Interspeech 2024
1-5 September 2024, Kos, Greece

Pinyin Regularization in Error Correction for Chinese Speech Recognition with Large Language Models

Zhiyuan Tang[1], *Dong Wang*[2], *Shen Huang*[1], *Shidong Shang*[1]

[1]Tencent Ethereal Audio Lab, Tencent, China
[2]Center for Speech and Language Technologies, BNRist, Tsinghua University, China
atomtang@tencent.com, wangdong99@mails.tsinghua.edu.cn

Abstract

Recent studies have demonstrated the efficacy of large language models (LLMs) in error correction for automatic speech recognition (ASR). However, much of the research focuses on the English language. This paper redirects the attention to Chinese. Firstly, we construct a specialized benchmark dataset aimed at error correction for Chinese ASR with 724K hypotheses-transcription pairs, named the Chinese Hypotheses Paradise dataset (ChineseHP), which contains a wide range of scenarios and presents significant challenges. Subsequently, we conduct a preliminary evaluation using the dataset for both direct-prompting and fine-tuning pre-trained LLMs. Furthermore, we propose a straightforward method of Pinyin regularization for prompts, which involves the transcription of Pinyin directly from text hypotheses. The experimental results reveal that Pinyin regularization consistently enhances the error-correcting ability of LLMs when compared with those without regularization. The dataset is available on the website[1].

Index Terms: speech recognition, error correction, large language model

1. Introduction

Automatic speech recognition systems (ASR) are extensively employed in a multitude of applications, including voice search, voice command, and transcription services. Nonetheless, the efficacy of ASR can be significantly influenced by a range of elements, such as background noise, speaker accents, and the fidelity of the audio signal. Errors in the ASR output, particularly in challenging environments, can be adverse to the functionality of downstream applications. Therefore, implementing subsequent error correction processes plays a vital role in enhancing the precision of ASR outputs.

The practice of employing a language model (LM) to rescore the N-best hypotheses from ASR beam search decoding is a common technique to identify the candidate with the lowest perplexity, according to various studies [1, 2, 3, 4, 5]. However, such LM rescoring merely chooses the optimal candidate, thereby neglecting the valuable information contained in the remaining hypotheses. A potentially more advantageous strategy involves merging the N-best hypotheses to generate a new prediction, which is anticipated to be more accurate than the initial candidates [6, 7, 8, 9, 10, 11].

Recently, large language models (LLMs) have begun to leverage their capacity for understanding language to assist in error correction in a generative style [12, 13]. In particular, a benchmark for generative error correction (GER) tailored to the English language has been introduced. This benchmark is specifically designed to generate correct transcription directly from N-best hypotheses produced by ASR systems. Additionally, a dataset named HyParadise has been created, containing 316K pairs of hypotheses and transcription, to facilitate the evaluation of GER performance. Building on this, subsequent research has expanded the GER benchmark to encompass a broader range of common noisy conditions encountered in ASR. This expansion has led to the development of a new dataset, Robust HyParadise (RobustHP), comprising 113K pairs for a more comprehensive evaluation of hypotheses and transcription.

While most of existing research primarily focuses on the English language, this study redirects attention to the Chinese language. The objective of this paper is to create a benchmark dataset tailor-made for error correction in Chinese, dubbed the Chinese Hypotheses Paradise dataset (ChineseHP), which comprises 724K hypotheses-transcription pairs. The ChineseHP dataset spans a wide array of contexts and presents a significant challenge for error correction.

Given that Chinese is a logographic language, the pronunciation of its characters does not inherently correspond to their written form. Pinyin, the romanization system for standard Chinese, is extensively utilized throughout China for teaching the language. Additionally, Pinyin is a common method for entering Chinese characters on computers and smartphones. The prevalent use of Pinyin renders it an accessible resource for LLMs to comprehend Chinese. Moreover, Chinese is rich in homophones, and numerous characters share similar initial or final phonetic elements. Consequently, a Pinyin transcription derived directly from the text hypothesis tends to exhibit a lower error rate than the text hypothesis itself, which is advantageous for the task of error correction. Therefore, we propose a Pinyin regularization method to be applied both in prompting pre-trained LLMs and during the fine-tuning phase to bolster the stability and efficacy of LLMs in error correction for Chinese ASR. Experimental outcomes consistently demonstrate that Pinyin regularization significantly improves the language model's ability to correct errors in Chinese language materials.

