From Stutter to Clarity: A Whispering LLaMA Approach for Stuttering

Jirayu Burapacheep, Farzaan Kaiyom, Valerie Fanelle
Stanford University
{jirayu, farzaank, vfanelle}@stanford.edu

Abstract

Recent advancements in automatic speech recognition (ASR) have struggled with recognizing stuttered speech. This project benchmarks state-of-the-art ASR models on the LibriStutter dataset and explores improving recognition using two-stage systems combining acoustic and language models. Off-the-shelf models like Whisper, GPT-4o, and LLaMA 2 are evaluated, and the Whispering LLaMA framework is fine-tuned on stuttered speech. Results show that language models improve performance, especially for smaller acoustic models, and fine-tuning further reduces the word error rate. An ablation study reveals the impact of the number of hypotheses passed to the language model. This work highlights the potential of two-stage ASR systems for enhancing stuttered speech recognition.

1 Introduction

Recent years have seen remarkable developments in pretrained end-to-end automatic speech recognition (ASR) models (Karita et al., 2019; Baevski et al., 2020; Gulati et al., 2020; Radford et al., 2022). These advancements have been driven by the increasing availability of large-scale datasets, improvements in deep learning architectures, and the development of sophisticated training techniques. End-to-end ASR models, which streamline the traditional ASR pipeline by directly mapping input audio to text, have shown significant promise in various applications, including voice assistants and transcription services.

However, despite these advancements, these systems continue to struggle with unusual speech patterns, including stuttering, due to training primarily on fluent speakers' data (Mitra et al., 2021). For example, Wav2Vec 2.0 (Baevski et al., 2020) transcribes the speech "t-t-t-to those people" as "TTETETETUTO THOSE PEOPLE". Stuttering is a speech disorder that impacts people's ability to communicate effectively. Stuttering can manifest in various forms: prolongations, repetitions, blocks, interjections, and broken words (See Table 1). Therefore, this limitation underscores the need for more inclusive ASR systems that can accurately recognize and transcribe stuttered speech, ensuring equal access and usability for individuals who stutter.

Disfluency	Examples
Blocks	I was **[pause]** sleeping
Prolongations	I was **ssssleeping**
Interjection	I **uh** was **uhm uh** sleeping
Sound Repetition	I **w-w-w-was** sleeping
Word Repetition	I **was was** sleeping

Table 1: **Types of stuttering disfluencies and their examples.** Blocks indicate involuntary pauses within a phrase. Prolongations indicate prolonged sounds. Interjection indicates the addition of fabricated words or sounds. Sound repetition refers to a repetition of phoneme, and word repetition refers to a repetition of any word.

Recent developments introduce two-stage systems that leverage language models to improve the performance of acoustic models. By incorporating both audio and language understanding, these systems have significantly enhanced speech recognition capabilities. Furthermore, the two-stage paradigm can accommodate domain shifts like low-resource languages or disfluent speech since only the language model is required to be fine-tuned on the new dataset (Li et al., 2022; Liu et al., 2021; Yu et al., 2023). This modularity is particularly motivating for the stuttered speech recognition task, as language models can effectively capture the unique patterns and disfluencies present in stuttered speech and correct them.

In this project, we aim to (1) establish a benchmark for ASR models on stuttered speech and (2)

