Contextualized Speech Recognition: Rethinking Second-Pass Rescoring with Generative Large Language Models

Yixuan Tang , Anthony K.H. Tung
Department of Computer Science, National University of Singapore
{yixuan, atung}@comp.nus.edu.sg

Abstract

Automatic Speech Recognition (ASR) systems have witnessed notable advancements in recent years. Contextualized ASR tasks require recognizing speech not as isolated utterances but within the broader context in which they occur. Conventional approaches often employ a second-pass paradigm to re-rank initial transcriptions, yet they risk propagating errors across candidate hypotheses, thereby compromising recognition precision. In this study, we introduce a novel framework that diverges from typical second-pass rescoring methods. Given n-best hypotheses, we leverage prompting with a large language model for contextualized second-pass generation. Besides pursuing higher accuracy, we aim to explore the performance boundaries without substantially altering the underlying pre-trained speech and language models. We investigate the effectiveness of the proposed paradigm through zero-shot prompting and strategic low-rank adaptation tuning. On the multi-accent spoken reading comprehension benchmark SQuAD-SRC, both prompting and fine-tuned models outperform the 1-best ASR hypothesis, achieving notable relative Word Error Rate (WER) improvements of 13.6% and 45.9%, respectively. The results suggest that the proposed approach enhances transcription accuracy and contextual understanding.

1 Introduction

Automatic Speech Recognition (ASR) models play a crucial role in various applications, ranging from virtual assistants and transcription services to accessibility tools. Despite significant recent progress in ASR [Radford *et al.*, 2023], recognizing speech in diverse contexts remains a challenge. Contextualized ASR tasks involve interpreting speech not just as individual utterances but within the broader context in which they occur. For instance, when transcribing short segments of lengthy lecture videos, context can provide essential information, such as topics and background details. Most existing contextualized ASR work focuses on contextual biasing towards a predefined list of rare named entities in specific domains [Sathyendra *et al.*, 2022; Sun *et al.*, 2023; Fu *et al.*, 2023]. In contrast, our work focuses on contextualized ASR with entire text passages, involving the comprehension of intricate linguistic subtleties, adaptation to varying speaking styles, and the ability to discern contextual cues, all contributing to the heightened complexity of the ASR task.

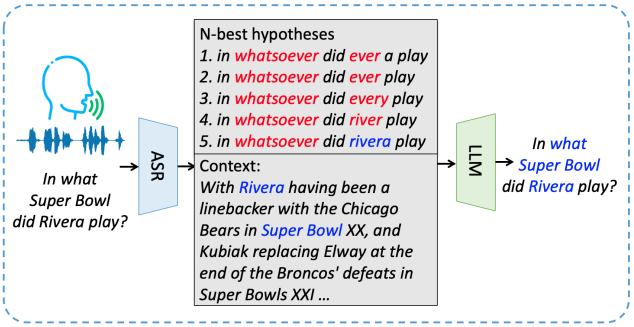

Figure 1: An illustrative example showcasing that second-pass generation with n-best ASR hypotheses and context allows it to exploit LLM for enhanced speech recognition.

Efforts to address contextualized ASR complexities fall into two main categories. One direction involves the fusion of acoustic representations extracted from the speech encoder and text representations of the context during decoding. These fusion approaches usually demand computationally expensive re-training of the entire network end-to-end, to enable better interaction between semantic spaces of speech and text modalities. In our work, we explore adapting the pre-trained speech model without extensive re-training, which proves more efficient, especially given the abundance of text-only training data.

