LLM-Driven Hypothesis Set Refinement for Enhanced ASR Post-Processing

Chen-Han Wu and Kuan-Yu Chen
National Taiwan University of Science and Technology, Taiwan
E-mail: {m11115119, kychen}@mail.ntust.edu.tw

***Abstract*—In high-performing automatic speech recognition (ASR) systems, post-processing plays a critical role alongside the ASR model itself. Recent studies increasingly utilize n-best hypothesis lists as inputs to post-processing models. While much of the existing research focuses on improving model architectures, we observe that the composition of the n-best list itself can significantly impact post-processing outcomes. Large language models (LLMs) have demonstrated strong performance across various speech-related tasks, often outperforming traditional methods. Although some studies have explored the use of LLMs for ASR post-processing, these efforts have primarily focused on selecting the hypothesis with the lowest error rate or directly correcting recognition errors. In this paper, we focus on the task of hypothesis set refinement (HSR)—selecting a high-quality subset of hypotheses from the original candidate list to better support downstream processing. We propose a novel LLM-based HSR method that leverages the semantic reasoning capabilities of LLMs to select a more informative, diverse, and error-tolerant hypothesis set. This refined set facilitates subsequent post-processing methods such as reranking and error correction. To validate our approach, we conduct experiments on two widely used ASR benchmarks: AISHELL-1 and LibriSpeech. Furthermore, we investigate the potential of LLMs to directly perform reranking and correction, demonstrating how LLMs can be effectively employed throughout the ASR post-processing pipeline to enhance overall system performance.**

I. Introduction

In automatic speech recognition (ASR), models are typically trained using paired speech and textual transcriptions [1]–[4]. While this training paradigm has proven effective, it often limits the model's ability to capture semantic relationships, resulting in relatively high recognition error rates. To address this limitation without modifying the ASR model itself, various post-processing techniques have been proposed to enhance the final output [5]–[10]. Among these, two primary approaches, *reranking* and *correction*, have gained significant attention.

Reranking and correction tackle the problem from different perspectives. Reranking involves re-scoring each hypothesis generated through beam search decoding and selecting the most promising one from the n-best hypothesis list [11]–[13]. In contrast, correction focuses on directly identifying and modifying semantically or syntactically incorrect tokens, words, or phrases in the top-1 hypothesis, offering greater flexibility [14]–[16]. Earlier correction methods often employed a two-step process: error detection followed by correction, designed to prevent unnecessary changes to already accurate parts of the sequence. More recently, advances in deep learning have enabled the development of end-to-end correction models, some of which take the entire candidate list as input rather than relying solely on the top-1 output [17], [18].

The rapid advancement of large language models (LLMs) has further reshaped the landscape of ASR post-processing [8], [16]. In reranking, LLM-based models offer rich semantic representations, enabling more accurate scoring at both the token and sentence levels. In correction, their powerful contextual understanding improves the accuracy of both detection and modification, often allowing a single model to handle both subtasks seamlessly.

Despite these innovations, most existing research—whether traditional or LLM-based—has primarily focused on improving reranking or correction mechanisms. However, an often overlooked but crucial factor is the quality of the input n-best hypothesis list itself. Prior work has proposed diversity-based selection algorithms that aim to increase the informativeness of the list by maximizing the diversity among hypotheses [19]. While this approach is intuitively appealing, results have shown that excessive diversity can inadvertently exclude correct or near-correct hypotheses, ultimately degrading performance. To mitigate this, strategies for generating pseudo ground truths—such as selecting the top-1 hypothesis, the most representative hypothesis, or a corrected version of the top-1—have been explored [19]. Nevertheless, achieving a balance between diversity and pseudo-average error rate remains challenging.

Inspired by these insights, we focus on the task of *hypothesis set refinement* (HSR) and present a novel LLM-based mechanism to improve the composition of the n-best hypothesis list for downstream ASR post-processing. HSR is defined as the task of selecting a high-quality subset of n hypotheses from an original m-best list, where $n \ll m$, in a manner that maximizes the effectiveness of subsequent post-processing steps such as reranking or error correction. Unlike traditional list construction approaches that rely solely on ASR decoding scores, HSR emphasizes three key criteria: semantic completeness, hypothesis diversity, and potential correctness. To address this task, we propose an LLM-based HSR method that leverages the semantic reasoning capabilities of large language models to evaluate multiple hypotheses simultaneously and select a more informative, diverse, and error-tolerant subset. This refined list serves as a higher-quality input for downstream post-processing models, including rerankers and correctors.