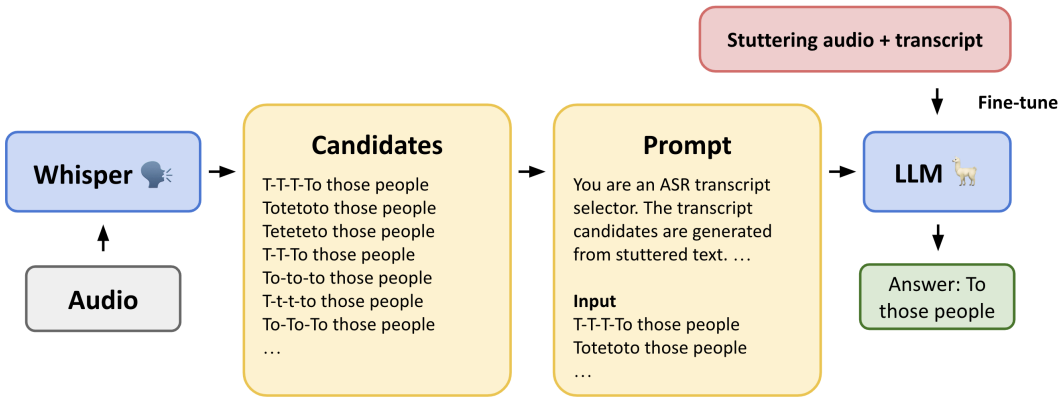

Figure 1: Overview of the two-stage ASR framework for stuttered speech recognition, inspired by Whispering LLaMA framework. The acoustic model generates multiple transcription hypotheses, which are then passed to a large language model to produce a refined final transcription.

explore different ways to incorporate two-stage speech recognition systems, both off-the-shelf and fine-tuned on stuttered speech data, to improve recognition performance on this task.

2 Related Works

Automatic Speech Recognition. Recent advancements in ASR, driven by deep neural networks, have led to substantial performance improvements. Notable architectures include the Conformer (Gulati et al., 2020), wav2vec 2.0 (Baevski et al., 2020), and Whisper (Radford et al., 2022). Many two-stage systems have been proposed to include language models to rescore the transcriptions (Xia et al., 2017; Guo et al., 2019; Hu et al., 2021; Yang et al., 2021; Salazar et al., 2020). The Whispering LLaMA framework (Radhakrishnan et al., 2023a) integrates Whisper's acoustic model and the Alpaca language model (Taori et al., 2023) to generate enhanced transcriptions. Specifically, Whisper produces multiple candidate transcriptions (hypotheses) and acoustic embeddings from the audio input, which are then conditioned by the Alpaca model to yield the final transcription, circumventing traditional ranking-based rescoring.

Disfluent Speech Recognition. Previous efforts to enhance ASR performance on disfluent speech have focused on classifying disfluency events. These systems, trained on stuttering identification datasets including SEP-28K (Lea et al., 2021), KSoF (Bayerl et al., 2022), and LibriStutter (Kourkounakis, 2021), adjust, ignore, or remove classified disfluency (Zayats et al., 2016) or provide their locations to the ASR for better handling (Shonibare et al., 2022). Adapting ASRs to disordered, dysarthric, and accented speech has also been studied (Tobin and Tomanek, 2021; Shor et al., 2019). Some progress has been made in directly enhancing ASRs for stuttering, such as fine-tuning models on stuttered speech data (Lea et al., 2023) and tuning acoustic model weights and insertion penalties (Mitra et al., 2021).

3 Methodology

In this section, we describe our approach to benchmarking two-stage ASR systems on stuttered speech. We focus on evaluating both off-the-shelf models and fine-tuned systems to determine their effectiveness in recognizing stuttered speech patterns. We analyze the impact of varying the number of transcription hypotheses on the overall transcription quality, using word error rate to evaluate performance. See Figure 1 for a visualization of the two-stage framework that we are evaluating.

Off-the-shelf models. For the off-the-shelf evaluation, we utilize state-of-the-art ASR models, specifically multiple sizes of the Whisper model (Radford et al., 2022). These models generate multiple transcription hypotheses for each stuttered speech sample, which are then passed to large language models (LLMs) such as GPT-4o (OpenAI, 2024) and LLaMA 2 (Touvron et al., 2023). The LLMs integrate the linguistic information to produce a final, refined transcription. We also evaluate end-to-end ASR models such as Wav2Vec 2.0 and Whisper models to compare the differences in performance.

