LibriSpeech. PLL based rescoring [2] with BERT outperforms GPT-2 giving upto 16.9% relative improvement over the first-pass. Improvements with BERT (1.26%) relative to GPT-2 can be attributed to the bi-directional encoding and is consistent with [2, 4].

All configurations of discriminative training improves upon the log-likelihood baseline systems. GPT-2 with MWER fine-tuning provides a relative WER improvement of up-to 5.9%. Similarly, BERT with PLL based MWER fine-tuning

| Model | Loss | Latency [ms] | test-clean | test-other |
|---|---|---|---|---|
| Whisper [19] | - | - | 5.67 | 12.88 |
| GPT-2 | CE | 49 | 4.77 (15.87%) | 11.19 (13.12%) |
| BERT [2] | CE | 6115 | **4.71 (16.93%)** | 11.19 (13.12%) |
| GPT-2 | MWER | 49 | 4.52 (20.34%) | 10.89 (15.45%) |
| GPT-2 | MWER + CE | 49 | 4.49 (20.79%) | 10.90 (15.37%) |
| BERT | MWER | 6115 | 4.36 (23.10%) | 10.93 (15.14%) |
| BERT | MWER + CE | 6115 | **4.31 (23.99%)** | 10.81 (16.07%) |
| RescoreGPT | MWER | 34 | 4.72 (16.75%) | 11.03 (14.36%) |
| with Attn. pooling | MWER | 35 | 4.65 (17.99%) | 11.60 (13.35%) |
| RescoreBERT [5] | MWER | 33 | 4.37 (22.93%) | 10.80 (16.15%) |
| with Attn. pooling | MWER | 33 | 4.38 (22.75%) | **10.78 (16.30%)** |
| 10-best Oracle | - | - | 3.63 (35.98%) | 9.22 (28.42%) |

**Table 1**: WER results for different models. Relative % WER-reduction is shown in the brackets respect to the 1st pass model. Latency numbers present average time spent in milliseconds for rescoring 10 hypothesis of sequence length 64 for various models on a single NVIDIA T4 GPU.

provides up-to 8.49%. It is interesting to note that the gap between GPT-2 and BERT after MWER training is increased from 1.26% to 4.01% relative. This suggests that the discriminative training can exploit the bi-directional information more efficiently. In addition to the MWER loss objective, we also experiment using linear combination of MWER loss and cross-entropy (CE) loss (MWER $+ \alpha \times$ CE) to stabilize training as suggested in [13, 11] ($\alpha = 0.01$ tuned over $\{0.01, 0.1, 1\}$). While we did not observe significant gains in combining the cross-entropy objective along with the MWER loss for GPT-2, we notice improvements in case of the BERT.

Further, RescoreGPT model gives up-to 1.4% improvements over the baseline GPT-2 model. RescoreBERT [5] gives relative improvements of up-to 7.21% in comparison with the baseline BERT model. This finding suggests (i) using the last-token embedding for rescoring in-case of causal language models is less effective, (ii) bi-directional encoding and [CLS] representation of BERT plays an important role in discriminative training. This also suggests that including sequence probability in MWER computation is crucial in case of generative, causal language models like GPT-2. For attention pooling variants, in both the cases, i.e., RescoreBERT and RescoreGPT, we do not observe significant gains.

We also perform latency profiling to demonstrate the advantages and disadvantages of each of the explored architectures. Table 1 shows model rescoring latency for ten random hypothesis with fixed length of 64 tokens. Rescoring latency was computed using pytorch in eager execution mode with a single NVIDIA T4 GPU. As expected, using PLLs from BERT model to rescore the hypothesis takes two magnitudes more time than other rescoring models. This is essentially due to the need to evaluate the model for each token. GPT-2 has roughly 40% higher latency than the pooled counterparts (RescoreGPT and RescoreBERT). This added latency can be attributed to the evaluation of the vocabulary softmax demonstrating that predicting the score directly can cut off the latency significantly even for 100 million parameter models. It is also worthwhile to note that some portion of the GPT model's latency could be reduced by inferring the model during the 1st pass decoding itself, which cannot be done with any bidirectional model (e.g. BERT).

