POST-ASR CORRECTION USING LARGE LANGUAGE MODELS

BLIND

ABSTRACT

Automatic Speech Recognition (ASR) systems have become increasingly integral in various applications, yet they still struggle with specific types of errors that can degrade transcription accuracy. This paper addresses the challenge of improving ASR outputs by leveraging Large Language Models (LLMs) for post-ASR correction. The LLMs that were used in this paper are LLaMa-2-7B, Gemma-7B, and Mistral-7B. Our approach focuses on two types of transcription: literal, which captures speech exactly as spoken, and clean, which ensures grammatical correctness and semantic coherence. This paper uses few-shot prompting to mainly identify and combine the most repetitive sequences of sentences across all given logical hypotheses from ASR outputs. Moreover, we use word error rate (WER) to assess the results of our models. Our findings indicate that targeted post-ASR correction enhances transcription accuracy, offering valuable insights for the development of more robust ASR systems.

***Index Terms*—** Automatic Speech Recognition, Large Language Models, Word Error Rate, Prompt Engineering, Ensemble

1. INTRODUCTION

Automatic Speech Recognition (ASR) technology has undergone significant advancements, becoming indispensable in numerous applications, ranging from virtual assistants and transcription services to accessibility tools and customer service automation. However, despite these technological advances, ASR systems continue to face considerable challenges, particularly when transcribing speech characterized by accents, varying speech rates, background noise, and other complicating factors. These challenges often lead to errors at various levels, lexical, semantic, and syntactical, thereby affecting the overall quality and usability of transcriptions.

The accuracy of ASR systems is crucial for ensuring effective human-computer interactions and enhancing user experiences across different domains. Errors in ASR output not only disrupt communication but also create significant obstacles to accessibility, particularly for users relying on speech-based technologies for essential tasks. Therefore, improving ASR accuracy is not just a technical challenge but a vital step towards making technology more inclusive and user-friendly.

Integrating speech as a primary interaction modality in intelligent systems significantly enhances accessibility and user experience, making technology more intuitive and user-friendly. However, transitioning from text-based interfaces to voice-enabled systems presents challenges, primarily due to the error susceptibility of ASR technologies. ASR systems often produce errors such as grammatical mistakes, word misspellings, and nonsensical outputs, complicating users' ability to detect and correct spoken input errors.

This paper addresses these issues through post-ASR correction using state-of-the-art Large Language Models (LLMs). Unlike traditional ASR models, which may not fully capture the nuances of spoken language, LLMs offer sophisticated capabilities for understanding and generating natural language, making them ideal candidates for refining ASR outputs. Our research focuses on two main types of transcription:

- **Literal Transcription**: Ensures the transcription reflects exactly what was said, including any grammatical inaccuracies [1].
- **Clean Transcription**: Aims for grammatically correct and semantically coherent transcriptions, even if this involves slight deviations from the exact speech [2].

We propose a multifaceted approach that leverages n-best hypotheses generated by ASR systems. By applying prompt-tuning techniques and ensemble approaches, we correct and refine these hypotheses to improve overall transcription accuracy. Our contributions include a detailed analysis of ASR output errors and their impact on transcription quality, as well as the development of prompt-tuning strategies for LLMs to correct these errors.

2. RELATED WORK

Large Language Models (LLMs) are specifically designed to process, understand, and generate human-like text [3]. While they excel at generating high-quality text, they require adaptation to effectively handle errors in Automatic Speech Recognition (ASR). The challenge is to develop robust methodologies that enhance LLMs' capabilities in correcting these errors, ensuring accurate and reliable speech interactions.

The literature offers numerous examples of LLMs being employed to address ASR transcription errors through methods such as distillation and re-scoring [4, 5, 6]. For instance, Futami et al. [4] utilize BERT to generate soft labels for training ASR models, while Kubo et al. [6] focus on transferring semantic knowledge embedded within vector representations. These methods underscore the potential of LLMs to enhance ASR performance by leveraging advanced natural language processing (NLP) techniques. Another innovative approach employs an LLM with prompt tuning for Contextual Spelling Correction (CSC) in ASR systems [7]. Specifically, the Flan-PaLM model[8] is adapted for specific tasks by appending task-specific embeddings to the input text, allowing efficient adaptation without modifying the model's weights.