The second direction involves a two-pass paradigm, where the first pass generates n-best hypotheses, and the second-pass subsequently re-scores these candidates to output the best transcript. While discriminative training with a Minimum Word Error Rate (MWER) during the second-pass rescoring phase enhances performance, it introduces a constraint by limiting the scope of candidate hypotheses to those generated in the initial pass. This limitation holds the potential to propagate errors across all candidate choices, leading to sub-optimal transcription outcomes. Figure 1 illustrates a scenario where the entity "Super Bowl" is inaccurately recog-

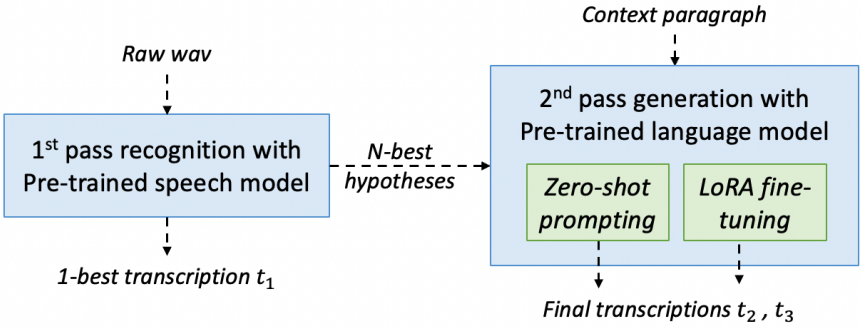

Figure 2: Overview of the proposed two-stage paradigm. The second-pass generation involves two settings: zero-shot prompting and fine-tuning with low-rank adapters.

nized in all n-best hypotheses, rendering second-pass rescoring ineffective even when the correct phrase is provided in the context passage.

To address these challenges, we introduce a paradigm-shifting framework that redefines the rescoring paradigm in ASR. Rather than confining the second-pass process to a predetermined set of candidate hypotheses, our approach harnesses the extensive linguistic knowledge embedded within large language models (LLMs) to generate contextually coherent transcriptions. This innovative methodology excels in capturing subtle contextual nuances and seamlessly adapting to diverse speech patterns.

Prompting emerges as a potent strategy for exploiting LLMs for ASR, providing users with the means to guide the model towards specific desired responses based on carefully crafted prompts. In our work, we leverage *Mistral* [Jiang *et al.*, 2023] and textual prompts for second-pass generation, incorporating n-best hypotheses and context paragraphs as input. Additionally, to optimize LLM performance on contextualized ASR, we introduce a further refinement by fine-tuning the model with low-rank adaptation (LoRA) [Hu *et al.*, 2022]. LoRA introduces adaptive learning rates across different layers, enhancing the LLM's capacity to generalize across diverse speech patterns and contextual variations. Through experimentation and evaluation, we show that our proposed methodology distinctly outperforms existing approaches, resulting in improved transcription accuracy and contextual fidelity. On the contextualized ASR dataset SQuAD-SRC [Tang and Tung, 2023], our proposed second-pass generation with zero-shot prompting and fine-tuning achieves a noteworthy relative WER improvement of 13.6% and 45.9%. We will release the code via https://github.com/tangyixuan/2ndPassContextASR.

To summarize, the main contributions of this paper are as follows:

- We propose a novel framework that seamlessly integrates large language models into a second-pass generation paradigm, overcoming the inherent limitations of existing solutions.
- By designing prompting and leveraging LoRA fine-tuning, we efficiently enhance the LLM's capacity to discern subtle contextual nuances and optimize its performance for contextualized ASR tasks.
- Through rigorous experimentation and evaluation, we demonstrate that our proposed approach achieves notable improvements in transcription accuracy and contextual understanding.

2 Problem Formulation

In specialized domains, such as medical, legal, or academic settings, the accuracy of ASR is significantly influenced by contextual cues embedded within textual paragraphs. The recognition of domain-specific information and terminologies demands a nuanced contextual understanding to ensure accurate transcription.

To formalize this task, let w denote a raw speech utterance, and p represent the corresponding contextual textual paragraph. Our objective is to develop a framework that generates a transcription $\hat{t}$ aligned with both w and p, i.e. $\hat{t} = f(w, p)$. Specifically, considering a list of n-best hypotheses generated in the first pass, denoted as $T = \{t_1, t_2, ..., t_n\}$, we aim to optimize the second-pass generation with respect to the context p, represented as $\hat{t} = g(T, p)$.

