LLM-BASED POST-ASR ERROR CORRECTION FOR DISORDERED SPEECH

Hangyi Wen, Mikiyas Assefa*, Anas Semsayan*, Eduardo Feo-Flushing*

Carnegie Mellon University
School of Computer Science
{hangyiw,massefa,asemsaya,efeoflus}@andrew.cmu.edu

ABSTRACT

Automatic speech recognition (ASR) systems achieve near-human accuracy on typical speech, but performance on disordered speech remains poor, with conversational word error rates (WER) often exceeding 50%. This gap creates serious accessibility barriers for individuals with communication disorders. We present the first systematic study of large language model (LLM)-based post-ASR error correction for disordered speech. Using the APROCSA corpus of conversational aphasic speech, we evaluate three complementary strategies: (i) multi-ASR fusion, where hypotheses from ten state-of-the-art ASRs are consolidated by LLMs; (ii) few-shot prompting for single-hypothesis correction; and (iii) supervised fine-tuning with parameter-efficient adapters. Results show that LLMs substantially reduce WER and improve semantic similarity, with fusion achieving up to 46% relative WER reduction and few-shot prompting exceeding 53%. By leveraging mainstream ASRs and applying lightweight LLM correction, our approach makes powerful recognition technology more accessible to speakers with disordered speech, lowering barriers to everyday communication.

***Index Terms*—** Speech recognition, Disordered speech, Large language models, Error correction, Accessibility

1. INTRODUCTION

Despite the impressive progress of automatic speech recognition (ASR) systems in recent years, their performance on speech from individuals with communication disorders remains alarmingly poor, with word error rates (WER) often exceeding 50% in conversational settings [1]. Disorders such as dysarthria, apraxia of speech, and aphasia introduce atypical articulation patterns, irregular rhythm, and reduced intelligibility that current ASR systems, trained primarily on typical, read speech, struggle to accommodate. This creates serious accessibility barriers, particularly in real-world applications where conversational interaction is essential [2].

Efforts to adapt ASR models to disordered speech, such as fine-tuning or reinforcement learning on specialized datasets, have shown measurable WER reductions [2]. However, these approaches require costly data collection and computational resources, and their gains often do not generalize across disorder types or conversational contexts. At the same time, mainstream ASR systems trained on typical speech continue to advance rapidly due to their scale and broad availability of training data, leaving disordered-speech ASR development comparatively lagging.

We therefore propose a complementary strategy: post-processing the outputs of strong, general-purpose ASR systems with large language models (LLMs). This approach inherits the rapid progress of state-of-the-art ASR while introducing a lightweight correction layer that can adapt to disordered speech with minimal supervision. LLM-based post-ASR correction requires no retraining of acoustic models, can prioritize semantic preservation over surface-level accuracy, and can flexibly operate on either single-hypothesis transcripts or multi-ASR fusions [3, 4, 5, 6, 7, 8]. These properties make it a practical and scalable path toward accessible speech recognition in everyday settings.

Our contributions are threefold. First, we present the first systematic study of LLM-based multi-ASR fusion for aphasic speech, evaluating hypotheses from ten state-of-the-art recognizers. Second, we introduce a lightweight few-shot prompting approach for single-hypothesis correction that requires no retraining and is practical for clinical or resource-limited settings. Third, we analyze how ASR quality, hypothesis diversity, and example selection influence performance, offering novel insights into the design of robust post-ASR correction. Together, these contributions demonstrate that LLMs can deliver substantial gains in both WER and semantic similarity, establishing post-processing as a promising new strategy for making mainstream ASRs more accessible to individuals with communication disorders.

2. RELATED WORK

LLMs have been explored in ASR, either by integrating them into the recognition pipeline or by using them for auxiliary tasks. One direction is to build LLM-based ASR systems that directly consume audio tokens, effectively replacing traditional language models [9]. LLMs have also been used as

*These authors contributed equally.

fallibility predictors, estimating the likelihood of recognition errors and guiding ASR model training [10]. In multilingual and code-switching scenarios, synthetic text generated by LLMs has been shown to substantially augment training corpora, achieving up to a 19% relative reduction in WER when native data is scarce [7]. These studies demonstrate that LLMs can improve ASR accuracy both as end-to-end models and as tools for data augmentation or error prediction.

While these studies focus on integrating LLMs into the ASR pipeline itself, another important line of work applies LLMs after recognition, where they act as post-processing modules for error correction. LLM-based post-processing generally falls into two categories: fine-tuning and prompt engineering. In fine-tuning, a pretrained model such as T5, GPT, or LLaMA is adapted on paired ASR hypotheses and reference transcripts [11]. This can be achieved through full parameter updates, which have been shown to substantially reduce WER on English, Japanese, and Chinese benchmarks [3, 8, 12], or through parameter-efficient methods such as LoRA/adapters, where only $\approx$0.1–1% of weights are trained [13]. Prompt engineering, in contrast, guides the LLM with carefully designed instructions and prompts [14]. Approaches include re-ranking prompts, where the LLM selects the best candidate from an N-best list [15], and generative prompts, where the LLM rewrites or corrects the hypotheses directly [5, 4]. Both fine-tuning and prompting can be further enhanced by ensembling, which aggregates hypotheses from multiple ASR systems or LLM outputs and applies voting techniques, often yielding relative WER reductions exceeding 30% [16].

Although LLM-based integration has improved ASR performance on typical speech, their effectiveness for atypical or disordered speech remains limited, with general-purpose systems often yielding conversational WERs above 50% [1]. Impairments such as slurred articulation, inconsistent rhythm, and atypical pronunciations are rarely represented in training corpora, leading to particularly poor recognition in spontaneous speech, where error rates can be nearly double those of read speech. In real-world applications such as conversational voice assistants [6], preserving semantic meaning is often more important than exact word matching, yet conventional evaluation metrics like WER fail to capture this dimension of performance.

