Error Propagation in ASR Pipelines: Investigating Language Model Limitations and Amplification Effects

Hanna Han
Department of Linguistics and Philology
Uppsala University
hanna.han.4075@student.uu.se

Abstract

Language Models (LMs) play a crucial role in enhancing the performance of Automatic speech recognition (ASR) systems by aiding the decoding process in acoustic models. Furthermore, LMs are often employed in post-correction to improve transcription accuracy. This study specifically examines the impact of LMs during the post-correction phase of the ASR workflow. A significant limitation of LMs is their inability to detect errors generated in earlier stages, such as by the acoustic model. When LMs receive erroneous input, they may produce syntactically and semantically incorrect predictions, potentially amplifying initial errors and degrading the overall transcription quality. For the experiments, an ASR system consisting of separate modular components is utilized. Two types of LM, an n-gram model and a transformer-based model, are applied to the post-correction of outputs from the acoustic model using two different test datasets. Evaluation metrics include Word Error Rate (WER) and SeMaScore (Sasindran et al., 2024). The results show that post-correction with both LMs often leads to degraded performance in terms of both metrics. In particular, the n-gram model significantly worsens the transcription quality, demonstrating the amplification of acoustic model errors during post-correction process.

1 Introduction

ASR systems provide valuable services by making digital content more accessible and improving user experiences in various applications, such as translation, transcription generation, voice search, and customer service. Before the introduction of End-to-End (E2E) ASR models (Prabhavalkar et al., 2023) such as Whisper (Radford et al., 2023), which are based on a transformer architecture (Vaswani, 2017), ASR systems were typically constructed with several modular components, including acoustic models, LMs, and pronunciation dictionaries that work together to transcribe speech into text. These traditional systems heavily relied on statistical methods, but their modularity allowed the integration of deep learning techniques into specific components to improve performance. In contrast, E2E ASR models streamline the process by using a single neural network that directly maps audio input to text output, bypassing the need for separate components such as acoustic models and LMs. Although this unified architecture simplifies model development and maintenance, it reduces the ability to customize individual components.

The main focus of this research is to explore the role of language models in ASR systems, specifically their effect during external post-correction. In particular, the study examines whether LMs amplify errors introduced by the acoustic model and assesses their effect in improving ASR performance. To conduct a thorough comparison, it is essential to have full control over the ASR pipeline, allowing a detailed evaluation of the system's performance both with and without the use of language models. Thus, we chose a traditional ASR model provided by Kaldi (Povey et al., 2011) [1], which was pre-trained on the 960 hours of the Librispeech dataset (Panayotov et al., 2015). More details on the ASR model will be provided in Section 3.2.

The experiments are conducted using two distinct test datasets in English: a Librispeech clean test set and Common Voice test set provided by OpenSLR [2] and Mozilla community respectively. A detailed description of these datasets is provided in Section 3.1. Two types of language models are used for the external post-correction: a statistical n-gram model and a transformer-based model. By

[1] https://kaldi-asr.org/
[2] https://openslr.org/

comparing these two models, which differ in their training approaches—one relying on deep learning and the other on statistical methods—this research aims to inform future decisions in ASR model development.

For the evaluation of ASR output, WER and SeMaScore (Sasindran et al., 2024) are used. WER is a widely used metric for assessing ASR accuracy. However it does not fully capture the contextual relevance, nor does it correlate with the performance of downstream tasks like large language models (LLMs). To gain a more comprehensive understanding of the ASR system's performance, both WER and SeMaScore are employed to evaluate the accuracy of words and the contextual relevance of the transcriptions.

2 Related works

In recent years, various neural network techniques such as Connectionist Temporal Classification (CTC) (Graves et al., 2006), Recurrent Neural Network Transducer (RNN-T) (Graves, 2012), and attention-based encoder-decoder (AED) architectures (Chan et al., 2016; Chorowski et al., 2015) have been integrated into ASR systems (Dhanjal and Singh, 2024) to enhance performance. They have made significant contributions to the field, such as simplifying system architectures, improving transcription accuracy, and enabling the handling of long contextual sentences among other benefits.

Despite the critical role that LMs play in ASR systems, relatively little research has been devoted to systematically exploring their impact or advancing their development for improving ASR performance. This gap has inspired my study, which explores effective approaches for incorporating LMs into ASR systems. Specifically, this research examines the role of LMs in addressing erroneous transcriptions produced by acoustic models, marking an essential first step in understanding their potential before delving into more advanced methods of leveraging LMs in future studies.