3 Methodology

3.1 Framework Overview

As illustrated in Figure 2, our approach to contextualized speech recognition follows a two-stage pipeline. In the initial stage, the proposed framework generates n-best candidate hypotheses using a pre-trained ASR model. The top-1 hypothesis serves as the baseline for comparison. Subsequently, a large language model is employed for the second-pass generation process. Instead of computing the probability of each hypothesis individually, all n-best hypotheses, along with the contextual information and an instructional prompt, are fed into the LLM for optimal transcription generation. We explore the LLM's performance in contextualized ASR under two settings: zero-shot prompting and fine-tuning with LoRA [Hu *et al.*, 2022].

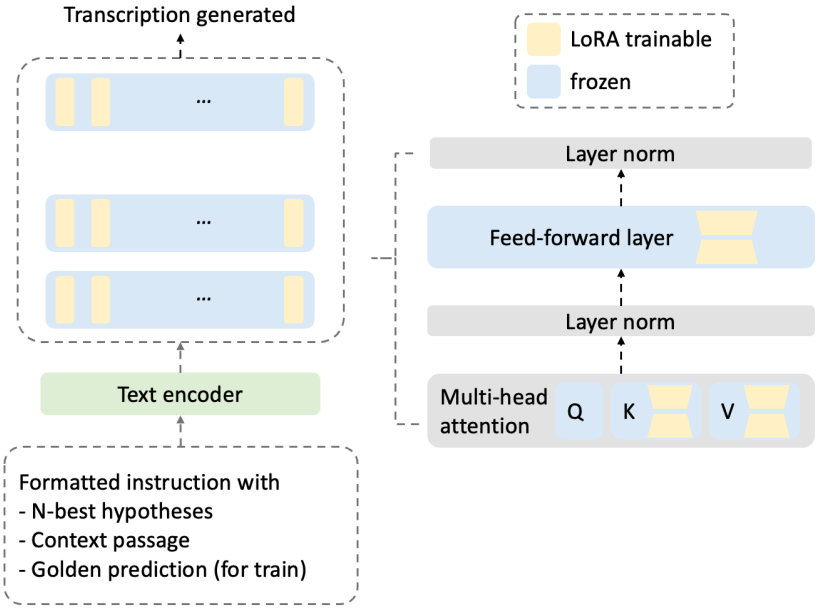

Figure 3: Model structure for second-pass regeneration with zero-shot prompting or low-rank adaptation.

3.2 Candidate Hypotheses Generation

To initiate the ASR pipeline, we leverage the capabilities of the pre-trained *Whisper* [1] [Radford *et al.*, 2023], an advanced auto-regressive transformer architecture designed for robust speech recognition. The *Whisper* model is configured to process raw audio waveforms, producing a set of n-best hypotheses during the decoding phase. Employing a beam search strategy, the model iteratively extends the sequence of tokens while maintaining a hypothesis set consisting of the top N sequences. At each step, the model samples the next token from its predicted probability distribution and updates the set based on the extended sequences' probabilities. This iterative process continues until reaching a maximum length or convergence, resulting in a diverse set of n-best hypotheses that represent alternative transcriptions, capturing various potential interpretations of the audio utterances. Among these hypotheses, the transcription with the highest probability is selected as the 1-best hypothesis baseline.

3.3 Second-Pass Re-Generation

Model Structure

Figure 3 illustrates the detailed model structure of our proposed framework. A distinctive feature is the shift from discriminative second-pass rescoring to second-pass generation. Following the acquisition of candidate hypotheses, we employ a pre-trained generative large language model to integrate both contextual information and speech-related details embedded within these hypotheses. It is noteworthy that, although we leverage *Mistral-7B-Instruct-v0.1* [2] [Jiang *et al.*, 2023] in our experiments, the framework is designed to accommodate any generative LLM. Unlike conventional rescoring strategies that treat each hypothesis independently, our approach involves concatenating all n-best hypotheses. We then feed them, along with the context passages structured by the prompt, to the LLM. This enables a more holistic integration of diverse interpretations and linguistic variations present in the candidate hypotheses, thereby enhancing the model's ability to generate contextually coherent transcriptions.

