Interspeech 2024
1-5 September 2024, Kos, Greece

Improving Speech Recognition with Prompt-based Contextualized ASR and LLM-based Re-predictor

Nguyen Manh Tien Anh[1], *Thach Ho Sy*[1,2]

[1]VinBigdata, Vietnam
[2]Hanoi University of Science and Technology, Vietnam
v.anhnmt2@vinbigdata.org, thach.hs182769@sis.hust.edu.vn

Abstract

In recent years, advancements in automatic speech recognition (ASR) systems have led to their widespread use in applications such as call center bots and virtual assistants. However, these systems encounter challenges in adverse speech conditions, lack of contextual information, and recognizing rare words. In this paper, we propose a novel architecture to tackle these limitations by integrating Large Language Models (LLMs) and prompt mechanisms, aiming to enhance ASR accuracy. By using a pre-trained text encoder with a text adapter for task-specific adaptation and an efficient LLM-based re-prediction mechanism, our method has shown remarkable results in various real-world scenarios. Our proposed system achieves an average relative word error rate improvement of 27% for conventional tasks, 30% for utterance-level contextual tasks, and 33% for word-level biasing tasks compared to a baseline ASR system on multiple public datasets.

Index Terms: speech recognition, large language model, contextual biasing

1. Introduction

In recent years, significant advancements have been witnessed in end-to-end automatic speech recognition (ASR) systems, primarily propelled by the rapid evolution of deep neural networks. The extensive utilization of these systems in various production applications, including call center bots and virtual assistants, is due to their superior capability of transcribing spoken languages in terms of efficiency and accuracy, surpassing that of conventional ASR systems [1]. However, despite the advancements, even state-of-the-art ASR systems encounter performance challenges when exposed to adverse conditions in the speech domain, such as speaker accent [2] and background noise [3]. Moreover, the lack of labeled speech training data in specific domains poses a significant obstacle, particularly since end-to-end ASR systems heavily rely on data. Consequently, these systems struggle to recognize rare words or phrases within such domains. Another critical issue is the lack of contextual information, which impedes even well-trained acoustic models from accurately transcribing speech. Thus, relying solely on acoustic-driven approaches may be restrictive in certain scenarios. In contrast, human speech recognition demonstrates remarkable robustness, as it additionally incorporates contextual information and linguistic knowledge to interpret confusing or distorted phrases. Consequently, there is a growing interest in research aimed at integrating contextual knowledge or text adaptation mechanisms, similar to human cognitive processes, into end-to-end ASR models to enhance transcription accuracy.

Common methods for incorporating textual information into ASR systems typically involve utilizing external language models (LMs) trained on text corpora. These text-based models are utilized for shallow fusion [4, 5] or re-scoring the n-best hypotheses [6, 7] generated by ASR systems, aiming to modify the posterior probabilities predicted by the ASR system. Another strategy, known as contextual biasing [8, 9], guides the recognition process toward contextual words or phrases through various mechanisms. Furthermore, the advent of Large Language Models (LLMs) [10] and their key component, prompts [11], has recently revolutionized the field of Natural Language Processing (NLP). These models have demonstrated exceptional performance across a range of downstream text-processing tasks [12, 13], often requiring minimal fine-tuning [14, 15]. Consequently, there is a growing interest in harnessing these recent advancements in LLMs and prompt mechanisms to enhance ASR model performance. Several studies explore the use of LLMs for ASR error correction [16, 17, 18], where n-best hypotheses generated by the ASR system are utilized to predict true transcriptions. This deviates from conventional language model re-scoring methods, wherein the output transcription is consistently determined by selecting the hypothesis with the highest posterior probability. However, this approach solely relies on the n-best hypotheses from ASR, lacking attention to acoustic and contextual information. Consequently, LLMs utilized for ASR correction may yield over-corrected results, as they may infer outcomes that diverge significantly from the n-best hypotheses, leading to hallucination issues. Other studies utilize prompt mechanisms from LLMs to inject contextual information into the ASR system [19, 20]. Although the integration of prompts for contextualized ASR shows promise in achieving improved predictions, the stability of the results remains a concern. Additionally, the inclusion of contextual information during the inference phase is essential to enhance ASR accuracy.