The experiments conducted in this paper extend the work of Zeping and Jinbo (2023) who investigated the use of LLMs for post-correction in ASR systems with a focus on enhancing transcription accuracy. In their study, the authors employed a pre-trained E2E ASR model provided by the WeNet speech community (Yao et al., 2021) and explored the in-context learning capabilities of ChatGPT and GPT-4 for error correction. To provide a clearer environment for evaluating the role of LMs, we adopts Kaldi's ASR model, which follows a traditional modular architecture because E2E ASR models have challenges in isolating the specific contribution of LMs. This approach allows greater control over the integration of LMs within the ASR pipeline. Unlike the LLMs employed in Zeping and Jinbo (2023) experiments, this research compares an n-gram-based model and a transformer-based model. In Zeping and Jinbo's study, three versions of ChatGPT and GPT-4 were tested, based on the hypothesis that their advanced in-context learning capabilities would yield better transcription accuracy. However, their findings showed that the integration of LLMs into ASR pipelines resulted in higher WER compared to outputs without LLM correction. These results highlight significant challenges for LLMs in this context, despite their success in other natural language processing (NLP) tasks. While Zeping and Jinbo (2023) focused primarily on WER as the evaluation metric, this study expands the analysis by incorporating SeMaScore. SeMaScore considers both an error rate and semantic similarity. Before computing sentence similarity between ground truth (GT) and hypothesis (H), a sentence alignment technique is utilized to map the Levenshtein distance between GT and H. Firstly, GT and H are aligned based on the character-level edit distance and afterwards identified corresponding words or groups of words (segments). When assessing obtained mapped segments, Match error rate (MER) is leveraged to assess character-level mistakes by multiplying them with (1 - MER). This combined computation offers a more comprehensive evaluation of ASR performance, including both word-level accuracy and contextual understanding.

3 Experiments

As an initial experiment, we divided the Librispeech audio files into subsets of 100, 300, and 500 samples and processed each subset through the ASR model without using the language model for decoding. A similar approach was applied to the Common Voice dataset. Table 1 presents the WER and SeMaScore of all three subsets across both the Librispeech and Common Voice datasets. The results demonstrate consistent performance across the subsets, indicating that the quality of

the test datasets is balanced and does not introduce significant variability. Subsequent experiments were conducted with all 500 audio samples from each dataset to ensure consistency in analysis.

	Librispeech		Common Voice	
	WER	SeMaScore	WER	SeMaScore
100	45	0.75	63	0.59
300	42	0.75	65	0.58
500	41	0.76	65	0.58

Table 1: WER(%) and SeMaScore(range: 0 to 1) results for the Librispeech and Common Voice datasets. (WER: lower values are better, SeMaScores: values closer to 1 are better)

3.1 Data setup

We utilized two types of test datasets for the experiments. The first is the Librispeech ASR corpus (Panayotov et al., 2015), a read-speech dataset derived from LibriVox audiobooks. From the test-clean dataset which features high-quality audio characterized by clear pronunciation, slower speech, and minimal background noise, we randomly selected 500 audio files. The second dataset is sourced from Common Voice, a speech corpus created by users reading text derived from various public domain sources, such as blog posts, books, movies, and other speech corpora. From the test-valid dataset, we randomly extracted 500 audio samples. The term, "valid" indicates that at least two reviewers have listened to the audio, and the majority agree that the audio matches the corresponding transcription.

Both datasets required pre-processing before being fed into the ASR model. First, we converted all audio samples to the WAV format with a sampling rate of 16kHz as a standard input for speech processing systems. Additionally, we standardized the content of the transcription files across both datasets by converting all text to uppercase, removing punctuation marks (e.g. question marks, commas, and periods) and ensuring alignment between audio file names and their corresponding transcriptions.

3.2 ASR Model

Kaldi is an open-source toolkit for speech recognition that provides a hybrid ASR system, effectively combining the strengths of probabilistic sequence modeling (e.g. Hidden Markov Models) with neural networks (e.g. TDNN-F). This architecture bridges the gap between traditional statistical methods and modern neural approaches, offering flexibility and robustness. we utilized the Librispeech ASR model pre-trained on 960 hours of Librispeech data (Panayotov et al., 2015). The model comprises modular components: feature extraction, the acoustic model and the language model. This decision was motivated by the need to isolate and analyze the role of the language model as a distinct component within the pipeline. we adhered to the standard Kaldi recipe to run the ASR model with my pre-processed test datasets and corresponding transcriptions.