Zero-Shot Prompting

""" <S> [INST] The n-best hypotheses for a speech utterance from an ASR model are:
{hypo}
The content is a question for the following passage:
{title}{para}
Please report the true transcript of the speech utterance. [/INST]
The true "transcript of the speech utterance is: {gold}
</s>"""

Figure 4: Prompt template designed for zero-shot prompting and LoRA fine-tuning for second-pass generation. {gold} is provided during training and it's not available during inference.

Zero-shot prompting stands out as a powerful technique for directing large language models to perform specific tasks without the need for explicit fine-tuning or additional training data [Liu *et al.*, 2023]. This approach relies on well-crafted prompts to guide the model towards generating responses

[1] https://github.com/openai/whisper
[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

aligned with the intended objectives. Figure 1 illustrates the desired scenario of second-pass generation using prompting with an LLM, utilizing context to enhance the quality of generated transcripts.

In this work, we employ zero-shot prompting by presenting the model with multiple candidate hypotheses and context and instructing it to generate the correct transcription. The success of this strategy depends on the pre-trained LLM's intrinsic understanding of linguistic nuances, syntax, and semantics. While the core concept is straightforward, the process of designing effective prompts demands ingenuity and resourcefulness to ensure high-quality outcomes.

Despite the apparent simplicity of prompts, the intricacies of prompt design are nontrivial. The challenge lies in effectively infusing contextual cues without introducing redundant tokens. To strike this balance, we experimented with various versions of prompt templates. The template shown in Figure 4 represents the final decision to be adopted in our paradigm, where each individual hypothesis is separated by a newline character when formatted into the prompt. This template includes a simple description of the input contents and a direct instruction to generate the true transcription.

Our exploration also delved into few-shot prompting with demonstrations and chain-of-thought prompting, though yielding minimal gains. We attribute this to the fact that one example with 10-best hypotheses and context already reaches an effective modeling sequence length. The formulated instruction for one example already comprises a mean of 408 tokens.

Low-Rank Adaptation

To optimize the LLM for contextualized ASR tasks, we employ fine-tuning on *Mistral-7B-Instruct-v0.1* using instructions generated on the training set using the same prompt template as described in zero-shot prompting. However, full fine-tuning on an LLM is computationally expensive. Drawing inspiration from the success of low-rank adaptation in other NLP tasks [Hu *et al.*, 2022], we adopt LoRA fine-tuning to further adapt the LLM.

In the process LoRA fine-tuning, we strategically adjust only the Key-Value (KV) pairs within multi-head attention layers and weights in feed-forward layers of each transformer block. This focused tuning enhances the model's ability to discern subtle contextual nuances, incorporate domain-specific vocabulary, and optimize its performance for contextualized ASR applications while significantly reducing computational costs.

Training Objective

To train the prompt template filled with concatenated n-best hypotheses, context, and reference transcript, we utilize an auto-regressive loss. This loss encourages the model to generate specified tokens sequentially, and the optimization of auto-regressive training loss ($\mathcal{L}$) is defined as:

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{|y|} \log P(\hat{y}_i | \hat{y} < i; \theta) \quad (1)$$

where $|\cdot|$ computes the cardinality of a sequence, $\hat{y}$ represents the predicted sequence, y denotes the ground truth sequence, and θ embodies the trainable parameters of LLM. This negative log loss formulation guides the model to produce sequences that align closely with the ground truth.

# N-best hypos	WER	Relative Improvement
1	24.2	baseline
3	23.1	4.5 %
5	22.3	7.9 %
10	20.9	13.6 %

Table 1: Performance of zero-shot prompting on SQuAD-SRC dev dataset with different number of n-best hypotheses used in the prompt for second-pass generation.

4 Experiments

In this section, we provide a comprehensive overview of the experimental setup, detailing the dataset, training specifics, baseline model, and evaluation metrics employed. Subsequently, we present the main experimental findings, delving into the impact of the number of candidate hypotheses generated in the first pass and analyzing the performance gain achieved by increasing the number of instruction pairs used for fine-tuning the model.