The remainder of the article is structured in the following manner. Section 2 introduces the Chinese Hypotheses Paradise dataset. Section 3 details the Pinyin regularization method. Section 4 outlines the experimental framework and findings. The paper is concluded in Section 5.

2. Chinese Hypotheses Paradise dataset

The ChineseHP dataset is collected from the ASR outputs of a Chinese-adapted and distilled version of the well-known Whisper Large V2 [14], named Belle-distilwhisper-large-v2-zh[2].

[1]https://github.com/tzyll/ChineseHP

[2]https://huggingface.co/BELLE-2/Belle-distilwhisper-large-v2-zh

10.21437/Interspeech.2024-987

The dataset encompasses four representative Chinese corpora: Aishell-1 [15], Wenetspeech [16], Aishell-4 [17], and Kespeech [18]. The dataset's statistical details are presented in Table 1. It is evident that the dataset contains a diverse range of situations, encompassing regular read speech, broadcast news, meetings, telephone conversations, as well as various accents and dialects. Specifically, the Aishell-1 corpus comprises standard reading speech, while Wenetspeech contains multiple domains from the internet, and includes the sub-corpora test_net and test_meeting for test, which features broadcast news and meeting speech correspondingly. Aishell-4 serves as a telephone conversation corpus, and Kespeech focuses on dialects. Considering Wenetspeech and Kespeech have significantly more data than Aishell-1 and Aishell-4, we limited our sample to 200K utterances from each to maintain balance within the dataset.

For each audio sample, a beam size of 10 was utilized during ASR decoding to generate the top 10 hypotheses. These hypotheses were then converted to simplified Chinese, deduplicated, and paired with the correct transcriptions to create hypotheses-transcription pairs.

Table 1: *Statistics of the Chinese Hypotheses Paradise dataset (ChineseHP). Wenetspeech/test contains two sub-corpora, i.e., test_net and test_meeting.*

Dataset	Description	Subset		
		train	dev	test
Aishell-1	reading style, clean	120,098	14,326	7,176
Wenetspeech	multi-domain, noisy	200,000	13,825	33,143
Aishell-4	meeting, overlapped	97,317	3,959	10,423
KeSpeech	accent, dialect	200,000	4,407	19,723
Total		617,415	36,517	70,465

Table 2: *Character error rate (CER) and Pinyin error rate (PinyinER) of test sets in ChineseHP. Wenetspeech/test contains two sub-corpora, i.e., test_net and test_meeting.*

	Aishell-1	Wenetspeech	Aishell-4	KeSpeech
CER%	5.84	11.97/16.07	25.27	29.83
PinyinER%	1.46	6.37/11.40	20.37	11.03

3. Pinyin regularization

3.1. Pinyin system

Hanyu Pinyin, often referred to simply as Pinyin, is a romanization system for Mandarin Chinese, typically comprising 23 initials, 24 finals, and 5 tones, including the neutral tone. While various Pinyin systems may exhibit minor variations in the number of initials and finals, they all adhere to the same foundational rules. In this paper, we utilize the Pinyin system as presented in *pypinyin*[3], with the modification that we employ "ü" in place of "v" and "en" instead of "n" for the final, as these are more prevalently used in China. The list below displays the initials and finals in Pinyin where the finals include several blends of basic finals, for example, "ua" is a composite of "u" and "a":

- Initials: b, c, ch, d, f, g, h, j, k, l, m, n, p, q, r, s, sh, t, w, x, y, z, zh
- Finals: a, ai, an, ang, ao, e, ei, en, eng, er, i, ia, ian, iang, iao, ie, in, ing, iong, iu, o, ong, ou, u, ua, uai, uan, uang, ue, ui, un, uo, ü, üe

[3]https://pypi.org/project/pypinyin

The pronunciation of Chinese characters is represented through a combination of initials and finals. For instance, the Chinese character for "you", which is written as "你", is articulated as "ni3" in Pinyin. Here, "n" serves as the initial, "i" acts as the final, and "3" denotes the tone. It is worth noting that certain Chinese characters can consist solely of a final, without an accompanying initial. Additional scenarios that may create confusion include:

- Homophone: different characters have the same pronunciation, some even the same tone, e.g., "桌 (desk)" and "捉 (catch)" are both pronounced as "zhuo1".
- Heteronym: same character has different pronunciations, e.g., "都" is pronounced as "dou1 (all)" and "du1 (capital)".
- Alliteration: some initial but different finals, e.g., "桌 zhuo1(desk)" and "助 zhu4 (help)" have the same initial "zh".
- Rhyme: different initials but same final, e.g., "都 dou1 (all)" and "狗 gou3 (dog)" have the same final "ou".