With the rapid advancements in LLMs, there is a significant potential to enhance ASR performance through the advanced NLP capabilities of these models. Despite the challenges posed by the larger number of parameters in newer LLMs, which can complicate traditional distillation and rescoring methods, these models offer a vital capability: in-context learning, which opens up novel applications. Recent studies demonstrate the effectiveness of meta-soft prompts, which require minimal input data to achieve superior results with pre-trained models by training only a few parameters [9]. Our approach leverages this capability by using examples in the prompt to facilitate few-shot learning with LLMs. Furthermore, while many existing methods focus on specific types of errors, such as spelling corrections, they often do not address the broader spectrum of transcription inaccuracies that can occur in ASR systems. Our methodology aims to fill this gap by employing advanced LLMs for post-ASR correction, ensuring a comprehensive approach to improving ASR transcription accuracy.

Filling these gaps is crucial for enhancing the practicality and reliability of ASR systems. By developing methods that require fewer resources and can handle a wider range of transcription errors, we can make ASR technology more accessible and effective across different domains. Our research builds upon these foundations by leveraging state-of-the-art LLMs for post-ASR correction, specifically targeting errors generated during transcription (See §3). Unlike previous studies, our approach utilizes the n-best hypotheses generated by ASR systems and applies prompt-tuning techniques and ensemble methods to correct and refine these hypotheses. This method not only addresses a broader range of errors but also requires fewer training examples, making it more practical for real-world applications. Through this research, we aim to bridge the gap between ASR technology and its practical application, enhancing the accuracy and reliability of speech transcriptions across various domains.

3. METHODOLOGY

Given the significant challenges in improving ASR transcription accuracy and reliability, particularly in dynamic and resource-constrained environments, it is essential to develop robust post-ASR correction methods. Building on the insights from previous research, we leverage advanced LLMs combined with prompt-tuning techniques and ensemble methods to address these challenges effectively.

3.1. Dataset and Model Selection

This paper uses HyPoradise [10] as the dataset for training and testing the LLMs. The authors of the dataset implemented a decoding strategy on several popular ASR datasets to generate paired data containing a 5-best hypotheses list and a single ground-truth transcription. Hyporadise is a compilation of various datasets, including LibriSpeech[11], CHiME-4[12], WSJ[13], SwitchBoard[14], CommonVoice[15], Tedlium-3[16], LRS2[17], ATIS[18], and CORAAL[19]. Each of these datasets has distinct characteristics, as detailed in their respective publications. They differ in file size, number of instances, sentence length, background noise, and most importantly, the average Word Error Rate (WER) across instances. Additionally, the output sentences, or ground truth, from each dataset exhibit differences in factors such as grammatical errors. Figure 1 shows sample outputs from two files:

- "i want a flight from memphis to seattle that arrives no later than three p m"
- "and and so i w i was looking on the front porch i saw debris and trash can flowing"

Fig. 1: Outputs from Chime4 and CORAAL

Moreover, the WER for each dataset individually can be very different, from 3.6% to 30.6%. Therefore, it is preferred to use every dataset in the HyPoradise independently.

One of the models selected for this paper is LLaMa-2-7B, which has state-of-the-art results [10]. Two other prominent open-source models for our study are Mistral-7B and Gemma-7B [20, 21]. These models are chosen due to their diversity in architecture, their availability, and favorable reviews in the existing literature. Specifically by utilizing diverse pre-trained models, the expectation is to yield better ensemble results. The models are prompted to either find the best transcription input or generate their own if none of the transcriptions made sense.

3.2. Error Categorization and Causes

Errors in ASR systems can be broadly categorized into three layers:

1. **Environmental Errors (Layer 1)**: These are caused by external factors such as background noise, speaker accent, and speech rate, which the ASR system cannot control.
2. **Model and External Interaction Errors (Layer 2)**: These result from a combination of ASR model inefficiencies and external factors, such as slight mispronunciations or overlapping speech.