We evaluate the effectiveness of the LLM-based HSR by integrating it into several existing post-processing models.

979-8-3315-7206-8/25/$31.00

Additionally, we explore the use of LLMs for direct reranking and correction tasks, adapting prompt-based strategies for these purposes [10], [16]. All experiments are conducted on two widely used ASR benchmarks, AISHELL-1 [20] and LibriSpeech [21], compiled from the HypR benchmark [10]. Experimental results demonstrate that employing LLMs for hypothesis set refinement yields superior outcomes compared to conventional strategies. Furthermore, the combination of LLM-based HSR with LLM-based post-processing achieves the best overall ASR performance, highlighting the transformative potential of large language models in ASR post-processing pipelines.

II. Related Work

A. ASR Post-processing Methods

1) N-best Reranking: In ASR, the n-best list generated during decoding often contains better candidates than the top-1 output. To exploit this, n-best reranking techniques have been developed to re-evaluate hypotheses using more sophisticated scoring mechanisms. Early reranking approaches applied linear interpolation between acoustic and language model scores—methods that were computationally efficient but limited in capturing higher-level semantics. More recent advancements fall into three main categories: token-level, sentence-level, and comparison-based methods.

Token-level methods assess each candidate at the token level using pretrained language models such as GPT-2 [22] and BERT [23], enabling fine-grained evaluation of fluency and correctness. Sentence-level methods encode the entire hypothesis to capture global semantics. Notable examples include PBERT [13] and RescoreBERT [24], which combine contextual embeddings with original model scores for improved sentence-level evaluation. Comparison-based approaches directly compare hypotheses using Siamese or cross-encoder architectures, such as BERT_{sem} and $\text{BERT}_{\text{alsem}}$ [12], often leveraging contrastive learning to differentiate subtle semantic differences. Collectively, these reranking strategies have substantially improved the handling of linguistically ambiguous or semantically rich ASR outputs.

2) Error Correction: Despite significant progress in ASR models, recognition errors remain common due to noise, accents, and modeling limitations. To mitigate these errors, a range of post-processing correction methods has been proposed. Early approaches relied on rule-based systems or statistical models to fix frequent error patterns, but these lacked robustness and failed to account for contextual semantics.

Recent methods adopt neural sequence-to-sequence frameworks, treating ASR outputs as noisy text and learning to generate corrected versions. One prominent model is BART [14], which uses a denoising autoencoder with an encoder-decoder structure. BART has demonstrated strong performance in correcting insertion, deletion, and substitution errors by capturing both local and global contextual dependencies. Further work has extended BART-based architectures with domain adaptation and hybrid modules, improving coherence and robustness, particularly in noisy or domain-specific ASR scenarios. These approaches have shown significant improvements in word error rates and overall transcript quality.

Interaction with ChatGPT

User: 假設我有一些 n-best Hypotheses List 要做 Re-ranking 任務，如果這些 n-best Hypotheses List 有什麼樣的特性有可能會使 Re-ranking 後的結果更準確?

ChatGPT: 這是一個非常關鍵的問題... 以下是各項特性及其可能影響：1. High Diversity（高度多樣性）... 2. High Coverage of the Correct Hypothesis（包含正確答案的覆蓋度高）... 3. Rich and Discriminative Features(富含可判別性特徵)... 4. Score Separability（分數可區分性）... 5. Reasonable High-Confidence Hypotheses（合理可信的高信心句）... 6. Contextual and External Knowledge Utilization（語境與外部知識的可利用性）... 。

Fig. 1: ChatGPT interaction illustrating its understanding of useful properties for effective hypothesis set refinement.

B. Large Language Models

Large language models (LLMs) have revolutionized natural language processing by scaling model parameters and training data, resulting in substantial gains in understanding and generation. Models such as GPT-4 (OpenAI) [25], PaLM (Google) [26], and LLaMA (Meta) [27] have demonstrated few-shot learning, multilingual reasoning, and even cross-modal capabilities, signaling a transition toward general-purpose intelligence.