These various factors can easily perplex ASR systems, leading to mistakes when confronted with any potential disruptions, including noise, accents, or dialects. Regarding errors in ASR output, while the character may be incorrect, the corresponding Pinyin—transcribed directly from the text hypothesis—is frequently accurate, either partially or wholly, often yielding a reduced error rate compared to the text hypothesis alone, as illustrated in Table 2. This is clearly advantageous for the task of error correction.

3.2. Pinyin-regularized prompts

We have devised a pair of prompt styles: the first is employed for immediate engagement with pre-trained LLMs, such as ChatGPT[4] as referenced in this study, while the second style is tailored for the fine-tuning process of a pre-trained LLM.

3.2.1. Direct prompt

The first prompt is specifically crafted to engage pre-trained LLMs in a direct manner. In this paper we employ ChatGPT, specifically GPT-3.5. Given ChatGPT's proficiency in English, we have utilized an English variant of the prompt as depicted in Figure 1 (upper). The prompt is structured to incorporate text hypotheses and the Pinyin transcribed directly from the text hypotheses can be optionally included. To decrease the hallucination of the language model output, we instruct the model to reply using a JSON format, which gives more stability and controllability.

3.2.2. Prompt in fine-tuning

As for fine-tuning LLMs for Chinese error correction, this paper favors the chatGLM model [19], given its considerable focus on the Chinese language. The training data is formatted in a prompt-response manner, which involves incorporating the pairs of hypotheses and transcription from ChineseHP into the Chinese prompts, as illustrated in Figure 1 (lower).

4. Experiments

4.1. Direct-prompting ChatGPT

We perform experiments to assess how various prompts affect the performance of ChatGPT in error correction. Taking into account the number of optimal hypotheses for either text or Pinyin, along with the kinds of Pinyin, we have crafted 9 distinct prompts. These are detailed in Table 3.

[4]https://chat.openai.com

```
Given a list of the top N-best text outputs and N-best Pinyin transcriptions
from a Chinese Automatic Speech Recognition (ASR) system, your task is to use
your language understanding capabilities to correct potential errors. Analyze
the text and Pinyin, identify discrepancies, and provide the most accurate and
contextually appropriate correction. Consider factors like homophones, tonal
variations, and context to ensure the corrected output accurately represents
the intended message.

N-best text outputs:
1. 今天天气很好
2. 今天天气很早
3. 今天天其很好
4. 今天甜气很好
5. 景天天气很好

N-best Pinyin transcriptions:
1. j in t ian t ian q i h en h ao
2. j in t ian t ian q i h en z ao
3. j in t ian t ian q i h en h ao
4. j in t ian t ian q i h en h ao
5. j ing t ian t ian q i h en h ao

Please output the corrected text directly without any explanation. Remember not
to change the number of characters in the corrected text. If you are not
confident in the correction, please select the best option from the N-best list.
If you are confident in the correction, please provide the corrected text
directly.

The output should be in the following JSON format:
{{"text": "your corrected text"}}
```

```
基于提供的 N-best 文本候选和 N-best 拼音转录，纠正文本并直接输出结果。
(Based on the provided N-best text candidates and N-best Pinyin transcriptions,
correct the text and output the result directly.)

N-best 文本候选:
(N-best text candidates)
1. 今天天气很好
2. 今天天气很早
3. 今天天其很好
4. 今天甜气很好
5. 景天天气很好

N-best 拼音转录:
(N-best Pinyin transcriptions)
1. j in t ian t ian q i h en h ao
2. j in t ian t ian q i h en z ao
3. j in t ian t ian q i h en h ao
4. j in t ian t ian q i h en h ao
5. j ing t ian t ian q i h en h ao
```

Figure 1: *Examples of Pinyin-regularized prompts for direct-prompting ChatGPT (upper) and fine-tuning ChatGLM (lower). The red part is only used when Pinyin regularization is enabled. The gray part in parentheses is just for translation. The corrected text is supposed to be "今天天气很好 (Today's weather is very good)".*

Considering economy and efficiency, we selectively sampled approximately 2,000 hypotheses-transcription pairs from the test set of each ChineseHP dataset to assess the efficacy of various prompts. We use the character error rate reduction (CERR) to measure the improvements different methods offered over the baseline. The variations in CERR across different prompts for each dataset are depicted in Figure 2.