3. **Internal Model Errors (Layer 3)**: These are intrinsic errors generated by the ASR system during transcription, such as misrecognition of words or phrases.

These error layers impact each other. For example, background noise (Layer 1) can lead to misheard words and context misunderstandings (Layer 2), which in turn can cause misspellings and other transcription errors (Layer 3). Figure 2 depicts an illustration of various errors in each layer. The focus of our work in this paper addresses the errors in Layer 3, which are extensively discussed in literature [22, 23, 24].

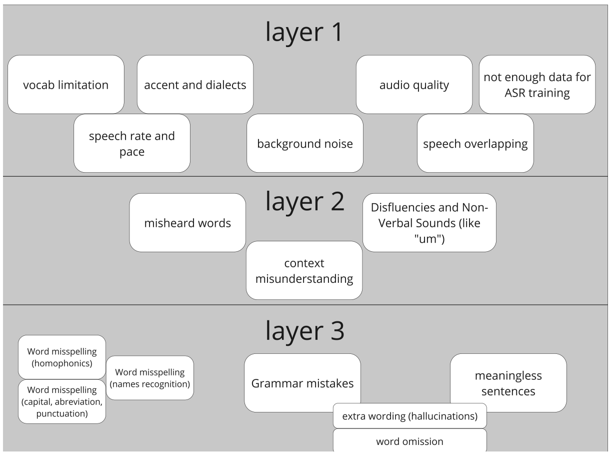
layer 1
vocab limitation
accent and dialects
audio quality
not enough data for ASR training
speech rate and pace
background noise
speech overlapping
layer 2
misheard words
Disfluencies and Non-Verbal Sounds (like "um")
context misunderstanding
layer 3
Word misspelling (homophonics)
Word misspelling (names recognition)
Word misspelling (capital, abreviation, punctuation)
Grammar mistakes
meaningless sentences
extra wording (hallucinations)
word omission

Fig. 2: Causes of ASR Errors Categorized into three layers

3.3. In-Context Learning

Few-shot prompting is preferred for post-ASR correction using LLMs over other techniques since it requires little to no additional labeled data, saving time and computational resources. Prompting is straightforward, as it involves no modifications to the model's architecture, making it user-friendly and accessible even to those without specialized knowledge or powerful hardware. LLMs, with their vast pre-trained knowledge, can generalize well from a few examples, handling a wide range of tasks effectively. Prompting also avoids overfitting and allows for rapid experimentation and immediate feedback, fostering broader adoption and flexibility in usage, especially for the wide range of data, as explained in §3.1.

3.4. Proposed Approaches

Transcription Method. Based on our dataset analysis, individual words in the output are found with a very high probability—approximately 97% within one of the input hypotheses. This indicates that the transcription type employed will primarily be literal, verbatim transcription that accurately captures and documents every spoken word, including pauses, fillers, and non-verbal sounds, in a recorded conversation or speech. However, in certain instances, such as the example where a single word, most probably a preposition, is missing from all inputs, minor cleaning for grammatical accuracy is required. This necessity for clean transcription to correct grammatical errors, though infrequent, is observed in most files. Therefore, we will predominantly use verbatim transcription with occasional clean transcription for minor grammatical adjustments.

- i am not sure what you have on your list
- i am not sure what you have on your list
- i am not sure what you have on your list
- i am not sure what you have in your list
- i am not sure what you have on the list

- **Output:** i am not sure what you have on your **own** list.

Fig. 3: Words not found in any hypothesis

As demonstrated in Figure 3, the word "own" is found in the output sentence only, so using a minor cleaning for the sentence would modify the generated sentence to match the output. To generate these outputs, an optimal prompt must be created for each LLM.