4.1 Experimental Settings

Dataset We evaluate the proposed two-stage paradigm for contextualized ASR on the SQuAD-SRC [Tang and Tung, 2023] dataset, employing both zero-shot prompting and LoRA fine-tuned models. The SQuAD-SRC dataset is a multi-accent spoken reading comprehension dataset, comprising 98,169 spoken questions. A textual passage is provided for each question which contains the answer. The audio utterances are recorded by 24 qualified speakers from six different countries, including the US, UK, China, India, Japan, and Thailand, thereby providing a challenging set of diverse English accents. In our experiments, the spoken questions are taken in the raw audio waveforms, while the corresponding textual passages offer contextual information. The primary objective for contextualized ASR is to transcribe the spoken questions' text.

Implementation Details In our experimental setup, we initiate by creating training instructions, formed by incorporating hypotheses and context through the prompt template specifically designed for contextualized ASR. We vary the number of n-best hypotheses employed in each example, exploring values of n = 1, 3, 5, and 10. The model is initialized with *Mistral-7B-Instruct-v0.1*. Subsequently, we systematically assess the performance of the proposed framework using different quantities of training instructions, ranging from 100, 500, 1000, to 2000. The selection of the batch size, which can be 1, 2, 4, or 8, depends on the number of hypotheses in each example and consequently influences the string length. Employing the paged Adam optimizer, we investigate learning rates of 2×10^{-4}, 1×10^{-4}, and 1×10^{-3}, choosing the optimal value. The model is trained for 5 epochs, with early stopping implemented to evaluate the checkpoint dis-

# training examples	hours	Whisper-tiny		Whisper-base		Whisper-small		Whisper-medium	
		WER	RI	WER	RI (%)	WER	RI (%)	WER	RI (%)
0	0	24.2	baseline	19.7	baseline	17.0	baseline	16.6	baseline
100	0.2	19.4	19.8%	14.9	24.4%	14.7	13.5%	14.6	12.0%
500	0.8	15.0	38.0%	13.0	34.0%	12.8	24.7%	13.3	19.9%
1000	1.4	13.6	43.8%	12.3	37.6%	11.6	31.8%	12.0	27.7%
2000	2.9	13.1	45.9%	12.1	38.6%	11.5	32.4%	11.9	28.3%

Table 2: Performance of LoRA finetuned model for second-pass generation with different amounts of training data from SQuAD-SRC train set on different configurations of backbone ASR model. "# Pairs" denotes the number of hypotheses-transcription pairs, and "Duration" represents the corresponding speech duration.

Reference	Which hotel did the **Panthers** stay at for the **Super Bowl?**
N-best hypotheses	1. What hotel did the **Panther** stay at for the **sugar bowl**? 2. What hotel did the **panther** stay at for the **soup bowl**? 3. What hotel did the **panther** stay at for the **supper belll**? 4. What hotel did the **panther** stay at the **supper bell**? 5. What hotel did the **Panthers** stay at for the **soup bowl**?
Zero-shot prompting	What hotel did the **panther** stay at for the **sugar bowl**?
LoRA w #100 training egs	What hotel did the **panther** stay at for the **sugar bowl**?
LoRA w #1k training egs	What hotel did the **Panthers** stay at for the **Super Bowl**?
LoRA w #2k training egs	What hotel did the **Panthers** stay at for the **Super Bowl**?

Table 3: Examples of transcripts recognized with / without second-pass generation under different settings.

playing the best performance on the development set. The training process is executed on one A100 card.

Baseline To establish a baseline, we use the 1-best candidate hypothesis with the highest probability during beam-search decoding. Specifically, we deploy state-of-the-art *Whisper* [Radford *et al.*, 2023] models for candidate hypothesis generation.

Evaluation Metrics To quantify the effectiveness of speech recognition, the standard word-error-rate (WER) is used as the evaluation metric. WER is calculated as the sum of substitutions, deletions, and insertions divided by the total number of words in the reference text:

$$\text{WER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{Total Words in Reference}}$$

Additionally, we present relative WER improvement by comparing each approach against the baseline. The relative improvement computes the percentage improvement in WER by comparing the WER of the proposed method with the WER of baseline approach.

$$\text{Relative Improvement} = \frac{\text{Baseline WER} - \text{Improved WER}}{\text{Baseline WER}}$$

4.2 Zero-shot Prompting Performance Analysis

Table 1 presents experimental results of zero-shot prompting on the SQuAD-SRC development dataset, showcasing the impact of varying the number of n-best hypotheses used in the prompt for second-pass generation. The baseline performance with 1-best hypothesis is recorded at 21.1% WER.