To solve the limitation above, we propose a novel architecture that leverages both LLMs and prompt mechanisms to efficiently adapt contextual information for enhancing ASR accuracy. Concretely, a text encoder is added to the ASR system to improve hypothesis prediction by utilizing the contextual prompt capabilities. Moreover, to leverage the pre-trained text encoder's high generalization capability, its parameters are frozen and integrated with a trainable text adapter to adapt to the ASR task. The encoded prompts are then injected into the ASR encoder using a cross-attention mechanism. By utilizing additional contextual information, this prompt-based strategy significantly improves final and n-best hypotheses accuracy predicted by the ASR system. Subsequently, n-best hypotheses with corresponding acoustic information and contextualized prompts, are fed into LLMs to re-predict the ASR transcription, instead of relying solely on text-based n-best hypotheses. This integration of efficient prompts enables LLMs to overcome over-corrected

10.21437/Interspeech.2024-1762

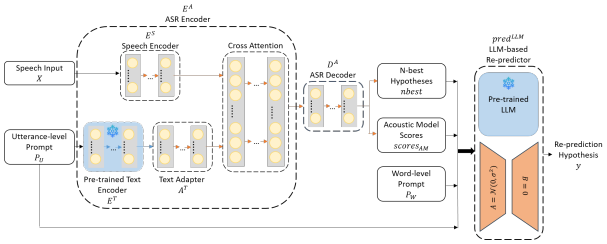

Figure 1: *The architecture of the proposed system.*

problems while preserving fidelity to spoken language. Furthermore, we present the definition of contextual data with its augmentation strategies employed for the prompt-based ASR and LLMs, thereby enhancing the overall performance of the system. Our proposed architecture shows significant improvements on LibriHeavy and other public datasets by including additional contextual information attributes and re-prediction mechanisms. Concretely, our proposed system achieves an average relative word error rate (WER) improvement of 27% for conventional tasks, 30% for utterance-level contextual tasks, and 33% for word-level biasing tasks compared to a baseline ASR system on the above public datasets.

The remainder of this paper is structured as follows: Section 2 explains our methodology and contextual prompt strategy. Section 3 outlines the experimental setup, followed by Section 4, which presents the analysis of results. Section 5 finally presents the conclusions of this paper.

2. Methodology

2.1. System Architecture

The architecture of our proposed system is illustrated in Fig. 1. Our system consists of three main components: ASR encoder, ASR decoder, and LLM-based re-predictor. Specifically, the ASR encoder and ASR decoder collaborate to establish prompt-based contextualized ASR, a novel architecture that incorporates additional contextual information during the prediction of transcription. Subsequently, the LLM-based re-predictor retrieves all informative elements from ASR output including n-best hypotheses, acoustic information, and context prompts to re-predict the true transcription.

2.1.1. Prompt-based Contextualized ASR

The prompt-based contextualized ASR consists of an ASR encoder E^A and an ASR decoder D^A. In contrast to traditional ASR, in which E^A solely processes the speech input X through a speech encoder E^S, a pre-trained text encoder E^T is added to represent contextual information at the utterance-level P_U. Specifically, to preserve pre-trained knowledge and prevent over-fitting issues, we freeze E^T and integrate it with a trainable text adapter A^T during the training phase. In A^T, multiple fully connected layers with biases are interspersed with a non-linear layer. To capture intricate relationships between acoustic features f_S derived from E^S and textual features f_T extracted by E^T, a multi-head cross-attention mechanism was implemented. Concretely, in this mechanism, f_T functions as key/value pairs, while f_S functions as queries, producing the cross-attention feature f_{attn}. Consequently, D^A receives not only acoustic embeddings f_S for predicting the output transcription, as in conventional ASR, but also incorporates contextual text embeddings f_T. It's essential to highlight that the entire system is adaptable and can be trained with various ASR encoder architectures, ASR decoder architectures, and ASR objective functions. Equation 1 expresses the forward process of prompt-based contextualized ASR:

$$\begin{aligned} f_S &= E^S(X) \\ f_T &= A^T(E^T(P_U)) \\ f_{attn} &= E^A(f_S, f_T) \\ nbest, scores_{AM} &= D^A(f_{attn}) \end{aligned} \tag{1}$$

where $nbest$ represents the n-best hypotheses and $scores_{AM}$ denotes the list of logit scores corresponding to the generated $nbest$ from the prompt-based contextualized ASR.