LLMs now achieve state-of-the-art performance across a wide range of tasks, including summarization, translation, question answering, and code generation [28]. They are also increasingly applied to speech-related tasks such as ASR correction and spoken dialogue generation, due to their strong semantic understanding and flexible reasoning abilities [16]. These developments highlight the potential of LLMs as powerful building blocks in ASR post-processing pipelines.

C. Hypothesis Set Refinement

While most ASR post-processing research focuses on reranking or correcting recognition results, a growing line of work investigates improving the composition of the n-best hypothesis list itself. This task, known as *Hypothesis Set Refinement* (HSR), aims to select a more effective subset of hypotheses to facilitate downstream processing.

One representative method is diversity-based selection, which computes pairwise edit distances among hypotheses and selects a subset with the highest overall diversity [19]. The intuition is that a diverse set may help rerankers or correctors distinguish between competing interpretations. However, excessive diversity can lead to the exclusion of correct or near-correct hypotheses, thus hurting final performance. To address this, several heuristics for selecting pseudo ground truths—such as the top-1 hypothesis, the most central candidate, or a corrected top-1 version—have been proposed [19].

请从以下 30-best hypotheses list 中选出 10 句 hypotheses，使其在后续 reranking 时的表现优于直接使用 top-10-best 进行 reranking。
请严格按照输出格式范例输出，请勿输出其他内容。
输入格式为 JSON：

```
[
  {
    "idx": 0,
    "hypo": "hypothesis 0 context",
    "score": score0
  },
  {
    "idx": 1,
    "hypo": "hypothesis 1 context",
    "score": score1
  },
  ...
]
```

其中：
- "idx" 是该句 hypothesis 的索引（范围 0～29）。
- "hypo" 是该句的转录文字。
- "score" 是该句的 am+lm 分数。

请输出一个 list，包含 10 条选出的 hypothesis 的"idx"。
输出格式范例：
[idx0, idx1, ..., idx9]
其中：
- idx0～idx9 必须在 0～29 之间的整数。
- idx0～idx9 按升序排序。

输入：

(a) Chinese Prompt

Please select 10 hypotheses from the following 30-best hypotheses list, such that their performance in subsequent reranking is better than directly using the top-10-best for reranking.
Strictly follow the output format example provided. Do not output anything else.
The input format is JSON:

```
[
    {
        "idx": 0,
        "hypo": "hypothesis 0 context",
        "score": score0
    },
    ...
]
```

Where "idx" is the index (range 0–29), "hypo" is the transcription text, "score" is the combined am+lm score.
Please output a list containing the "idx" values of the 10 selected hypotheses. Output format example:

```
[idx0, idx1, ..., idx9]
```

Where:
- idx0 to idx9 must be integers between 0 and 29.
- idx0 to idx9 must be sorted in ascending order.

(b) English Prompt

Fig. 2: Prompts used in the proposed LLM-based hypothesis set refinement (LLM-HSR) method.

III. Proposed Methods

We propose an LLM-based ASR post-processing framework consisting of two key components. First, we introduce an LLM-based hypothesis set refinement method to select a high-quality subset of n hypotheses from an original m-best list, where $m \gg n$. The goal is to identify hypotheses that not only have low error rates but also enhance the effectiveness of downstream post-processing tasks such as reranking and error correction. Second, we explore the direct application of LLMs for ASR post-processing, including n-best reranking and error correction in a zero-shot manner.

A. LLM-based Hypothesis Set Refinement

Motivated by the impressive zero-shot capabilities of LLMs across a wide range of tasks, we investigate their potential to perform hypothesis set refinement. Specifically, we use ChatGPT-4o as the backbone of our LLM-based hypothesis set refinement (LLM-HSR) method, due to its strong semantic reasoning and practical accessibility [25]. While ChatGPT-4o is used in our implementation, the methodology can be extended to other LLMs [26], [27].

As a preliminary step, we evaluate whether LLMs possess relevant knowledge to support effective hypothesis set refinement. To this end, we conduct a qualitative probing experiment, with the interaction results summarized in Figure 1. In its responses, ChatGPT-4o identifies six desirable characteristics for a good hypothesis list: (1) high diversity, (2) high coverage of the correct hypothesis, (3) rich and discriminative features, (4) score separability, (5) reasonable high-confidence hypotheses, and (6) contextual and external knowledge utilization. Some criteria align with prior findings [19], suggesting that LLMs demonstrate a meaningful understanding of what constitutes a strong hypothesis set.