Fine-tuning on stuttered speech. To study the impact of fine-tuning the two-stage systems on stut-

tering data, we perform an experiment to fine-tune the Whispering LLaMA framework on the LibriStutter dataset (Kourkounakis, 2021). We follow the fine-tuning methodology where the Whisper models generate multiple hypotheses and audio features. The audio features are integrated into the Alpaca model (Taori et al., 2023) through an adapter module. Specifically, the weights of the Whisper models and the LLM, except for the adapter module, are frozen, and we train the adapter module so that the LLM can correctly output the final transcription from the generated hypotheses through a completion prompt.

4 Experiments

In this section, we detail the experiments conducted to evaluate the performance of two-stage ASR systems on stuttered speech.

4.1 Dataset

We use the LibriStutter dataset (Kourkounakis, 2021) of synthetic stuttered speech, which includes 20 hours of English data with over 4,000 annotated samples from 50 speakers (23 males and 27 females). The dataset provides transcriptions for each sample, facilitating the training and evaluation of speech recognition models. LibriStutter focuses on repetitions, blocks, and prolongations, excluding non-speech sounds and unclear audio. It is split into training, validation, and test sets (80%, 10%, 10%). The samples have a minimum duration of 2 seconds, a maximum duration of 29 seconds, and an average duration of 14.2 seconds. Although the synthetic nature of the repetitions may raise concerns about the dataset's realism, analysis of mel spectrograms revealed the presence of noise and distortions in repeated segments, mitigating this issue (See Figure 2).

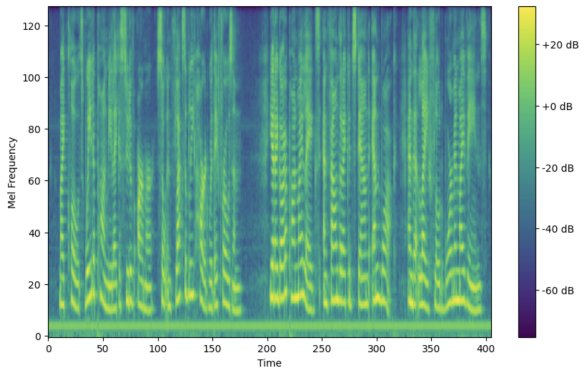

Figure 2: Audio example showing synthetic repetitions with noise

4.2 Evaluation Metric

We employ Word Error Rate (WER) as the primary evaluation metric to assess the performance of ASR systems. WER measures the rate of errors in the transcription by comparing it to the reference text and is calculated using the following equation:

$$\text{WER} = \frac{S + D + I}{N}$$

where S, D, I, and N indicate the number of substitutions, deletions, insertions, and the total number of words in the reference, respectively. Following the methodology in Radford et al. (2022), we normalize the output transcripts using Whisper's BasicTextNormalizer to ensure consistency and accuracy in WER calculation. The normalization removes punctuation, converts text to lowercase, and handles other formatting issues to create a uniform basis for comparison between the transcription and the reference text.

4.3 Baselines

For the baselines, we consider popular and recent ASR methods, which are Wav2Vec 2.0 (Baevski et al., 2020), Whisper (Radford et al., 2022), and WhisperingLlama (Radhakrishnan et al., 2023b). These conventional ASR technologies have been widely employed but often fall short when faced with stuttered speech, primarily because they are not designed to handle the irregular speech patterns that stuttering presents. For the WhisperingLlama result, we are using the paper's released checkpoint to evaluate its performance.

4.4 Experiment Details

For off-the-shelf experiments, we consider Whisper models with different sizes, including Tiny (39M parameters), Base (74M parameters), Small (244M parameters), and Medium (769M parameters). We generate multiple transcription hypotheses for each audio sample using beam search with a beam size of 1, temperature of 0.7, and no length penalty The generated hypotheses are then passed to the GPT-4o or LLaMA 2 language models for refinement. For the LLM generation of LLaMA 2, we use greedy sampling.