2.1.2. LLM-based Re-predictor

To enhance the extraction of information from $nbest$ list in the prompt-based contextualized ASR, we propose a technique that utilizes LLMs to directly re-predict a true transcription, rather than selecting the best candidate from $nbest$. However, relying solely on $nbest$ for LLMs may lead to "over-correction" cases, lacking fidelity to spoken language due to the lack of both acoustic and contextual information. Consequently, we introduce an LLM-based re-predictor $pred^{LLM}$ that incorporates $nbest$, acoustic information $scores_{AM}$, and contextual information for improved ASR re-prediction. Concretely, $nbest$ list, along with their respective $scores_{AM}$ from the ASR output, are utilized to create a final prompt, combining contextual information at both the utterance-level P_U and word-level P_W. Moreover, given the inefficiency of fully fine-tuning large pre-trained models, we employ LoRA [14] techniques to adapt the LLM to our ASR re-prediction task. LoRA avoids tuning the entire set of parameters by introducing a neural module with a small number of extra trainable parameters, approximating full parameter updates. This allows for efficient learning of the proposed model mapping without affecting the pre-trained parameters of the LLM. Specifically, LoRA reparameterizes each model layer through matrix multiplication by injecting low-rank decomposition matrices, as depicted in Fig. 1. Therefore, the LLM-produced representations remain undistorted during task-specific tuning. Simultaneously, the adapter module acquires the ability to re-predict the final transcription based on the prompt encompassing $nbest$, $scores_{AM}$, P_U, and P_W. With effective training, we can incorporate the LLM into the ASR re-prediction approach, relying on its understanding of task specifics and capability to recognize relationships within the comprehensive prompt information. Equation 2 represents the forward process to generate re-prediction hypothesis y of the LLM-based Re-predictor:

$$y = pred^{LLM}(nbest, scores_{AM}, P_U, P_W) \tag{2}$$

2.2. Prompts for Contextualized ASR and LLMs

In this paper, we define three types of prompts: word-level prompt, utterance-level prompt, and re-prediction prompt. The word-level prompt consists of a list of words or phrases intended for enhancement. It aims to improve the system's recognition accuracy for specific target words and phrases, such as contact names, application names, locations, or music playlists. Consequently, the word-level prompt serves as an ASR personalization strategy, leveraging the context of a specific user to significantly boost ASR accuracy for that user.

In contrast, the utterance-level prompt contains more semantic and context-related information expressed in the form of sentences rather than a list. It can be further categorized into prior prompt and instruct prompt. The prior prompt maintains a

Utterance-level prompt	Prior prompt	close the window and open the asphalt game
	Instruct prompt	this is a vinfast voice assistant system
Word-level prompt		vinfast, asphalt
Target transcription		hey vinfast which company is currently developing the asphalt game

Table 1: *Example of prompts with their target transcription.*

logical and close relationship with the target transcription, often comprising historical texts from the conversation or preceding sentences in an audiobook transcription. On the other hand, the instruct context represents the topic of the target transcription, providing contextual information for the transcription. Consequently, the utterance-level prompt contributes to an overall enhancement in ASR accuracy across a broader context and domain, focusing not only on specific words or phrases, as with the word-level prompt. Prompt samples for both utterance-level and word-level, along with their corresponding target transcription, are presented in Table 1.

Finally, we design an efficient re-prediction prompt that combines all informative attributes, encompassing n-best hypotheses, acoustic information, utterance-level context, and word-level context. This comprehensive prompt enables the LLM to re-predict the final transcription more effectively. With the inclusion of full-information prompts, LLM can achieve better results without being affected by "over-correction" problems and maintain fidelity to spoken language by utilizing additional acoustic and contextual information. The design of the re-prediction prompt for LLM is expressed as follows:

"Re-predict the final transcription from the automatic speech recognition (ASR) results using this information with the corresponding description below. Note that any missing information is denoted by "None". Please re-predict the final transcription only, do not add any explanations or additional content: ### N-best hypotheses is a list of hypotheses predicted by the ASR system: $\{1 \sim N$ *utterances*$\}$ *### Acoustic scores is a list of logit scores from the ASR system corresponding to the above hypotheses:* $\{1 \sim N$ *acoustic scores*$\}$ *### Utterance-level text describes the topic or logical relationships with the true transcription:* $\{$*texts*$\}$ *### Word-level text provides a list of words or phrases requiring improved recognition by the ASR if possible:* $\{$*[words/phrases]*$\}$*"*