Building on this insight, we design prompt templates for both Chinese and English, as illustrated in Figure 2. Both prompts instruct the LLM to analyze a given m-best list and return a subset of n selected hypotheses in the form of their indices. The target is to produce a refined list that performs better in downstream tasks than the original top-n hypotheses.

To improve reliability and reduce hallucinations, we format both input and output using JSON, consistent with recent findings that structured formats help constrain LLM behavior [29]. After the LLM outputs the selected subset, the refined hypothesis list is passed to downstream post-processing models for performance comparison against the original n-best list.

B. LLM-based ASR Post-processing Strategies

Beyond hypothesis set refinement, we further investigate the use of LLMs for ASR post-processing, focusing on n-best reranking and error correction. These experiments aim to evaluate whether LLMs can independently identify or improve hypotheses in a zero-shot setting. For reranking, the LLM is presented with a set of hypotheses and prompted to output the index of the most likely correct sentence. Example prompts for Chinese and English datasets are shown in Figure 3.

Error correction, by contrast, poses greater challenges. Our

LLM-Based Reranking Example on the AISHELL-1 Dataset
你是一个优秀的语言模型，你的任务是从下列十个句子中挑选出一句最正确的句子，你只需要回答数字题号 (1 ∼ 10)： 请严格的输出题号 () 中的数字，请勿输出其他内容。 (1) 甚至出现交易几乎停滞的情况 ... (10) 甚至出现交易几乎停职的情况 最正确的句子是：
ChatGPT: (1)

(a) Prompt for LLM-based reranking in Chinese.

LLM-Based Reranking Example on the LibriSpeech Dataset
You are an excellent language model. Your task is to select the most correct sentence from the following ten sentences. You only need to answer with the number (1–10): Please strictly output only the number inside the parentheses (), and do not output any other content. (1)stuffed into you his belly counselled him ... (10)stuffed in to you his belly counselled him The most accurate sentence is:
ChatGPT: (3)

(b) Prompt for LLM-based reranking in English.

Fig. 3: Prompt examples for the LLM-based reranking task on (a) Chinese and (b) English datasets.

preliminary results show that performance is highly sensitive to prompt phrasing. To address this, we carefully craft separate prompts for Chinese and English to ensure clarity and effectiveness in each language. These prompts are illustrated in Figure 4.

IV. Experiment Setup

As mentioned earlier, we use ChatGPT-4o as the large language model in this study, accessed via the official OpenAI API for all interactions. All experiments are conducted on the HypR benchmark [10], which provides ASR-generated hypothesis lists along with detailed scoring information. To evaluate the performance of our proposed method across different languages, we compile the Chinese AISHELL-1 [20] and English LibriSpeech [21] datasets from HypR. For each utterance, the ASR model produces a 50-best hypothesis list using beam search. Due to computational resource constraints, we limit the candidate set to the top 30 hypotheses for each speech utterance in our experiments.

Our experimental design is two-fold:

- **Hypothesis Set Refinement**: We investigate whether LLMs can be used to select a refined subset of 10 hypotheses from the top 30 candidates, aiming to improve the effectiveness of subsequent ASR post-processing.
- **Post-processing Evaluation**: We compare traditional and LLM-based post-processing strategies using different input sets, including the original top-10 and top-30 lists, diversity-based HSR selections, and our proposed LLM-HSR results. We evaluate the effectiveness of each post-processing method based on its ability to select or correct ASR recognition results.

LLM-Based Correction Example on the AISHELL-1 Dataset
你是一个优秀的语言模型，你的任务是参考下列的十个由语音辨识系统产生的句子，根据语意生成出最有可能的结果： 请严格的将生成的句子放置于「」中，请勿输出其他内容。 (1) 甚至出现交易几乎停滞的情况 ... (10) 甚至出现交易几乎停职的情况 最正确的句子是：
ChatGPT: 「甚至出现交易几乎停滞的情况」

(a) Prompt for LLM-based correction in Chinese.