## 6. CONCLUSION

In this work, we proposed and explored several discriminative training techniques (based on MWER criterion) in application to ASR second-pass rescoring designed for pre-trained language models, particularly GPT-2 and BERT. We explored three configurations (i) typical sequence probability based MWER loss, (ii) tuning last token embedding with MWER loss, and (iii) attention pooling of all the output embeddings. Experiments were conducted using publicly available dataset (LibriSpeech) and models. Comparisons are carried out to assess the role of bi-directional encoding and its relevance to discriminative training. Our results suggests that using the last token embedding of GPT-2 model is not as effective as using [CLS] token in BERT models. However, we demonstrate that with sequence probability based MWER training of GPT-2 model, the gap is much closer to the BERT counterpart. We find that discriminative training helps exploit the bidirectional information in a better way for rescoring. Clear advantages in terms of latency is demonstrated for pooling techniques with trade-offs for WER. In future, we plan to scale the size of rescoring models and check the applicability of zero-shot prompting with discriminative models.

## 7. REFERENCES


[1] Hongzhao Huang and Fuchun Peng, "An empirical study of efficient asr rescoring with transformers," *arXiv preprint arXiv:1910.11450*, 2019.

[2] Joonbo Shin, Yoonhyung Lee, and Kyomin Jung, "Effective sentence scoring method using bert for speech recognition," in *Asian Conference on Machine Learning*. PMLR, 2019, pp. 1081–1093.

[3] Takuma Udagawa, Masayuki Suzuki, Gakuto Kurata, Nobuyasu Itoh, and George Saon, "Effect and analysis of large-scale language model rescoring on competitive asr systems," *arXiv preprint arXiv:2204.00212*, 2022.

[4] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff, "Masked language model scoring," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2699–2712.

[5] Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko, "Rescorebert: Discriminative speech recognition rescoring with bert," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6117–6121.

[6] Tongzhou Chen, Cyril Allauzen, Yinghui Huang, Daniel Park, David Rybach, W Ronny Huang, Rodrigo Cabrera, Kartik Audhkhasi, Bhuvana Ramabhadran, Pedro J Moreno, et al., "Large-scale language model rescoring on long-form data," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] Prashanth Gurunath Shivakumar, Jari Kolehmainen, Yile Gu, Ankur Gandhe, Ariya Rastrow, and Ivan Bulyko, "Distillation strategies for discriminative speech recognition rescoring," 2023.

[8] Jari Kolehmainen, Yile Gu, Aditya Gourav, Prashanth Gurunath Shivakumar, Ankur Gandhe, Ariya Rastrow, and Ivan Bulyko, "Personalization for bert-based discriminative speech recognition rescoring," 2023.

[9] Yile Gu, Prashanth Gurunath Shivakumar, Jari Kolehmainen, Ankur Gandhe, Ariya Rastrow, and Ivan Bulyko, "Scaling laws for discriminative speech recognition rescoring models," 2023.

[10] Rohit Prabhavalkar, Tara N Sainath, Yonghui Wu, Patrick Nguyen, Zhifeng Chen, Chung-Cheng Chiu, and Anjuli Kannan, "Minimum word error rate training for attention-based sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4839–4843.

[11] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al., "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[12] Takaaki Hori, Chiori Hori, Shinji Watanabe, and John R. Hershey, "Minimum word error training of long short-term memory recurrent neural network language models for speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5990–5994.

[13] Ankur Gandhe and Ariya Rastrow, "Audio-attention discriminative language model for asr rescoring," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7944–7948.

[14] Ke Hu, Ruoming Pang, Tara N Sainath, and Trevor Strohman, "Transformer based deliberation for two-pass speech recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 68–74.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[16] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[17] Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu, "Prompting large language models for zero-shot domain adaptation in speech recognition," 2023.

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

[19] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.