Most Frequent Words (MFWs). We attempt to generate an output based on the most frequently occurring words in the input hypotheses. Nonetheless, we observe a high WER. Upon reviewing the generated sentences, we identify a minor issue where the LLM occasionally produces grammatically incorrect sentences, even when the inputs are grammatically sound. For example, as illustrated in Figure 4, the differing number of words due to the presence of the preposition "in" leads the LLM to compare dissimilar words, resulting in grammatically incorrect sentences and the repetition of certain words. Based on the error discussed in Figure 4, it becomes evident that phrases like "pay in" should be treated as cohesive units and included intact in the generated output. This realization underscores the importance of considering the most repetitive sequences of words rather than focusing solely on individual words when addressing such errors.

- this opposition did not pay in her
- the supposition did not pay in her
- this opposition did not pay her
- this opposition did not pain her
- the supposition did not pain her

- **Generated:** the opposition did not pay her her.

Fig. 4: Example using the MFWs approach

Sequence of Words (SoW). Leveraging the SoW in an LLM involves synthesizing an output sentence from a set of five similar input sentences by identifying and prioritizing the most repetitive sequence among them. This method ensures that the generated output maintains coherence and fidelity to the common phrases and structures found across the inputs. By focusing on the repetitive sequences rather than individual words in isolation, the LLM can effectively capture the context and intent embedded within the collective input sentences. This approach enhances the model's ability to produce outputs that are contextually appropriate and linguistically coherent, thus improving its overall performance in tasks requiring synthesis and correction of textual data.

Priorities for SoW. At times, we encounter sequences with nearly equal occurrences. Figures 5 and 6 illustrate an edge case. In such cases, our priority is to select the sequence of words that makes the most contextual sense, even if it is not the most frequently repeated among the input hypotheses. Additionally, when multiple sensible sequences are identified, preference is given to the sequence associated with hypotheses having higher am_scores. These hypotheses are provided by ASR in descending order of confidence, with the top-ranked hypotheses being prioritized by the LLM during output generation.

- the program **compiles** correctly
- the program combines correctly
- the program combiles correctly
- the program comes correctly
- the program combine correctly

- **Output:** the program **compiles** correctly

Fig. 5: Edge-case Example for MFWs approach

- the fleshy **food** is edible
- the fleshy **food** is edible
- the fleshy fruit is edible
- the fleshy fruit is edible
- the fleshy fruit is edible

- **Output:** the fleshy **food** is edible

Fig. 6: Edge-case Example for MFWs approach

Final Prompt. Based on the comprehensive analysis presented in the preceding subsections of the approach section, an optimal prompt for is formulated, as shown in Figure 7.

> "Given five hypotheses, synthesize a unified hypothesis by identifying and integrating the most frequently recurring sequences of sentences across all hypotheses, prioritizing sequences that align cohesively with the broader contextual meaning of the sentences. Additionally, assign slightly higher importance to sentences appearing at the beginning of the hypotheses. In cases where no common sequences are found among the hypotheses, select the hypothesis that exhibits logical coherence."

Fig. 7: Prompt Example

3.5. Process Procedure

The results of each model are extracted and taken for the future ensemble stage, which takes the best output out of the three models; furthermore, these are compared to the original 'output' column taken from the dataset, which is the ground truth, for further analysis. Some examples are shown below in which all three models had the same result.

- **Input:**
  - about half these managers are in the us
  - about half these managers are in the us
  - about half these managers are in the us
  - about half these managers are in the us
  - about half of these managers are in the us

  Output: about half these managers are in the u s

  Prediction: about half of these managers are in the us

- **Input:**
  - union officials expect ratification
  - union officials expect ratification
  - union officials expect ratification
  - union officials expect ratification
  - union officials expect ratification

  Output: union officials expect ratification

  Prediction: union officials expect ratification

3.6. Ensemble

To perform ensembling, we use a variant of decision fusion. As the name implies, decision fusion combines the decisions made by multiple classifiers to achieve a consensus that surpasses the accuracy of individual classifier decisions. The primary objective of decision fusion is to enhance the overall performance of the classification task. In our case, the models are ensembled to create the most agreed upon output. This is done by prompting each model, obtaining the outputs, and performing a weighted majority voting to obtain a cumulative outcome, similar to Decision Fusion[25]. The models are first evaluated individually to find which model is best on its own. The ensemble can then consistently be performed: When two models agree on an output that output will be chosen; otherwise the model that does best on its own is chosen.