LLM-Based Correction Example on the LibriSpeech Dataset
You are given the top 10 hypotheses generated by an Automatic Speech Recognition (ASR) system for a single spoken utterance. These hypotheses are listed below in descending order according to their ASR posterior scores, with <hypothesis1> being the highest-ranked. <hypothesis1>stuffed into you his belly counselled him</hypothesis1> ... <hypothesis10>stuffed in to you his belly counselled him</hypothesis10> Your task is to review these 10 hypotheses and produce the most accurate corrected transcription of the original utterance. Use information from across the hypotheses to identify and fix possible recognition errors. Return only the final corrected transcription as a single line of text. Do not include any explanations, comments, or any additional content—just the corrected result.
ChatGPT: stuff it into you his belly counselled him

(b) Prompt for LLM-based correction in English.

Fig. 4: Prompt examples for the LLM-based error correction task on (a) Chinese and (b) English datasets.

V. Experimental Results

A. Hypothesis Set Refinement

We first evaluate the effectiveness of different hypothesis set refinement strategies across both AISHELL-1 and LibriSpeech datasets. Tables I and II report character error rate (CER) results for various HSR methods, including original top hypotheses, diversity-based HSR strategies, and the proposed LLM-HSR method on the AISHELL-1 and LibriSpeech benchmarks, respectively. Top-1 and Top-10 are the original ASR-decoded top-ranked hypotheses, serving as the baseline. Div, Div+Top-1, Div+MBR, and Div+Cor are diversity-based methods adapted from [19] with different pseudo-references.

On both datasets, Div+Cor yields the best performance among the diversity-based methods, confirming prior findings that pseudo-references derived from corrected hypotheses are effective. However, we observe that only Div+Cor outperforms Top-10 when combined with traditional reranking and correction models on the AISHELL-1 dataset. In contrast, other diversity-based HSR methods consistently perform worse than Top-10 on both AISHELL-1 and LibriSpeech. These results suggest that diversity-based approaches may lack strong generalization ability. In comparison, our proposed LLM-HSR consistently surpasses all other HSR strategies, demonstrating the ability of large language models to leverage semantic understanding and contextual reasoning to select a high-quality, compact subset of hypotheses.

AISHELL-1	Top-1	Top-10	Div	Div+Top-1	Div+MBR	Div+Cor	LLM-HSR
Oracle	8.33	4.74	8.89	5.26	5.31	4.60	4.43
Reranking Method							
CLM	8.33	6.79	11.40	7.12	7.17	6.65	**6.61**
MLM	8.33	6.04	10.57	6.46	6.52	5.89	**5.76**
RescoreBERT	8.33	6.12	10.68	6.60	6.64	6.03	**5.85**
+MWER	8.33	6.11	10.70	6.59	6.66	6.01	**5.87**
+MWED	8.33	6.11	10.71	6.61	6.49	6.03	**5.84**
PBERT	8.33	5.89	10.44	6.42	6.49	5.83	**5.63**
$BERT_{sem}$	8.33	6.04	10.64	6.54	6.63	5.99	**5.79**
$BERT_{alsem}$	8.33	6.07	10.63	6.55	6.61	5.99	**5.82**
Correction Model							
BART	**7.44**	-	-	-	-	-	-
Proposed LLM-based							
Reranking	8.33	6.28	10.64	6.79	6.86	6.16	**6.11**
Correction	-	**5.20**	**5.03**	**5.57**	**5.65**	**5.22**	**5.10**

TABLE I: Experimental results on the AISHELL-1 dataset.