As we progressively increase the number of candidate hypotheses to 3, 5, and 10, we observe a consistent reduction in WER. Specifically, employing 3-best hypotheses yields a WER of 20.4%, representing a 3.3% relative improvement over the baseline. The trend continues, with 5-best hypotheses resulting in a WER of 19.5% (7.6% relative improvement), and 10-best hypotheses achieving the lowest WER of 18.8% (10.9% relative improvement).

These results indicate that incorporating a greater number of diverse hypotheses during the second-pass generation process enhances the model's ability to capture contextual nuances and improve overall transcription accuracy. The observed reduction in WER demonstrates the effectiveness of leveraging multiple hypotheses for contextualized ASR, highlighting the potential of our proposed framework in accommodating varied interpretations and improving performance on challenging spoken language understanding tasks.

4.3 LoRA Fine-tuned Performance

Table 2 presents the experimental results of our proposed paradigm after LoRA fine-tuning for second-pass generation, illustrating the impact of varying amounts of training data. We evaluate the paradigm using various configurations of *Whisper*, with the baseline being the best candidate hypothesis generated by the original *Whisper*. Each example is evaluated using 10 candidate hypotheses.

The results demonstrate a consistent trend across all back-

bone ASR configurations, showing a decrease in WER and an increase in relative improvement (RI) as the number of examples used for fine-tuning increases. This trend highlights the effectiveness of our framework across different backbone ASR sizes. Notably, even with just 100 training examples (0.2 hours total speech duration), a significant performance improvement is observed.

Comparing different *Whisper* configurations, *Whisper-tiny* achieves the most substantial gain with a 45.9% RI. After fine-tuning on 2000 examples, our paradigm significantly reduces the WER difference between *Whisper-tiny* and *Whisper-medium* from 7.6 to 1.2,. Fine-tuned *Whisper-tiny* surpasses the performance of the original *Whisper-medium*, which contains about 20 times more parameters than it.

These results underscore the impact of LoRA fine-tuning on the proposed framework's performance with minimal training data. The model effectively captures contextual information, leading to notable improvements in transcription accuracy. These findings highlight the potential of our approach to adapt to diverse speech patterns and enhance contextualized ASR tasks effectively.

4.4 Qualitative Analysis

Table 3 presents a qualitative analysis of ASR transcriptions under different settings, comparing the n-best hypotheses generated in the first pass with the results obtained through zero-shot prompting and LoRA fine-tuning.

In the n-best hypotheses, word errors are obvious, including wrongly recognized terms like "Panther" and "panther", "sugar bowl", "soup bowl", and "supper bell".

Under zero-shot prompting, the recognition is consistent with the prevalent errors in the n-best hypotheses, indicating limited improvement in transcription accuracy for some examples.

As the number of training examples used in LoRA fine-tuning increases, the model exhibits an improved ability to learn from context passages and correct mistakes present in all n-best hypotheses. Upon increasing the training data to 1,000 pairs, the LoRA fine-tuned model showcases significant improvement, accurately transcribing the reference question as "What hotel did the Panthers stay at for the Super Bowl." Further improvement is observed with 2,000 pairs, reinforcing the efficacy of LoRA fine-tuning in addressing errors present in the initial hypotheses and achieving enhanced transcription accuracy and contextual understanding.