4. EXPERIMENTAL RESULTS

4.1. Evaluation Metric

The existing literature shows that Word Error Rate (WER) is a common metric to evaluate our models. The output of the models is compared to the ground truth, and the difference is the WER in percentage (%) [10, 24]. To measure the WER of the baseline, we calculate the WER between the hypothesis with the highest AM score (which is always the first hypothesis) and the ground truth of the input. We also add the Oracle n-best (O_nb) WER to the comparison for better evaluation.

The observation from reviewing multiple datasets reveals that abbreviations consistently have character space-separation, leading to higher Word Error Rates (WER). To ensure equitable comparisons of LLM effects on output quality, we enforce uniform formatting rules across generated sentences: no space-separation. The models are also evaluated against a baseline with space separation and a baseline without space separation.

4.2. Numerical Results

In this paper, multiple attempts to use zero-shot learning with various language models failed to generate the desired outputs [26], resulting in a Word Error Rate (WER) of 100% for most sentences. This performance is due to the absence of examples to guide the model in the expected output format. Table 1 presents a comparison of the few-shot prompting performance of various Large Language Models (LLMs) against the baseline.

According to Table 1, we find that Mistral is the best standalone model. It consistently performs well and is better than the baseline in most cases. Further analysis of the cases where the baseline outperforms Mistral, we found that the datasets contain grammatically incorrect output sentences. An example of one of the grammatically incorrect sentences found in CORAAL, SWBD, and TD3 is shown in Figure 8.

> “and and so i w i was looking on the front porch i saw debris and trash can flowing”

Fig. 8: Wrong Grammar Output Example

LLaMA and Gemma, on the otherhand, do not perform well in most of the datasets. We attribute the weak performance to multiple reasons: They are not tuned for the task; they will sometimes attempt to answer the question rather than to transcribe, or swap English character representations of numbers into their algebraic equivalent.

The ensemble proved best results across all models. In particular, despite LLAMA and Gemma sometimes performing poorly on their own, combining them with Mistral outperformed all three models and in many cases the baseline.

5. CONCLUSION

In conclusion, our study showcases the effectiveness of advanced LLMs and ensemble decision fusion in enhancing ASR transcription quality. Mistral positions itself as a robust model, consistently improving transcription accuracy over baseline methods across various datasets. While LLaMa and Gemma demonstrate mixed performance individually, their integration in ensemble configurations proves beneficial, highlighting the complementary strengths of diverse LLM architectures. All models fix grammatical errors and stuttering, making them suitable for tasks where a transcription is unclean and needs a complete change. Future research should explore refining prompt-tuning techniques tailored to specific ASR challenges and expanding dataset diversity to strengthen generalization capabilities. Furthermore, integrating real-time feedback mechanisms and adapting LLMs to dynamic environments can further advance ASR technology’s reliability and applicability in practical settings.

Test Set	BL	BL_a	Mist	Gem	Lam	Ens	O_nb
atis	8.8	7.5	**5.4**	13.7	36.8	5.5	5.2
chime4	11.7	10.7	9.35	13.6	18.2	**9.1**	8.6
coraal	24.5	24.6	29.4	35.6	50.2	**28.9**	22.1
cv	17.5	17.5	**15.6**	21.8	24.6	**15.6**	11.7
lrs2	15.3	15.3	15.1	56.5	69.1	**14.8**	6.9
ls_clean	8.7	8.9	3.6	13.6	18.1	**3.3**	0.9
ls_other	10.8	11.0	6.2	17.2	23.6	**6**	2.8
swbd	17.9	17.8	24.7	33.8	55.2	**24.6**	12.7
td3	4.9	4.9	12	23.8	44.1	**11.3**	2.9
wsj	6.3	5.6	4.9	8.28	14.3	**4.7**	3.9

Table 1: Few-Shot Results of LLM models. The acronyms represent the following models: BL for BaseLine and BL_a for baseline with no space-separation, Mist for Mistral, Gem for Gemma, Lam for LLaMA, and finally Ens for ensemble. The performance metric used for comparison is WER in percentage (%). Finally, O_nb stands for Oracle n-best.