3. Experimental Setup

3.1. Training Datasets

To train the prompt-based contextualized ASR, we utilized the publicly available Libriheavy dataset [21], encompassing 50,000 hours of audiobook recordings. Distinguishing itself from conventional datasets, Libriheavy not only provides audio samples and corresponding ground truth transcriptions but also includes contextual information at the utterance level. Specifically, the textual context comprises transcriptions from preceding utterances, with a default length of 1000 bytes. Within Libriheavy, three training subsets—small (500 hours), medium (5000 hours), and large (50000 hours) are available. In this paper, we employed the medium subset containing 5000 hours for training the prompt-based contextualized ASR model. Furthermore, in pursuit of encompassing diverse scenarios encountered in ASR tasks, such as context availability, non-contextual information, domain-specificity, background noise, and multi-accent instances, we augmented our training corpus. This augmentation involved combining approximately 3000 hours of audio-

Test set		Baseline	**Contextualized ASR + LLM re-prediction**
WSJ		5.82	**2.93**
CommonVoice-*en*		11.86	**8.8**
CHiME4		13.21	**8.09**
Librispeech	test-clean	3.64	**3.16**
	test-other	7.68	**6.82**

Table 2: *WER (%) results of the baseline model and the proposed model on multiple test sets in conventional tasks.*

transcription data from distinct datasets with varied characteristics. Specifically, we combined Librispeech [22], a widely used dataset for standard ASR evaluation, Wall Street Journal (WSJ) [23], representing a domain-specific dataset, CHiME4 [24], designed for scenarios involving noise, and CommonVoice-*en* [25], which offers a multi-accent dataset. By employing this comprehensive training data preparation strategy, our evaluation results thoroughly describe the model's performance across different types of input data and assess its robustness under varied scenarios.

To create the training dataset for the LLM-based re-predictor, we employed the previously trained prompt-based contextualized ASR system to generate n-best hypotheses for each utterance within the aforementioned ASR dataset. Specifically, we utilized the beam search decoding algorithm with a beam size of 6 to generate n-best lists. Additionally, to ensure that the LLM considers not only the n-best hypotheses from the ASR system but also incorporates acoustic and contextual information, we included acoustic scores (e.g., logit scores) along with the n-best lists. Furthermore, both utterance-level and word-level contextual information were incorporated into the training dataset, if available. Concretely, to create a word-level biasing list, we perform a word frequency analysis on the training corpus, identifying the 5,000 words with the lowest frequency as rare words. Subsequently, a word-based contextual list is generated for each training sample by selecting the rare words present in the respective sample and expanding up to 100 random distractors. The final dataset designated for LLM fine-tuning encompasses more than 1,500,000 samples. Each sample consists of n-best lists, corresponding acoustic scores, additional contextual information, and the corresponding ground truth.

3.2. Model Selection

In the prompt-based contextualized ASR, we utilize a pre-trained BERT model [26] as the text encoder to capture contextual prompts at the utterance-level. Specifically, we employ the pre-trained bert-based-uncased model with 110M parameters. The ASR system adopts a neural transducer with a Zipformer [27] speech encoder, a stateless decoder, and a jointly trained network using the prunedRNNT [28] objective loss.

For the LLM re-prediction model, we select the pre-trained LLaMA [12] foundation model. This model has demonstrated notable efficiency in fine-tuning public NLP benchmarks. Concretely, we adopt the LLaMA model with 13B parameters for LoRA adaptation, considering it as the most suitable setup for our approaches.

3.3. Training Details

We conducted separate training for the prompt-based contextualized ASR and the LLM-based re-predictor. The contextualized ASR utilizes 80-dimensional mel filter bank features, and the vocabulary size, generated through Byte-pair encoding [29] subword algorithm, is set to 1024. The model is trained for 60

Test set		Baseline	Contextualized ASR	**Contextualized ASR + LLM re-prediction**
Libriheavy	test-clean	3.31	2.59	**2.32**
	test-other	6.98	5.43	**4.84**

Table 3: *WER (%) results of the baseline model, the prompt-based contextualized ASR, and the final proposed model in utterance-level contextual tasks.*

epochs, and the final inference checkpoints are obtained by averaging the results of the last ten epochs. During the inference phase, we employ the beam search decoding algorithm with a beam size of 6.

In the fine-tuning process of LLM-based re-predictor, a learning rate of $1e^{-4}$ and a batch size of 128 are employed. Regarding the LoRA configuration, the hyperparameter for rank r is set to 8. The LM-based re-predictor is trained for 10 epochs using the AdamW optimizer. In our experiments, both models are trained on 8 NVIDIA A100 GPUs.