For the fine-tuning experiments on the LibriStutter dataset, we follow the Whispering LLaMA methodology. We freeze the weights of both the Whisper Medium and Alpaca models and fine-tune only the residual adapter modules inside the Alpaca

model. Specifically, we generate candidate transcriptions and acoustic feature embeddings from the stuttered speech input. The acoustic feature embedding serves as input to the adapters, and the language model is trained to produce the ground truth transcription from the context of these candidate transcriptions, as we do not modify the weights of the acoustic model. We adopt the original hyperparameters from the Whispering LLaMA paper except for setting the learning rate to 10^{-5} as we are fine-tuning the model as opposed to training from scratch. The fine-tuning step takes about one day to fine-tune the system for 10 epochs, fitting within our spending budget.

5 Results

In this section, we present the results of our experiments. We evaluate the performance of off-the-shelf models, including combinations of the Whisper ASR model and large language models (LLMs) like GPT-4o and LLaMA 2.

Inclusion of language model generally improves the performance of vanilla ASR models. For our main experiments, we use the Whisper Medium model to generate transcription hypotheses, passing 10 hypotheses to GPT-4o and 1 hypothesis to LLaMA 2. For the fine-tuned Whispering LLaMA, we pass 15 hypotheses. We discuss the number of hypotheses selection in the later ablation section. The results in Table 2 show that incorporating a language model, either through the off-the-shelf combination of Whisper and GPT-4o or through the fine-tuned Whispering LLaMA approach, improves the transcription performance compared to the Whisper Medium baseline. The language models leverage information from the multiple transcription hypotheses and audio features to refine the final transcription.

Improvement is more prominent in smaller acoustic models. The improvement from incorporating a language model is more significant for smaller Whisper models, as shown in Table 3. This suggests that even smaller acoustic models can achieve competitive performance when combined with a large language model, as the language model can leverage its linguistic knowledge to infer the correct transcription from the hypotheses.

Ablation study on the number of hypotheses. We perform an ablation study on the number of hypotheses passed to the large language model.

Model	WER
Baselines	
Wav2Vec 2.0	22.50
Whisper (medium)	15.38
Off-the-shelf models	
Whisper (medium) → GPT-4o	**13.28**
Whisper (medium) → LLaMA 2	19.89
Fine-tuned model	
Fine-tuned Whispering LLaMA (medium)	**14.23**

Table 2: Performance of ASR models on LibriStutter. The word error rate (WER) is measured using the Hugging Face evaluation metric. **Bold** numbers indicate improved performance compared to its vanilla counterpart.

	WER	
Model size	**Without GPT-4o**	**With GPT-4o**
Tiny	20.93	15.00
Small	19.53	13.51
Base	18.99	14.31
Medium	15.38	13.28

Table 3: Impact of using GPT-4o on the WER of different Whisper model sizes for stuttered speech recognition.

Our findings indicate that increasing the number of hypotheses improves performance for GPT-4o but degrades performance for LLaMA 2, as shown in Figure 3. The best configuration for LLaMA 2 is to use a single hypothesis, which still performs worse than the baseline. This indicates a limitation in LLaMA 2's capability to infer the correct transcription from multiple hypotheses.

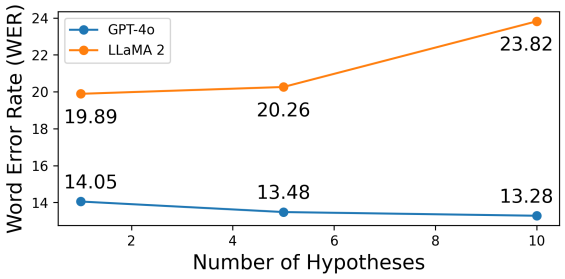

Figure 3: WER for different numbers of hypotheses with GPT-4o and LLaMA 2.

Fine-tuning Whispering LLaMA gives better results than the baseline. Fine-tuning the Whispering LLaMA system on stuttered speech data

improves its performance compared to the Whisper Medium baseline, as shown in Table 2. This validates that fine-tuning can help the language model better infer the correct transcription from the hypotheses. Notably, this fine-tuned and integrated system achieved similar WER to the conventional and segmented GPT-4o system that had the best WER, albeit based on the outdated and tinier Alpaca LLM, indicating that integrated approaches like ours work best. We also evaluated the fine-tuned Whispering LLaMA system on the LibriSpeech (Panayotov et al., 2015) dataset to ensure that its performance on normal speech is not degraded. The results show that the fine-tuned model's performance is comparable to the original Whisper model, suggesting that the fine-tuning process does not negatively impact its ability to transcribe non-stuttered speech (See Table 4).