LibriSpeech	Clean							Other						
	Top-1	Top-10	Div	Div+Top-1	Div+MBR	Div+Cor	LLM-HSR	Top-1	Top-10	Div	Div+Top-1	Div+MBR	Div+Cor	LLM-HSR
Oracle	4.64	2.43	6.27	2.89	3.04	2.59	2.32	12.00	8.14	10.70	8.84	9.14	8.26	7.86
Reranking Method														
CLM	4.64	3.79	7.87	4.08	4.24	3.89	**3.77**	12.00	10.23	13.35	10.92	11.07	10.36	**10.11**
MLM	4.64	3.55	7.59	3.87	4.03	3.64	**3.47**	12.00	9.82	12.91	10.42	10.62	9.87	**9.56**
RescoreBERT	4.64	3.68	7.69	3.93	4.08	3.71	**3.60**	12.00	9.98	13.09	10.52	10.79	10.01	**9.79**
+MWER	4.64	3.67	7.75	3.96	4.09	3.72	**3.59**	12.00	9.92	13.02	10.50	10.77	10.00	**9.72**
+MWED	4.64	3.66	7.74	3.98	4.12	3.73	**3.60**	12.00	9.97	13.02	10.57	10.81	10.03	**9.77**
PBERT	4.64	3.63	7.73	3.95	4.07	3.73	**3.56**	12.00	9.96	12.98	10.59	10.81	10.01	**9.74**
$BERT_{sem}$	4.64	3.76	7.79	4.04	4.13	3.79	**3.68**	12.00	10.16	13.15	10.75	11.00	10.22	**9.91**
$BERT_{alsem}$	4.64	3.78	7.80	4.07	4.14	3.80	**3.71**	12.00	10.16	13.14	10.72	10.97	10.22	**9.92**
Correction Model														
BART	**4.50**	-	-	-	-	-	-	**11.26**	-	-	-	-	-	-
Proposed LLM-based														
Reranking	4.64	**4.49**	7.88	4.86	5.02	4.67	4.64	12.00	10.70	13.34	11.41	11.67	10.81	**10.67**
Correction	-	**3.97**	4.16	4.22	4.24	4.08	4.02	-	9.80	9.73	10.36	10.28	10.00	**9.67**

TABLE II: Experimental results on the LibriSpeech dataset under clean and other conditions.

B. Post-processing Evaluation

1) Reranking Performance: We evaluate representative reranking models alongside the proposed LLM-based reranking strategy using various input hypothesis sets. The baseline models include CLM [11], MLM [11], the RescoreBERT series [24], PBERT [13], and the BERT-based comparison models [12]. Across both datasets, the performance trends remain consistent: stronger rerankers such as PBERT and RescoreBERT outperform weaker models like CLM, regardless of the input hypothesis set. Notably, reranking performance improves when the hypothesis list is refined using our proposed LLM-HSR, indicating that selecting a higher-quality candidate set substantially benefits downstream processing.

Despite this, the LLM-based reranking method still lags behind traditional fine-tuned rerankers, surpassing only the CLM baseline on the AISHELL-1 dataset. This shortcoming may be attributed to the fact that ChatGPT is not explicitly trained for reranking tasks, which can limit its ability to reliably distinguish between hypotheses, especially when the differences involve subtle or similarly distributed errors.

2) Correction Performance: BART is one of the most widely used models for ASR result correction [15]. In contrast to LLM-based reranking, our experiments reveal a distinct performance trend for LLM-based correction. Although diversity-based methods like Div typically yield poor reranking outcomes due to the high error rates in the hypothesis list, their correction performance remains relatively stable. This indicates that LLMs are capable of mitigating noisy inputs through strong semantic understanding and generative capabilities.

Across both datasets, LLM-based correction consistently outperforms most traditional reranking and correction models. On AISHELL-1, the best results are achieved by combining LLM-HSR with LLM-based correction. On LibriSpeech, however, applying LLM-based correction directly to the original top-10 list produces the best outcome on the test-clean subset. This suggests that when the initial hypotheses are already of relatively high quality, further refinement through selection may offer limited additional gains.

These results highlight the robustness of the LLM-based correction strategy, which likely stems from the model's inherent ability to generate fluent, coherent, and semantically accurate text. This generative strength enables it to recover correct transcriptions even from noisy or error-prone inputs. Overall, the findings suggest that correction tasks may be more naturally aligned with the capabilities of LLMs than discriminative tasks like reranking.

VI. Conclusion

We propose a novel ASR post-processing framework centered on hypothesis set refinement (HSR) using large language models. By leveraging ChatGPT-4o to select a more informative and diverse subset of hypotheses, our method consistently improves the performance of downstream reranking and correction models across Chinese and English datasets. While LLM-based reranking still underperforms specialized models, LLM-based correction demonstrates strong and robust results, particularly when paired with refined hypothesis sets. Our findings highlight the practicality and flexibility of LLMs in ASR post-processing and point to future opportunities in prompt optimization and integration with fine-tuned models.