5 Related Work

Contextualized Speech Recognition Addressing the challenges of contextualized ASR has been a focal point in recent research endeavors. Traditional approaches in contextualized ASR often centered on context biasing towards a predefined set of rare named entities [Sun *et al.*, 2023; Chang *et al.*, 2021; Le *et al.*, 2021; Sathyendra *et al.*, 2022; Fu *et al.*, 2023] in specific domains [Yu *et al.*, 2023; Li *et al.*, 2023; Yang *et al.*, 2023]. These conventional methods typically involve incorporating contextual information related to a handful of specific entities, tailoring ASR models to recognize and transcribe these entities more accurately. However, these approaches often fall short when confronted with the intricate linguistic subtleties present in extended passages or varied speaking styles . Our work takes a departure from this entity-centric focus and aims to contextualize ASR on a broader scale, dealing with entire text passages [Shenoy *et al.*, 2021; Chang *et al.*, 2021]. This shift introduces new challenges, including the need for the model to comprehend more nuanced context, adapt to diverse linguistic variations, and accurately discern contextual cues, contributing to the heightened complexity of the ASR task.

Leveraging Language Models for ASR Enhancement The integration of pre-trained language models to augment tasks in speech recognition has attracted considerable attention. Capitalizing on the extensive linguistic knowledge embedded within LLMs emerges as a promising avenue to enhance transcription accuracy and contextual understanding. Prior studies have explored various applications of language models in ASR, encompassing the fusion of hidden representations of acoustic and textual features in both shallow and deep manners [Huang *et al.*, 2023; Ogawa *et al.*, 2023; Chang *et al.*, 2021; Han *et al.*, 2022; Chang *et al.*, 2023], as well as the incorporation of modality adaptation layers to process speech encoder outputs as inputs for generative LLMs [Radhakrishnan *et al.*, 2023; Wu *et al.*, 2023; Shu *et al.*, 2023; Shukor *et al.*, 2023; Hono *et al.*, 2023; Chen *et al.*, 2023]. Many of these approaches require end-to-end training of speech-text pairs, a time-consuming process. Our work aligns with this overarching trend but introduces a novel approach by integrating LLMs into a second-pass generation paradigm, instead of second-pass rescoring [Li *et al.*, 2023; Shivakumar *et al.*, 2023]. This innovative strategy allows us to harness LLMs for contextually guided transcription generation, unlocking the potential to capture intricate contextual nuances and seamlessly adapt to diverse speaking styles. This marks a departure from traditional approaches that often necessitate exhaustive re-training, highlighting the efficiency and effectiveness of leveraging LLMs to enhance ASR capabilities in a more streamlined manner.

6 Conclusion & Future Work

To conclude, we propose a novel framework for contextualized ASR that diverges from traditional second-pass rescoring paradigms. By utilizing large language models for contextualized second-pass generation, our approach aims to improve contextual understanding and transcription accuracy. Our experiments on the SQuAD-SRC dataset demonstrate the effectiveness of the proposed framework. Both zero-shot prompting and fine-tuning with low-rank adaptation significantly outperformed the 1-best hypothesis, achieving relative word error rate reductions of 13.6% and 45.9%, respectively. Qualitative analysis further highlights the model's ability to correct errors that conventional second-pass rescoring cannot handle.

For future work, we plan to explore the generalization of the second-pass generation paradigm across different domains and languages. Incorporating domain information as context in the contextualized ASR framework could help address domain-specific speech recognition challenges.

Acknowledgments

This research is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001), the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative, and the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-029). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore and National Research Foundation, Singapore.

References

[Chang *et al.*, 2021] Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann. Context-aware transformer transducer for speech recognition. In *ASRU*, pages 503–510. IEEE, 2021.

[Chang *et al.*, 2023] Shuo-Yiin Chang, Chao Zhang, Tara N. Sainath, Bo Li, and Trevor Strohman. Context-aware end-to-end ASR using self-attentive embedding and tensor fusion. In *ICASSP*, pages 1–5. IEEE, 2023.

[Chen *et al.*, 2023] Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C. Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. SALM: speech-augmented language model with in-context learning for speech recognition and translation. *CoRR*, abs/2310.09424, 2023.

[Fu *et al.*, 2023] Xuandi Fu, Kanthashree Mysore Sathyendra, Ankur Gandhe, Jing Liu, Grant P. Strimel, Ross McGowan, and Athanasios Mouchtaris. Robust acoustic and semantic contextual biasing in neural transducers for speech recognition. In *ICASSP*, pages 1–5. IEEE, 2023.