3.2.1 Feature extraction

The Kaldi ASR model utilizes Mel-Frequency Cepstral Coefficients (MFCC) features. The purpose of feature extraction is to represent the audio signal in a way that captures its phonetic and linguistic content while discarding irrelevant details, such as noise. Kaldi also supports Cepstral Mean and Variance Normalization (CMVN) as well as identity vectors (i-vectors) methods.

CMVN is employed to normalize the extracted features, reducing variability caused by channel or environmental conditions. Meanwhile, i-vectors are used as speaker adaptation features, capturing speaker-specific vocal characteristics. These features help mitigate variability introduced by differences in speaker traits, environments, and recording conditions. By leveraging both CMVN and i-vectors, the Kaldi ASR model is better equipped to handle pronunciation and speaker variability, ultimately improving recognition accuracy. For my experiments, we generated the features of the test datasets by running the Kaldi's standard recipes, including the MFCC extractor, CMVN, and i-vector generator.

3.2.2 Acoustic model

The acoustic model used in this study is based on Factorized Time-Delay Neural Networks (TDNN-F) (Peddinti et al., 2015). TDNNs are a specialized type of feedforward neural network that processes input data across various time delays. This design enables the model to capture temporal dependencies in sequential data between different time steps. TDNN-F is an improved version of TDNNs that reduces computational complexity during both training and inference. The input features to the TDNN-F model are aug-

mented with i-vectors. The Hidden Markov Model (HMM) is employed to represent speech as a sequence of hidden states, such as phonemes or subphonemes. The HMM establishes a probabilistic relationship between these hidden states and the observed acoustic features.

3.2.3 Decoding

During decoding, the Viterbi algorithm given in the Kaldi's recipe is employed to align the observed audio features from the acoustic model with the most likely sequence of hidden states in the HMM. The ASR decoding process integrates the outputs of the acoustic model with probabilities from both the language model and the HMM to produce the most probable word sequence. Kaldi provides a weighted finite-state transducer (WFST) (Mohri et al., 2008) decoding framework for the Librispeech ASR model. In this framework, a 3-gram language model (either pruned or full) is incorporated. The language model probabilities guide the traversal of the graph during decoding. In the Librispeech ASR model, the pruned 3-gram LM is used during decoding while a recurrent neural network language model (RNN-LM) can be employed for rescoring. The 3-gram LM provides statistical probabilities for word sequences, guiding the decoder to select plausible word combinations. After the initial decoding pass, the RNN-LM can refine the scores, leveraging its ability to capture richer linguistic patterns and improving transcription accuracy. In my experiments, we conducted decoding under two cases: with and without the pruned 3-gram LM.

3.3 Post correction

3.3.1 Language models

For post-correction of the outputs from the acoustic model, we selected a n-gram model and a transformer-based model. For the n-gram model, we used a 3-gram model specifically designed for tasks such as decoding or rescoring within the LibriSpeech ASR pipeline. This model was pretrained on text from Project Gutenberg, ensuring that the training material does not overlap with the texts found in the test and development sets. The model is trained using Modified Kneser-Ney Smoothing (Kneser and Ney, 1995), improving its ability to handle rare word sequences.

For the transformer-based model, we chose LLaMA 3.2-1B, a state-of-the-art architecture widely used in numerous NLP tasks. Transformer models like LLaMA 3.2 leverage self-attention mechanisms to capture long-range dependencies in input texts, enabling them to generate coherent and contextually accurate outputs. Given its advanced language understanding, we anticipated that this model would improve the scores in either WER or SeMaScore.

3.3.2 Process

After decoding, the post-correction process is applied using both the 3-gram and the LLaMA model for each case (with and without the LM for decoding). For post-correction with the 3-gram model, we employed the KenLM toolkit. we created a Python script to handle the post-correction process, which consists of three main steps: preprocessing, finding alternative words, and correction. Additionally, a vocabulary list was loaded, which was provided along with the LibriSpeech ASR model by OpenSLR. This file contains a lexicon derived from the training datasets of the language models built for the ASR model.