3.4. Evaluation

To test the robustness of our proposed system in multiple scenarios, we evaluate our proposed system by word error rate (WER) metric within three tasks: conventional task, utterance-level contextual task, and word-level contextual task. For conventional task, the official test set of Librispeech, WSJ, CommonVoice, and CHiME4 was used. The default Libriheavy test-clean and test-other sets which include supplementary contextual information are adopted for evaluation in the utterance-level contextual task. Finally, we utilized the biasing list for the Librispeech corpus provided in [30] to evaluate our system in word-level contextual tasks. As outlined in [30], the biasing list for each utterance is created by identifying words belonging to rare-word list from the Librispeech corpus, and contains a certain number of distractor words.

The baseline ASR system for comparison is a conventional neural transducer without a text encoder and re-predictor component. It is trained with the dataset previously outlined.

4. Experimental Results

4.1. Conventional Task

We initially conducted experiments to assess the effectiveness of traditional settings where contextual information is unavailable during the inference phase. As depicted in Table 2, our proposed system achieves significant performance improvements in specific scenarios when compared to the baseline ASR. In particular, the contextualized ASR with LLM-based re-predictor shows a relative WER reduction of 49.65%, 25.80%, and 38.75% on the WSJ, a domain-specific corpus, CommonVoice, a corpus with diverse speaker accents, and CHiME, a dataset designed for scenarios involving noise, respectively. Additionally, for Librispeech, improvements of 13.18% and 11.19% in relative WER are observed on the standard test-clean and test-other sets. This suggests that even without access to contextual information in this setting, our proposed system still achieves superior results, benefiting from the LLM-based re-predictor. By incorporating both acoustic information and n-best hypotheses from the ASR, the LLM can refine the final result with increased accuracy, mitigating potential "over-correction" issues highlighted in [18, 16], especially evident in the Librispeech test sets. Consequently, these results demonstrate the robustness and generalization capabilities of our proposed system across various scenarios in a conventional task.

Test set		Baseline	**Contextualized ASR + LLM re-prediction**			
			N=10	N=100	N=500	N=1000
Librispeech	test-clean	3.64	2.36	**2.33**	2.83	3.19
	test-other	7.68	5.41	**5.36**	6.12	6.91

Table 4: *WER (%) results of the baseline model and the proposed model with different sizes of biasing list N in word-level contextual tasks.*

4.2. Utterance-level Contextual Task

Table 3 shows the WER results for our proposed system compared to the baseline ASR in the task focusing on utterance-level context. Specifically, the contextualized ASR with LLM-based re-predictor shows relative WER reductions of 29.90% and 30.65% on the Libriheavy test-clean and test-other sets when compared to the baseline ASR. Additionally, we report WER results for the prompt-based contextualized ASR without the re-predictor, using only the 1-best result from the beam search decoding algorithm to assess the effectiveness of integrating contextual information into ASR during training. We observe similar relative WER improvements of 21.75% and 22.20% on these two Libriheavy test sets for the prompt-based contextualized ASR without LLM re-prediction. This highlights the efficiency and robustness of utilizing the pre-trained text encoder to incorporate contextual information into the ASR system during training. The overall result in the second task demonstrates that integrating utterance-level contextual information into both ASR and LLM prompts significantly enhances overall ASR recognition accuracy compared to baseline ASR, which relies solely on acoustic features for transcription.

4.3. Word-level Contextual Task

We also evaluated the performance of our proposed system in a word-level biasing task. Table 4 shows the WER for our system with different sizes N of biasing lists in the Librispeech contextual task. In our experiments, the best result was achieved when employing a biasing list of 100 components, leading to a relative WER improvement of 35.98% and 30.20% compared to the baseline ASR. When the size of the biasing list increased, the observed WER improvements showed a noticeable decrease. Concretely, with $N = 1000$, the improvement in WER is 12.36% and 10.02% on the Librispeech test-clean and test-other. These results show that word-level biasing recognition via the LLM-based re-predictor is affected by the number of distractors N, even though there is still a noticeable improvement compared to the traditional method. The potential reason is that the LLM is fine-tuned on a relatively small word-biased list with fewer than 100 words, making the model sensitive when faced with a large number of distractors. Nevertheless, the overall result that incorporating word-level biasing prompts into the LLM-based re-predictor significantly aids the system in recognizing rare words during the inference phase, offering high flexibility through the prompt mechanism.