[Han *et al.*, 2022] Minglun Han, Linhao Dong, Zhenlin Liang, Meng Cai, Shiyu Zhou, Zejun Ma, and Bo Xu. Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection. In *ICASSP*, pages 8532–8536. IEEE, 2022.

[Hono *et al.*, 2023] Yukiya Hono, Koh Mitsuda, Tianyu Zhao, Kentaro Mitsui, Toshiaki Wakatsuki, and Kei Sawada. An integration of pre-trained speech and language models for end-to-end speech recognition. *CoRR*, abs/2312.03668, 2023.

[Hu *et al.*, 2022] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.

[Huang *et al.*, 2023] Kaixun Huang, Ao Zhang, Binbin Zhang, Tianyi Xu, Xingchen Song, and Lei Xie. Spike-triggered contextual biasing for end-to-end mandarin speech recognition. *CoRR*, abs/2310.04657, 2023.

[Jiang *et al.*, 2023] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023.

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

[Li *et al.*, 2023] Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. Prompting large language models for zero-shot domain adaptation in speech recognition. *CoRR*, abs/2306.16007, 2023.

[Liu *et al.*, 2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35, 2023.

[Ogawa *et al.*, 2023] Atsunori Ogawa, Takafumi Moriya, Naoyuki Kamo, Naohiro Tawara, and Marc Delcroix. Iterative shallow fusion of backward language model for end-to-end speech recognition. In *ICASSP*, pages 1–5. IEEE, 2023.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023.

[Radhakrishnan *et al.*, 2023] Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A. Kiani, David Gomez-Cabrero, and Jesper Tegnér. Whispering llama: A cross-modal generative error correction framework for speech recognition. In *EMNLP*, pages 10007–10016. Association for Computational Linguistics, 2023.

[Sathyendra *et al.*, 2022] Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P. Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann. Contextual adapters for personalized speech recognition in neural transducers. In *ICASSP*, pages 8537–8541. IEEE, 2022.

[Shenoy *et al.*, 2021] Ashish Shenoy, Sravan Bodapati, Monica Sunkara, Srikanth Ronanki, and Katrin Kirchhoff. Adapting long context NLM for ASR rescoring in conversational agents. In *Interspeech*, pages 3246–3250. ISCA, 2021.

[Shivakumar *et al.*, 2023] Prashanth Gurunath Shivakumar, Jari Kolehmainen, Yile Gu, Ankur Gandhe, Ariya Rastrow, and Ivan Bulyko. Discriminative speech recognition rescoring with pre-trained language models. *arXiv preprint arXiv:2310.06248*, 2023.

[Shu *et al.*, 2023] Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. Llasm: Large language and speech model. *CoRR*, abs/2308.15930, 2023.

[Shukor *et al.*, 2023] Mustafa Shukor, Corentin Dancette, Alexandre Ramé, and Matthieu Cord. Unified model for image, video, audio and language tasks. *CoRR*, abs/2307.16184, 2023.

[Sun *et al.*, 2023] Chuanneng Sun, Zeeshan Ahmed, Yingyi Ma, Zhe Liu, Lucas Kabela, Yutong Pang, and Ozlem Kalinli. Contextual biasing of named-entities with large language models. *CoRR*, abs/2309.00723, 2023.

[Tang and Tung, 2023] Yixuan Tang and Anthony K. H. Tung. Squad-src: A dataset for multi-accent spoken reading comprehension. In *IJCAI*, pages 5206–5214. ijcai.org, 2023.

[Wu *et al.*, 2023] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal LLM. *CoRR*, abs/2309.05519, 2023.

[Yang *et al.*, 2023] Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. Generative speech recognition error correction with large language models and task-activating prompting. pages 1–8, 2023.

[Yu *et al.*, 2023] Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth Gurunath Shivakumar, Yile Gu, Sungho Ryu, Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, Tuan Dinh, Ankur Gandhe, Denis Filimonov, Shalini Ghosh, Andreas Stolcke, Ariya Rastrow, and Ivan Bulyko. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. *CoRR*, abs/2309.15223, 2023.