During pre-processing, any <UNK> token at the end of a line is removed if present. This token typically appears when the acoustic model recognizes a noise but treats it as an unknown word in decoding. After pre-processing each sentence, words within the 3-gram window are scored based on language model probabilities. If the score of a word is below a certain threshold (e.g. -10), alternative words are generated by replacing it with candidates from the vocabulary list. Each candidate is then scored using the language model in context, and the top 5 alternatives are returned, sorted by their language model score. If sufficient 3-gram context exists, <UNK> tokens are replaced with the highest-scoring alternative. If no better alternatives are found, the original word is retained.

For post-correction with LLaMA 3.2 1-B, we created a following simple prompt instruction:

"Correct the following transcription for any spelling or grammatical errors and output only corrected transcription: {sentence}."

This instruction prompts the model to refine the text by correcting spelling and grammar errors and give the output containing only the corrected transcription. Parameter settings limit the number of new tokens generated to 20 additional tokens beyond the original input and use greedy decoding to ensure deterministic output. Despite the prompt specifying that only the corrected transcription

		Librispeech		Common Voice	
Decoding	Post-correction	WER	SeMaScore	WER	SeMaScore
No LM	None	41	0.76	65	0.58
	3-gram model	80	0.33	89	0.34
	LLaMA 3.2-1B	40	0.74	61	0.59
LM	None	**11**	**0.91**	26	0.79
	3-gram model	51	0.53	49	0.56
	LLaMA 3.2-1B	18	0.88	31	0.78

Table 2: WER(%) and SeMaScore(range: 0 to 1) for 500 samples from Librispeech and Common Voice datasets. No LM and LM indicate the decoding process with and without the use of the Language Model respectively (WER: lower values are better, SeMaScore: values closer to 1 are better).

should be returned, some outputs included extraneous sentences such as “I’ve corrected the transcription”, “Here is the corrected sentence” or even the prompt instruction itself along with the corrected transcription. These unrelated strings are manually removed and the corrected transcription is converted to uppercase before evaluating the output.

3.4 Results

Table 2 presents the scores of 500 audio samples from Librispeech and Common Voice test datasets in WER and SeMascores.

3.4.1 Results of decoding

Before post-correction, the output of the Librispeech dataset achieved better scores than the Common Voice in both WER and SeMaScore for the two decoding cases. This was expected as a better quality of audio generally yields better performance. The acoustic model performs more effectively with clean audio samples compared to those recorded in challenging environments, such as background noise. For each test dataset, the highest performance was observed when using the LM during decoding and without post-correction: for the Librispeech, the WER and SeMaScore were 11 and 0.91 respectively while for the Common Voice, they were 26 and 0.79. Among all the results, the best scores were achieved with the Librispeech outputs decoded with the LM without post-correction. This underscores the importance of using a language model in decoding to improve the performance of the ASR system. Table 3 shows an example output of each test dataset before post-correction (BP). More examples are provided in Appendix. Most common errors stemmed from misspellings, as the acoustic model relies on phoneme probabilities and selects acoustically similar phonemes. For example, "iz" instead of "is", "upp" instead of "up". Additionally, lexical and grammatical errors were observed, reflecting the acoustic model’s inability to handle syntactical or grammatical structure. Examples include "withers spade" instead of "with a spade".

The 3-gram language model used in the decoding process helps minimize phonetic errors by reranking acoustically similar words based on context. In addition, it improves fluency and addresses grammatical issues by ensuring consistency within 3-gram structures. Even though this 3-gram language model is pruned, it showed a significant improvement in both WER and SeMaScore compared to the outputs decoded without the LM. These results highlight the importance of incorporating a language model in the decoding process to improve transcription accuracy and contextual similarity.

3.4.2 Results of post-correction

The outputs decoded without the LM showed minimal improvement after post-correction with the LLaMA model. It slightly improved the WER scores for both test datasets, with minor gains of 1 and 4 respectively. However, for outputs decoded with the LM, post-correction using either language model not only failed to improve performance but actually resulted in decreased scores. In particular, the 3-gram model significantly worsened the scores for both test datasets. For the Librispeech and the Common Voice outputs, the WER increased by an average of 40 and 20 respectively. This pattern was consistent in both decoding cases.