Acknowledgment

This work was supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC-112-2628-E-011-008-MY3 and Grant NSTC-114-2640-B-002-005; and in part by the Empower Vocational Education Research Center, National Taiwan University of Science and Technology, through the Featured Areas Research Center Program under the Higher Education Sprout Project, Ministry of Education, Taiwan. We would like to thank the National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

References

[1] W.-C. Huang, C.-H. Wu, S.-B. Luo, K.-Y. Chen, H.-M. Wang, and T. Toda, "Speech recognition by simply fine-tuning bert," *Proc. of ICASSP*, 2021.

[2] F.-H. Yu, K.-Y. Chen, and K.-H. Lu, "Non-autoregressive asr modeling using pre-trained language models for chinese speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1474–1482, 2022.

[3] C.-E. Lin and K.-Y. Chen, "A lexical-aware non-autoregressive transformer-based asr model," in *Proc. of Interspeech*, 2023.

[4] K.-H. Lu and K.-Y. Chen, "A context-aware knowledge transferring strategy for ctc-based asr," in *Proc. of SLT*, 2023.

[5] M. Nie, M. Yan, and C. Gong, "Prompt-based re-ranking language model for asr," in *Proc. of Interspeech*, 2022.

[6] I. E. Kang, C. Van Gysel, and M.-H. Siu, "Transformer-based model for asr n-best rescoring and rewriting," in *Proc. of Interspeech*, 2024.

[7] A. D. Tur, A. Moumen, and M. Ravanelli, "Progres: Prompted generative rescoring on asr n-best," in *Proc. of SLT*, 2024.

[8] Y. Li, X. Qiao, X. Zhao, *et al.*, "Large language model should understand pinyin for chinese asr error correction," in *Proc. of ICASSP*, 2025.

[9] C.-H. Kuo and K.-Y. Chen, "Correcting, rescoring and matching: An n-best list selection framework for speech recognition," in *Proc. of APSIPA*, 2022.

[10] Y.-W. Wang, K.-H. Lu, and K.-Y. Chen, "Hypr: A comprehensive study for asr hypothesis revising with a reference corpus," in *Proc. of Interspeech*, 2024.

[11] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," in *Proc. of ACL*, 2020.

[12] D. Fohr and I. Illina, "Bert-based semantic model for rescoring n-best speech recognition list," in *Proc. of Interspeech*, 2021.

[13] S.-H. Chiu and B. Chen, "Innovative bert-based reranking language models for speech recognition," in *Proc. of SLT*, 2021.

[14] M. Lewis, Y. Liu, N. Goyal, *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. of ACL*, 2020.

[15] Y. Zhao, X. Yang, J. Wang, Y. Gao, C. Yan, and Y. Zhou, "Bart based semantic correction for mandarin automatic speech recognition system," in *Proc. of Interspeech*, 2021.

[16] R. Ma, M. Qian, M. Gales, and K. Knill, "Asr error correction using large language models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1389–1401, 2025.

[17] M. L. Quatra, V. M. Salerno, Y. Tsao, and S. M. Siniscalchi, "Flanec: Exploring flan-t5 for post-asr error correction," in *Proc. of SLT*, 2024.

[18] J. Liang, *Perl: Pinyin enhanced rephrasing language model for chinese asr n-best error correction*, 2024. arXiv: 2412.03230.

[19] C.-H. Wu and K.-Y. Chen, "An n-best list selection framework for asr n-best rescoring," in *Proc. of O-COCOSDA*, 2024.

[20] 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 *Proc. of O-COCOSDA*, 2017.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. of ICASSP*, 2015.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI*, 2019, Accessed: 2024-11-15.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of ACL-HLT*, 2019.

[24] L. Xu, Y. Gu, J. Kolehmainen, *et al.*, "Rescorebert: Discriminative speech recognition rescoring with bert," in *Proc. of ICASSP*, 2022.

[25] OpenAI, J. Achiam, S. Adler, *et al.*, *Gpt-4 technical report*, 2024. arXiv: 2303.08774.

[26] A. Chowdhery, S. Narang, J. Devlin, *et al.*, "Palm: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 1, Jan. 2023.

[27] A. Grattafiori, A. Dubey, A. Jauhri, *et al.*, *The llama 3 herd of models*, 2024. arXiv: 2407.21783.

[28] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," in *Proc. of NIPS*, 2020.

[29] G. Mialon, R. Dessì, M. Lomeli, *et al.*, "Augmented language models: A survey," *Trans. Mach. Learn. Res.*, vol. 2023, 2023.