Model	WER
Wav2vec 2.0	2.71
Whisper (large-v2)	2.7
Fine-tuned Whispering LLaMA (medium)	2.74

Table 4: Performance of ASR models on LibriSpeech Clean.

6 Qualitative Analysis

Qualitative analysis of the two-step systems reveals key insights into their strengths and weaknesses in transcribing stuttered speech. The Whisper ASR models, particularly the larger ones, generally capture the content words and overall meaning well (see Table 5). However, they tend to include filler words like "um", "ah", and "well", which are artifacts of disfluent speech. Table 5 shows an example where the Whisper model introduces filler words such as "was it" and "did not they" that are not present in the ground truth.

Incorporating language models like GPT-4o and LLaMA 2 helps refine the transcriptions and make them more fluent. GPT-4o corrects misspellings and generates semantically reasonable substitutions while largely maintaining the original meaning. However, LLaMA 2 is more prone to hallucinating words and making unusual substitutions, as demonstrated by the phrases "The sight of the f***ing gossiper" and "so resplendent and beauty" in the example in Table 5.

Increasing the number of hypotheses passed to the LLMs generally improves transcription quality by providing more context. With fewer hypotheses, there are more instances of omitted words, especially repetitions and filler words. The example in Table 5 illustrates how using only one hypothesis can lead to repetitive and less coherent transcriptions.

Despite the improvements, none of the models fully handle the disfluencies and repetitions characteristic of stuttered speech. Repetitions of sounds, syllables, and words are frequently omitted across all models. Fine-tuning the Whispering LLaMA framework on LibriStutter helps mitigate this issue to some extent. However, there is still room for improvement in developing more inclusive ASR systems that can accurately transcribe stuttered speech.

Model	Example
Whisper medium	"I kept the library as long as I could. We can sit on the stairs if you like. Which they proceeded to do, quite amiably..."
Whisper + filler words	"You must have thought me so rude, but indeed it was not my own fault, was it, Mrs. Allan? Did not they tell me that Mr. Tilney and his sister were gone out in a phaeton together?"
LLaMA 2 hallucinations	"The sight of the f***ing gossiper so resplendent and beauty acted upon him like magic..."
Fewer hypotheses	"I am back. All is well. Now listen. I am well. Now listen. I am well. Now listen."

Table 5: Examples of model outputs

7 Conclusion and Future Works

This project benchmarked state-of-the-art automatic speech recognition models on the task of recognizing stuttered speech using the LibriStutter dataset. We found that incorporating large language models into a two-stage framework generally improved performance over end-to-end ASR models. The language model helped refine transcriptions by leveraging information from multiple recognition hypotheses and audio features. This improvement was particularly noticeable for smaller acoustic models when paired with a powerful language model like GPT-4o. Additionally, fine-tuning the Whispering LLaMA framework on stuttered speech data further reduced the word error rate while maintaining performance in fluent speech.

For future works, extending this work to real-world stuttered speech datasets would also be valuable, as the synthetic nature of LibriStutter may not fully capture the complexity of authentic stuttering.

8 Examples

We include some sample LibriStutter audio files and their generated transcripts in this demo site.

Contributions

J.B. implemented off-the-shelf two-stage models for both GPT-4o and LLaMA 2 integration, fine-tuned the Whispering LLaMA model, evaluated all the models, wrote the proposal and milestones, Introduction, Related Works, Methodology, Experiments, Results, and Conclusion sections of the final report, and built the demo site. F.K. did initial dataset exploration, implemented loaders for the original WhisperingLlama with Libristutter, extracted hypotheses and audio features from Libristutter using Whisper for evaluation and fine-tuning, evaluated the fine-tuned model on LibriSpeech, and contributed to the Dataset and Results subsections. V.F. worked on team formation, proposal, identifying the Whispering LLaMA model and 3 datasets, setting up the LibriStutter dataset for implementation, and contributed to the Abstract, Introduction, Related Works, and Qualitative Analysis sections. J.B., F.K., and V.F. worked on the poster.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Sebastian Bayerl, Alexander Wolff von Gudenberg, Florian Hönig, Elmar Noeth, and Korbinian Riedhammer. 2022. KSoF: The kassel state of fluency dataset – a therapy centered dataset of stuttering. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1780–1787, Marseille, France. European Language Resources Association.