Table 3 represents example transcriptions of each test dataset after post-correction (AP). Examining hypotheses decoded without the LM, the post-correction with the 3-gram model frequently

Librispeech		
	Reference	“It is sixteen years since John Bergson died”
No LM	BP	“It iz sixtine years since Jon Burkes and dyed”
	AP(3-gram)	“It iz te chapel ago then said broughams me”
	AP(LLaMA)	“It is sixteen years since John Burkes died”
LM	BP	“It is sixteen years since John <UNK> and died”
	AP(3-gram)	“It is sixteen years since then Mitchel and died”
	AP(LLaMA)	“It is sixteen years since John died”
Common Voice		
	Reference	“Henderson stood up with a spade in his hand”
No LM	BP	“Hendersons stood upp withers spade in hiz hand”
	AP(3-gram)	“Hendersons stood up ah away work hand z”
	AP(LLaMA)	“Hendersons stood up with their spade in his hand”
LM	BP	“Henderson stood up with spade in his hand”
	AP(3-gram)	“Henderson stood up with him work hand hand”
	AP(LLaMA)	**“Henderson stood up with a spade in his hand”**

Table 3: Hypotheses of Librispeech and Common Voice datasets in two decoding and post-correction cases. (No LM: Hypotheses decoded without the pruned 3-gram model, LM: Hypotheses decoded with the pruned 3-gram model, BP: Hypothesis Before post-correction, AP: Hypothesis After post-correction)

produced unrelated or semantically nonsensical phrases, such as "It iz te chapel ago then said broughams me" and "Hendersons stood up ah away work hand z". It struggles especially when initial hypotheses contain serious errors, often amplifying local context errors instead of resolving them. In contrast, the LLaMA model performed better corrections, often mitigating major errors. For example, it corrected "Jon" to "John" and "withers spade" to "with their spade" However, some minor inaccuracies remained, such as substituting "Bergson" with "Burkes" and "a" with "their". For hypotheses generated with the LM during decoding, both models performed better, as the initial inputs contained fewer errors. The post-correction outputs were closer to the reference, with fewer severe inaccuracies. For instance, LLaMA successfully refined "with spade" to "with a spade" in the Common Voice dataset, demonstrating its ability to improve grammatical error. However, both models struggled with certain challenges, such as handling unknown tokens. The 3-gram model replaced <UNK> with “Mitchel” while the LLaMA model left it unchanged.

3.4.3 Relation between WER and SeMaScore

Generally, a lower WER corresponds to a higher SeMaScore, indicating that improved word-level accuracy (lower WER) improves the semantic accuracy of the output. In my experiments, changes in SeMaScore were often aligned with changes in WER. In other words, when the SeMaScore improved or declined, the WER tended to follow suit. However, this relationship is not strictly linear, as SeMaScore accounts for more than just word accuracy. It was particularly interesting to observe how the SeMaScore varied across different experimental cases and its alignment with the WER in the ASR performance evaluation.

4 Ethical considerations

Language diversity The use of an English test dataset in this experiment raises questions about language diversity. Research in NLP often emphasizes Indo-European languages, neglecting smaller or underrepresented language groups, such as those in Asia and Africa. This imbalance causes a lack of labeled data for these languages and influences the focus of researchers on more resource-rich languages, reinforcing a phenomenon called topic overexposure (Hovy and Spruit, 2016). Topic overexposure can lead to bias by building availability heuristics (Tversky and Kahneman, 1973) where researchers tend to prioritize topics with readily accessible data. To address this, future research should aim to incorporate datasets from diverse and underexposed languages.

Variations inside languages: Domain, dialect, ability Future research should also consider the linguistic and demographic variations present

within a single language, such as differences in domain, dialect, and speaker accents. ASR systems, including both LMs and acoustic models, often exhibit bias when trained exclusively on clean data from standardized or native speakers. This overfitting to majority-centered data can result in poorer performance, reduced accuracy, and demographic biases when working with non-standard dialects or regional accents. To improve ASR systems and enhance their fairness, it is vital to diversify the training and testing datasets across different demographic and domain groups.

Environmental perspective From an environmental point of view, traditional ASR systems generally have lower computational costs than E2E systems due to their modular architectures. E2E systems require larger datasets, more powerful hardware, and extended training periods, which contribute to higher energy consumption and environmental impact. Future research should explore strategies to enhance the efficiency of traditional ASR systems to bridge the performance gap with E2E systems while maintaining their environmental benefits. For instance, developing more energy-efficient algorithms, optimizing feature extraction methods, and refining language and acoustic modeling could reduce computational costs.