The figure illustrates that incorporating Pinyin within the prompt enhances the performance of ChatGPT, with accuracy directly correlating to the precision of the provided Pinyin, as indicated by $prompt3 > prompt2 > prompt1$ ($>$ means 'better than', same below). This effect is particularly notable when considering a singular best hypothesis, as demonstrated by $prompt7 > prompt6 > prompt5$.

In an attempt to discern whether duplicating the 1-best text hypothesis could serve as an effective alternative to Pinyin, we carry out an experiment, but the results indicated that this method was less effective when compared to the use of Pinyin, as shown by $prompt6 > prompt8$.

Moreover, we double the 1-best Pinyin hypothesis to determine the effect of additional Pinyin in the prompt. This leads to a modest enhancement in performance, with $prompt9 > prompt6$, although the gain was not as substantial as the initial inclusion of Pinyin.

Table 3: *Different prompts for ChatGPT. "$*$" means repeating the first candidate for N times. Ground-truth Pinyin is only used for analysis and unavailable in practice.*

Method	N for best hypotheses		
	Text	Pinyin	
		Transcribed	Ground-truth
Baseline	-	-	-
Prompt1	5	-	-
Prompt2	5	5	-
Prompt3	5	-	1
Prompt4	5	-	5*
Prompt5	1	-	-
Prompt6	1	1	-
Prompt7	1	-	1
Prompt8	2*	-	-
Prompt9	1	2*	-

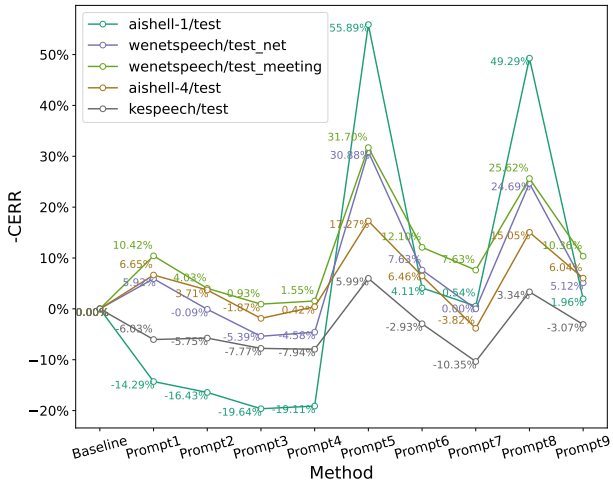

Figure 2: *Minus Character error rate reduction (-CERR) of different prompts for ChatGPT. Lower is better.*

Inspired by these findings, we experiment with replicating the ground-truth Pinyin hypothesis 5 times. However, this results in only a negligible improvement in performance: $prompt4 \approx prompt3$. This suggests that the 5-best text hypotheses, providing a variety of information, may offer a comparable regularization effect and that further repetition of the same Pinyin does not contribute significant additional insight.

4.2. Fine-tuning ChatGLM

We perform a series of experiments to assess the efficacy of various prompts for fine-tuning ChatGLM (Version 3[5] with 6 billion parameters). Taking into account the optimal number of hypotheses for text or Pinyin, we have formulated four distinct fine-tuning prompts, which are detailed in Table 4.

The training data is prepared by combining subsets of each dataset in ChineseHP, precisely 10K hypotheses-transcription pairs from Aishell-1/train, 20K from Wenetspeech/train, 20K from Aishell-4/train, and 20K from KeSpeech/train. The data list can be found on the ChineseHP website. We use 4 NVIDIA A100 GPUs to fine-tune the ChatGLM model with 2 epochs for

[5]https://github.com/THUDM/ChatGLM3

Table 4: *Results of different fine-tuning prompts for error correction with ChatGLM. o_{cp} and o_{nb} are compositional oracle and n-best oracle respectively, defined in Section 4.2. The base model is unable to perform the task precisely, so the CERs are not available.*