Anmol Gulati, Chung-Cheng Chiu, James Qin, Jiahui Yu, Niki Parmar, Ruoming Pang, Shibo Wang, Wei Han, Yonghui Wu, Yu Zhang, and Zhengdong Zhang, editors. 2020. *Conformer: Convolution-augmented Transformer for Speech Recognition*.

Jinxi Guo, Tara N. Sainath, and Ron J. Weiss. 2019. A spelling correction model for end-to-end speech recognition. In *ICASSP 2019*.

Ke Hu, Ruoming Pang, Tara N. Sainath, and Trevor Strohman. 2021. Transformer based deliberation for two-pass speech recognition.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Tedd Kourkounakis. 2021. LibriStutter.

Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Tien Dung Tran, Panayiotis Georgiou, Jeffrey Bigham, and Leah Findlater. 2023. From user perceptions to technical improvement: Enabling people who stutter to better use speech recognition. In *CHI 2023*.

Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey Bigham. 2021. PSep-28k: A Dataset for Stuttering Event Detection from Podcasts with People Who Stutter. In *ICASSP 2021*.

Ke Li, Jay Mahadeokar, Jinxi Guo, Yangyang Shi, Gil Keren, Ozlem Kalinli, Michael L. Seltzer, and Duc Le. 2022. Improving fast-slow encoder based transducer with streaming deliberation. In *ICASSP 2023*.

Linda Liu, Yile Gu, Aditya Gourav, Ankur Gandhe, Shashank Kalmane, Denis Filimonov, Ariya Rastrow, and Ivan Bulyko. 2021. Domain-aware neural language models for speech recognition. In *ICASSP 2021*.

Vikramjit Mitra, Zifang Huang, Colin Lea, Lauren Tooley, Sarah Wu, Darren Botten, Ashwini Palekar, Shrinath Thelapurath, Panayiotis Georgiou, Sachin Kajarekar, and Jefferey Bigham. 2021. Analysis and tuning of a voice assistant system for dysfluent speech. In *Proc. Interspeech 2021*, pages 4848–4852.

OpenAI. 2024. GPT-4o. Accessed: 06-07-2024.

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Jai Radhakrishnan, Rupak Sarkar, Tarun Arora, and Ashwini Kumar Yadav. 2023a. Whispering llama: A cross-modal generative error correction framework for speech recognition. *arXiv preprint arXiv:2305.13048*.

Srijith Radhakrishnan, Chao-Han Yang, et al. 2023b. Whispering LLaMA: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10007–10016. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Olabanji Shonibare, Xiaosu Tong, and Venkatesh Ravichandran. 2022. Enhancing asr for stuttered speech with limited data using detect and pass.

Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, Avinatan Hassidim, and Yossi Matias. 2019. Personalizing asr for dysarthric and accented speech with limited data. In *Interspeech 2019*. ISCA.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Jimmy Tobin and Katrin Tomanek. 2021. Personalized automatic speech recognition trained on small disordered speech datasets.

Hugo Touvron, Louis Martin, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Chao-Han Huck Yang, Linda Liu, Ankur Gandhe, Yile Gu, Anirudh Raju, Denis Filimonov, and Ivan Bulyko. 2021. Multi-task language modeling for improving speech recognition of rare words. In *IEEE Automatic Speech Recognition and Understanding (ASRU) 2021*.

Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G. 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 Rastow, and Ivan Bulyko. 2023. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE.

Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency detection using a bidirectional lstm.