6. REFERENCES

[1] Laurens van der Werff and Willemijn Heeren, “Evaluating asr output for information retrieval,” *Searching Spontaneous Conversational Speech*, p. 13, 2007.

[2] Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, “An end-to-end model from speech to clean transcript for parliamentary meetings,” in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2021, pp. 465–470.

[3] Manit Mishra, Abderrahman Braham, Charles Marsom, Bryan Chung, Gavin Griffin, Dakshesh Sidnerlikar, Chatanya Sarin, and Arjun Rajaram, “Dataagent: Evaluating large language models’ ability to answer zero-shot, natural language queries,” in *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, 2024, pp. 1–5.

[4] Hayato Futami, Hirofumi Inaguma, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, “Distilling the knowledge of bert for sequence-to-sequence asr,” *arXiv:2008.03822*, 2020.

[5] Takuma Udagawa, Masayuki Suzuki, Gakuto Kurata, Nobuyasu Itoh, and George Saon, “Effect and analysis

of large-scale language model rescoring on competitive asr systems," *arXiv:2204.00212*, 2022.

[6] Yotaro Kubo, Shigeki Karita, and Michiel Bacchiani, "Knowledge transfer from large-scale pretrained language models to end-to-end speech recognizers," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8512–8516.

[7] Gan Song, Zelin Wu, Golan Pundak, Angad Chandorkar, Kandarp Joshi, Xavier Velez, Diamantino Caseiro, Ben Haynor, Weiran Wang, Nikhil Siddhartha, Pat Rondon, and Khe Sim, "Contextual spelling correction with large language models," 12 2023, pp. 1–8.

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yuxuan Li, Xuezhi Wang, Mostafa Dehghani, et al., "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[9] Jen-Tzung Chien, Ming-Yen Chen, and Jing-Hao Xue, "Learning meta soft prompt for few-shot language models," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 57–62.

[10] Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng, "Hyporadise: An open baseline for generative speech recognition with large language models," *Advances in Neural Information Processing Systems*, 2024.

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

[12] Emmanuel Vincent, Shinji Watanabe, Jon Barker, and Ricard Marxer, "The 4th chime speech separation and recognition challenge," 2016.

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

[14] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, speech, and signal processing, ieee international conference on*. IEEE Computer Society, 1992, vol. 1, pp. 517–520.

[15] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *arXiv:1912.06670*, 2019.

[16] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve, "Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer: 20th International Conference*. Springer, 2018.

[17] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6447–6456.

[18] Charles T Hemphill, John J Godfrey, and George R Doddington, "The atis spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, PA*, 1990.

[19] Charlie Farrington and Tyler Kendall, "The corpus of regional african american language," 2021.

[20] Ayush Thakur and Raghav Gupta, "Introducing super rags in mistral 8x7b-v1," *arXiv:2404.08940*, 2024.

[21] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, et al., "Gemma: Open models based on gemini research and technology," *arXiv:2403.08295*, 2024.

[22] Yuanchao Li, Pinzhen Chen, Peter Bell, and Catherine Lai, "Crossmodal asr error correction with discrete speech units," *arXiv:2405.16677*, 2024.

[23] Junwei Liao, Sefik Eskimez, Liyang Lu, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng, "Improving readability for automatic speech recognition transcription," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 5, pp. 1–23, 2023.

[24] Martucci Giuseppe, Mauro Cettolo, Matteo Negri, and Marco Turchi, "Lexical modeling of asr errors for robust speech translation," in *Proceedings of Interspeech 2021*, pp. 2282–2286. 2021.

[25] Atsunori Ogawa, Naohiro Tawara, Marc Delcroix, and Shoko Araki, "Lattice rescoring based on large ensemble of complementary neural language models," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6517–6521.

[26] Yinheng Li, "A practical survey on zero-shot prompt design for in-context learning," in *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for NLP*. 2023, RANLP, INCOMA Ltd., Shoumen, BULGARIA.