Method	**N for best hypotheses**		**CER%↓** $-CERR\%\downarrow$				
	Text	Transcribed Pinyin	Aishell-1/test	Wenetspeech/test_net	Wenetspeech/test_meeting	Aishell-4/test	KeSpeech/test
Baseline	-	-	5.84	11.62	16.03	25.19	29.81
o_{cp}	-	-	$2.93_{-49.83}$	$6.89_{-40.71}$	$8.67_{-45.91}$	$13.34_{-47.04}$	$22.9_{-23.18}$
o_{nb}	-	-	$3.49_{-40.24}$	$8.48_{-27.02}$	$12.76_{-20.40}$	$20.06_{-20.37}$	$25.49_{-14.49}$
Finetune1	5	-	$5.14_{-11.99}$	$13.27_{14.20}$	$17.3_{7.92}$	$27.4_{8.77}$	$31.12_{4.39}$
Finetune2	5	5	$4.74_{-18.84}$	$12.28_{5.68}$	$16.33_{1.87}$	$25.61_{1.67}$	$28.73_{-3.62}$
Finetune3	1	-	$6.88_{17.81}$	$14.36_{23.58}$	$18.25_{13.85}$	$28.09_{11.51}$	$33.25_{11.54}$
Finetune4	1	1	$6.26_{7.19}$	$14.06_{21.00}$	$17.94_{11.92}$	$27.61_{9.61}$	$31.2_{4.66}$

each fine-tuning prompt. For computational efficiency, the low-rank adaptation (LoRA) approach [20] is applied to optimize a small number of parameters. The fine-tuning pipeline and LoRA configuration follows an open-source repository[6].

The results are presented in Table 4. Due to the limited number of instances in training data with transcriptions exceeding 100 characters, we confine our report to test set results with lengths under 100 characters concerning references, hypotheses, and LLM outputs to prevent hallucination. As in [13], we also provide a comparison with two oracle CERs for reference, namely: 1) the n-best oracle o_{nb}, which represents the CER of the "best candidate" in the N-best hypotheses, and 2) the compositional oracle o_{cp}, indicating the minimum achievable CER by utilizing "all tokens" available in the N-best hypotheses. The o_{nb} is indicative of the potential peak performance of re-ranking-based approaches, whereas o_{cp} signals the ceiling for corrections that utilize elements present in the list.

The table reveals that fine-tuning ChatGLM with 5-best text hypotheses leads to a marked performance enhancement on the Aishell-1/test dataset. However, the improvement in performance on more intricate settings, such as Wenetspeech/test, Aishell-4/test, and KeSpeech/test, remains minimal. This aligns with the findings from ChatGPT, highlighting the significant challenges associated with error correction in these scenarios. The promising aspect is that the base model is empowered to perform the task precisely via fine-tuning.

The performance observed with the 1-best text or Pinyin hypothesis is not as good as that achieved with ChatGPT. We conjecture that this is due to the inadequate model size and the limited amount of training data, which fail to capture the complementary information, leading to a tendency for the model to overfit to the training dataset.

On the other hand, ChatGLM's performance is consistently enhanced when Pinyin is incorporated into the fine-tuning process, regardless of whether the number of top hypotheses is 5 or 1. This is evidenced by the fact that $finetune2 > finetune1$ and $finetune4 > finetune3$. Such results indicate the promising role of Pinyin regularization in reducing the errors in Chinese ASR systems.

4.3. Case Analysis

Two cases are shown in Table 5 to illustrate the performance of different fine-tuning prompts for error correction with ChatGLM. Case 1 is from Aishell-1/test, which is a standard reading sample with few errors in the N-best list, while Case 2 is from KeSpeech/test, which has more errors due to the accent.

In Case 1, the one or two errors can be effectively corrected with Pingyin regularization even with the 1-best hypothesis, while the result of the 5-best text hypotheses without Pinyin gets some hallucination.

In Case 2, the errors are of a more complicated nature, making it challenging to recover the original intent from the given hypotheses. This complexity increases the probability of hallucination, leading to a decline in performance across all fine-tuning models. Nonetheless, Pinyin regularization continues to mitigate the issue, preventing the partial text from being incorrectly re-generated under the constraints of the Pinyin.

Table 5: *Case analysis of different fine-tuning prompts for error correction with ChatGLM. The red Pinyin is for the errors compared to the ground truth.*