5 Conclusion

This research investigated the role of LMs in traditional ASR pipelines, specifically their impact during post-correction. The experiments demonstrated that incorporating the LM during the decoding stage improved significantly transcription accuracy, as evidenced by lower WER and higher SeMaScore. While the 3-gram model (pruned) improve transcription quality in decoding, both the 3-gram and LLaMA models exhibit notable limitations in post-correction tasks. They struggle to consistently correct errors when the initial hypotheses from the acoustic model were flawed. In some cases, they even exacerbate those errors. The 3-gram model, in particular, frequently introduced unrelated or nonsensical phrases during post-correction, leading to substantial declines in WER and SeMaScore. Although the LLaMA model showed marginal improvements, it proved insufficient to meaningfully enhance transcription quality overall. While LLaMA models with higher parameters perform possibly better than the one used in these experiments, prioritizing optimization of the language model for the decoding process would likely yield greater benefits. This approach not only offers the potential to improve speed but also enhances the practicality of implementation in ASR systems.

These results underscore the importance of optimizing the use of LMs during the decoding stage to minimize transcription errors. Furthermore, relying solely on post-correction is not only insufficient to fully address these errors but may also amplify them. These findings highlight the inevitability of error propagation in ASR pipelines and the critical need to address the limitations of LMs in this context. This study can provide a valuable foundation for future research aimed at improving various ASR systems, including E2E approaches, by refining the role and application of language models. Especially during the decoding process to mitigate error propagation throughout the pipeline. Furthermore, the performance of alternative transformer models can be explored and compared to assess whether the observed limitations are inherent to specific language models or if improvements can be achieved through other model selection or optimization techniques.

References

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28.

Amandeep Singh Dhanjal and Williamjeet Singh. 2024. A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications*, 83(8):23367–23412.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2008. Speech recognition with weighted finite-state transducers. *Springer Handbook of Speech Processing*, pages 559–584.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech*, pages 3214–3218.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Luká Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlícek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The kaldi speech recognition toolkit. IEEE Signal Processing Society.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Zitha Sasindran, Harsha Yelchuri, and TV Prabhakar. 2024. Semascore: a new evaluation metric for automatic speech recognition tasks. *arXiv preprint arXiv:2401.07506*.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. *arXiv preprint arXiv:2102.01547*.

Min Zeping and Wang Jinbo. 2023. Exploring the integration of large language models into automatic speech recognition systems: An empirical study. In *International Conference on Neural Information Processing*, pages 69–84. Springer.

A Appendix

This section contains an additional table to support the analysis presented in Section 3.4.

Librispeech		
	Reference 1	"A cold lucid indifference reigned in his soul"
No LM	BP	"A colde lucid indifference arraigned in hiz soul"
	AP(3-gram)	"A colde grave intervals vanished himself court z"
	AP(LLaMA)	"A cold lucid indifference arrested in his soul"
LM	BP	"A cold lucid indifference reined in his soul"
	AP(3-gram)	"A cold sweat intervals vanished in sharply soul"
	AP(LLaMA)	"A cold lucid indifference reined in his soul"
	Reference 2	"Missus griffin however expressed the need for a little more light"
No LM	BP	"Missus gryphon however express the need phor a littlemore lyte"
	AP(3-gram)	"Missus gryphon today small it same of ho man inclusive"
	AP(LLaMA)	"Missus gryphon has however express the need phor a littlemore lyte"
LM	BP	"Missus griffith however expressed no need for a little more light"
	AP(3-gram)	"Missus griffith jenkins small it surprise for you little more light"
	AP(LLaMA)	"Missus griffith however expressed no need for a little more light"
Common Voice		
	Reference 1	"Without the dataset the article is useless"
No LM	BP	"Without that that' set the arted gorse useless"
	AP(3-gram)	"Without that knowledge hark in table Corbett bushes"
	AP(LLaMA)	"Without that the arted gorse useless"
LM	BP	"Without that upset the articles useless"
	AP(3-gram)	"Without that knowledge me boat mentioned"
	AP(LLaMA)	"Without that upset the articles, useless"
	Reference 2	"You're playing with dynamite"
No LM	BP	"Here playing with dynamite <UNK>"
	AP(3-gram)	"Here playing cards dynamite"
	AP(LLaMA)	"Here playing with dynamite"
LM	BP	**"You're playing with dynamite"**
	AP(3-gram)	**"You're playing with dynamite"**
	AP(LLaMA)	**"You're playing with dynamite"**

Table 4: Hypotheses of Librispeech and Common Voice datasets in two decoding and post-correction cases. (No LM: Hypotheses decoded without the pruned 3-gram model, LM: Hypotheses decoded with the pruned 3-gram model, BP: Hypothesis Before post-correction, AP: Hypothesis After post-correction)