5. Conclusion

We propose a novel architecture that integrates LLMs and prompt mechanisms to enhance ASR accuracy in multiple scenarios. By utilizing a pre-trained text encoder with a trainable text adapter and an efficient LLM-based re-predictor, our approach outperforms the baseline method on different public datasets. Our system achieves an average relative WER reduction of 27%, 30%, and 33% for the traditional task, utterance-level contextual task, and word-level biasing task respectively.

6. References

[1] B. Li, S.-y. Chang, T. N. Sainath, R. Pang, Y. He, T. Strohman, and Y. Wu, "Towards fast and accurate streaming end-to-end asr," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6069–6073.

[2] Y. Qian, X. Gong, and H. Huang, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2842–2853, 2022.

[3] C. Chen, N. Hou, Y. Hu, S. Shirol, and E. S. Chng, "Noise-robust speech recognition with 10 minutes unparalleled in-domain data," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4298–4302.

[4] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5828.

[5] D. Le, G. Keren, J. Chan, J. Mahadeokar, C. Fuegen, and M. L. Seltzer, "Deep shallow fusion for rnn-t personalization," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 251–257.

[6] L. Xu, Y. Gu, J. Kolehmainen, H. Khan, A. Gandhe, A. Rastrow, A. Stolcke, and I. 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.

[7] Y. Yu, C.-H. H. Yang, J. Kolehmainen, P. G. Shivakumar, Y. Gu, S. R. R. Ren, Q. Luo, A. Gourav, I.-F. Chen, Y.-C. Liu *et al.*, "Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[8] K. Huang, A. Zhang, Z. Yang, P. Guo, B. Mu, T. Xu, and L. Xie, "Contextualized end-to-end speech recognition with contextual phrase prediction network," *arXiv preprint arXiv:2305.12493*, 2023.

[9] K. Huang, A. Zhang, B. Zhang, T. Xu, X. Song, and L. Xie, "Spike-triggered contextual biasing for end-to-end mandarin speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[10] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023.

[11] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.

[12] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[13] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[15] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. A. Raffel, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1950–1965, 2022.

[16] Z. Min and J. Wang, "Exploring the integration of large language models into automatic speech recognition systems: An empirical study," in *International Conference on Neural Information Processing*. Springer, 2023, pp. 69–84.

[17] P. Dighe, Y. Su, S. Zheng, Y. Liu, V. Garg, X. Niu, and A. Tewfik, "Leveraging large language models for exploiting asr uncertainty," *arXiv preprint arXiv:2309.04842*, 2023.

[18] C. CHEN, Y. Hu, C.-H. H. Yang, S. M. Siniscalchi, P.-Y. Chen, and E. Chng, "Hyporadise: An open baseline for generative speech recognition with large language models," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: https://openreview.net/forum?id=cAjZ3tMye6

[19] S. Dingliwal, A. Shenoy, S. Bodapati, A. Gandhe, R. T. Gadde, and K. Kirchhoff, "Prompt-tuning in asr systems for efficient domain-adaptation," *arXiv preprint arXiv:2110.06502*, 2021.

[20] X. Yang, W. Kang, Z. Yao, Y. Yang, L. Guo, F. Kuang, L. Lin, and D. Povey, "Promptasr for contextualized asr with controllable style," *arXiv preprint arXiv:2309.07414*, 2023.

[21] W. Kang, X. Yang, Z. Yao, F. Kuang, Y. Yang, L. Guo, L. Lin, and D. Povey, "Libriheavy: a 50,000 hours asr corpus with punctuation casing and context," 2023.

[22] 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.

[23] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[24] E. Vincent, S. Watanabe, A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, vol. 46, pp. 535–557, 2017.

[25] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[27] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang, Y. Yang, Z. Jin, L. Lin, and D. Povey, "Zipformer: A faster and better encoder for automatic speech recognition," *arXiv preprint arXiv:2310.11230*, 2023.

[28] F. Kuang, L. Guo, W. Kang, L. Lin, M. Luo, Z. Yao, and D. Povey, "Pruned RNN-T for fast, memory-efficient ASR training," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 2068–2072. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-10340

[29] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, K. Erk and N. A. Smith, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[30] D. Le, M. Jain, G. Keren, S. Kim, Y. Shi, J. Mahadeokar, J. Chan, Y. Shangguan, C. Fuegen, O. Kalinli, Y. Saraf, and M. L. Seltzer, "Contextualized Streaming End-to-End Speech Recognition with Trie-Based Deep Biasing and Shallow Fusion," in *Proc. Interspeech 2021*, 2021, pp. 1772–1776.