Method	**Utterance**	**CER%↓**
	Case 1	
Ground truth	一线楼市成交量激ji1增zeng1	-
N-best list	一线楼市成交量基ji1增	11.11
	一线楼市成交量机ji1增	11.11
	一线楼市成交量积ji2增	11.11
	一线楼市成交量基ji1僧seng1	22.22
	一线楼市成交量基ji1升sheng1	22.22
Finetune1	一线楼市成交量继续保持	44.44
Finetune2	一线楼市成交量激增	0.00
Finetune3	一线楼市成交量激增	0.00
Finetune4	一线楼市成交量即增	11.11
	Case 2	
Ground truth	当dang1你ni3面对马凝ning2视shi4时shi2马则不愿前qian2行	-
N-best list	但dan4念nian4面对马宁ning2事shi4实shi2马则不愿意yi4牵qian1行	50.00
	但dan4你面对马宁ning2事shi4实shi2马则不愿意yi4牵qian1行	42.86
	当年nian2面对马宁ning2事shi4实shi2马则不愿意yi4牵qian1行	42.86
	但dan4念nian4面对马宁ning2事shi4实shi2马则不愿意yi4谦qian1行	50.00
	但dan4你面对马宁ning2事shi4实shi2马则不愿意yi4谦qian1行	42.86
Finetune1	但面对马宁实事求是地回答了问题	92.86
Finetune2	但面对马宁试马则不愿意前进一步	64.29
Finetune3	但面对马赛时马则不愿意前进一步	57.14
Finetune4	但面对马宁失事马则不愿意签新	57.14

5. Conclusion

In this study, we introduce a specialized benchmark dataset designed for error correction in Chinese ASR named the Chinese Hypotheses Paradise dataset (ChineseHP), comprising 724K hypotheses-transcription pairs. The dataset covers a wide range of real-world scenarios, representing a considerable challenge. Additionally, we develop baseline methods for prompting and fine-tuning pre-trained LLMs and propose a simple yet effective Pinyin regularization technique to enhance their robustness and performance. Moving forward, we plan to investigate more sophisticated fine-tuning approaches, develop more potent prompts, and utilize additional training data to further enhance the error correction capabilities of LLMs for Chinese ASR.

[6]https://github.com/liucongg/ChatGLM-Finetuning

6. References

[1] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, no. 3. Makuhari, 2010, pp. 1045–1048.

[2] E. Arisoy, A. Sethy, B. Ramabhadran, and S. Chen, "Bidirectional recurrent neural network language models for automatic speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5421–5425.

[3] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using bert for speech recognition," in *Asian Conference on Machine Learning*. PMLR, 2019, pp. 1081–1093.

[4] C.-H. H. Yang, L. Liu, A. Gandhe, Y. Gu, A. Raju, D. Filimonov, and I. Bulyko, "Multi-task language modeling for improving speech recognition of rare words," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 1087–1093.

[5] 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.

[6] J. Guo, T. N. Sainath, and R. J. Weiss, "A spelling correction model for end-to-end speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5651–5655.

[7] K. Hu, T. N. Sainath, R. Pang, and R. Prabhavalkar, "Deliberation model based two-pass end-to-end speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7799–7803.

[8] Y. Leng, X. Tan, R. Wang, L. Zhu, J. Xu, W. Liu, L. Liu, T. Qin, X.-Y. Li, E. Lin *et al.*, "Fastcorrect 2: Fast error correction on multiple candidates for automatic speech recognition," *arXiv preprint arXiv:2109.14420*, 2021.

[9] K. Hu, T. N. Sainath, Y. He, R. Prabhavalkar, T. Strohman, S. Mavandadi, and W. Wang, "Improving deliberation by text-only and semi-supervised training," *arXiv preprint arXiv:2206.14716*, 2022.

[10] R. Ma, M. J. Gales, K. M. Knill, and M. Qian, "N-best t5: Robust asr error correction using multiple input hypotheses and constrained decoding space," *arXiv preprint arXiv:2303.00456*, 2023.

[11] K. Hu, B. Li, and T. N. Sainath, "Scaling up deliberation for multilingual asr," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 771–776.

[12] C. Chen, Y. Hu, C.-H. H. Yang, H. Liu, S. M. Siniscalchi, and E. S. Chng, "Generative error correction for code-switching speech recognition using large language models," *arXiv preprint arXiv:2310.13013*, 2023.

[13] C. Chen, Y. Hu, C.-H. H. Yang, S. M. Siniscalchi, P.-Y. Chen, and E.-S. Chng, "Hyporadise: An open baseline for generative speech recognition with large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] 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.

[15] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[16] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng *et al.*, "Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6182–6186.

[17] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu *et al.*, "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," *arXiv preprint arXiv:2104.03603*, 2021.

[18] Z. Tang, D. Wang, Y. Xu, J. Sun, X. Lei, S. Zhao, C. Wen, X. Tan, C. Xie, S. Zhou *et al.*, "Kespeech: An open source speech dataset of mandarin and its eight subdialects," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[19] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "Glm: General language model pretraining with autoregressive blank infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 320–335.

[20] 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.