To address these challenges, recent work has explored adapting LLM-based ASR to disordered speech by fine-tuning on mixed corpora of typical (LibriSpeech) and impaired (Euphonia) data [2]. Incorporating semantic accuracy alongside syntactic accuracy as a reward signal in reinforcement learning with human feedback (RLHF) further improved results, yielding up to 75% global accuracy. Nonetheless, significant room for improvement remains, and to our knowledge, no prior research has investigated the feasibility of LLM-based post-ASR error correction specifically adapted to disordered speech. This gap motivates the present study.

3. METHODOLOGY

We evaluate three strategies for LLM-based post-ASR error correction on disordered speech: (1) multi-ASR fusion, (2) few-shot single-hypothesis correction, and (3) supervised fine-tuning. All experiments are conducted on the APROCSA corpus [17], which contains six conversational samples from individuals with aphasia (20 minutes each). The speech is characterized by disfluencies, word-finding difficulties, and atypical prosody. Ground-truth transcripts remove fillers and stutters while preserving meaningful repetitions, providing a realistic benchmark for ASR correction. Evaluation relies on WER as the standard accuracy measure, and SBERT cosine similarity score [18] as a semantic similarity metric to capture meaning preservation beyond exact word matching.

For methods that require correction examples (few-shot correction and fine-tuning), the conversational speech was first segmented into utterance-level fragments and aligned with ground-truth transcripts using the JiWER library, yielding paired samples of ASR hypotheses and references. From this pool, three selection strategies were considered: (i) *Random Selection*, where examples were drawn at random from known errors; (ii) *Exhaustive Phonemes*, where examples were chosen to cover a broad range of phonetic sounds using the CMU Pronouncing Dictionary; and (iii) a *Data-Driven* strategy that prioritized examples by phonetic diversity and the presence of common filler words. This setup enabled us to assess whether broader phonetic coverage or targeted diversity yielded more robust correction performance. All LLM inference was solely text-based (no audio signal inputted) and performed with a temperature of 0 to ensure deterministic outputs, and random seeds were fixed for reproducibility. All source code, LLM prompts, sampling functions, and transcript data used in the experiments are available on GitHub[1].

3.1. Multi-ASR Fusion

We evaluate a fusion approach in which transcripts from multiple recognition services are combined by an LLM. Transcripts were obtained from ten systems: AssemblyAI Slam-1, AWS Transcribe, Microsoft Azure Speech, Deepgram Nova-3, ElevenLabs Scribe v1, GCP Chirp 2, Gemini 2.5 Pro Audio Understanding, Gladia Solaria, Speechmatics Ursa 2, and Whisper v3. These hypotheses were passed to three LLMs (Gemini 2.5 Pro, GPT-4.1, and DeepSeek R1) tasked with producing a fused transcript.

To assess the effect of input diversity, we sampled subsets of ASR services of varying sizes, generating up to 100 random combinations for each size. The fused transcripts (concatenated across all six samples) were then evaluated with WER. This analysis also allowed us to examine how both the number and the quality of input transcripts influence the relative WER gains achieved through fusion.

[1] https://github.com/cmu-impactlab/LLM-Corrector-for-Aphasic-ASR

3.2. Few-Shot Correction

We evaluate a more computationally efficient, single-hypothesis correction approach using few-shot prompt engineering, as a practical alternative to multi-ASR fusion. Using GPT-4.1, we corrected the individual outputs from the ten ASR services described above separately. Prompts were guided by small sets of correction examples, with three selection strategies compared as described above: *Random Selection*, *Exhaustive Phonemes*, and a *Data-Driven* strategy. We further analyzed the effect of varying the number of examples (2, 4, 6, 8, and 10) on performance. To avoid bias, utterances used as examples were excluded from final scoring.

3.3. Fine-Tuning

We investigate supervised fine-tuning (SFT) of an LLM to post-process ASR output, asking whether a parameter-efficient adapter can match or exceed the quality of fusion or prompting, and how the choice of training examples influences outcomes. From the longest one of the six APROCSA samples, we constructed paired datasets of ASR hypotheses and reference transcripts. Two subsets of 26 utterances ($\approx$ 40% of the sample) were selected using the *Random Selection* and *Exhaustive Phonemes* strategies described above, and separate models were trained on each.

Training followed an Alpaca-style [19] instruction format, with each JSONL entry containing an `instruction`, an `input` (raw ASR hypothesis), and an `output` (ground truth). We fine-tuned Qwen2.5-14B using LoRA (rank $r = 16$) applied to attention and MLP layers. Training used the TRL `SFTTrainer` with mixed precision (bf16) on 2 $\times$ A100 GPUs. Performance was evaluated in terms of WER against ground truth and relative WER improvement over raw ASR (excluding utterances used in training), consistent with the evaluation setup of the previous experiments.

4. RESULTS AND DISCUSSION

We report results for the three approaches introduced in Section 3: multi-ASR fusion, few-shot correction, and supervised fine-tuning. Performance is evaluated using WER and SBERT text similarity score to capture both surface-level accuracy and semantic preservation. WER improvements in our results represent relative percentages.

Figure 1 shows that LLM-driven fusion yields substantial gains for disordered speech. Individual ASR services varied widely, with WERs between 18–37% (mean 26%) and semantic similarity between 82–90% (mean 87%). Fusing these hypotheses reduced errors significantly: GPT-4.1 achieved the best performance with 14% WER and 93% semantic similarity, a 46% relative improvement in WER over the mean ASR baseline. DeepSeek R1 and Gemini 2.5 Pro also performed strongly, reaching 16% and 17% WER respectively, both with 92% semantic similarity.

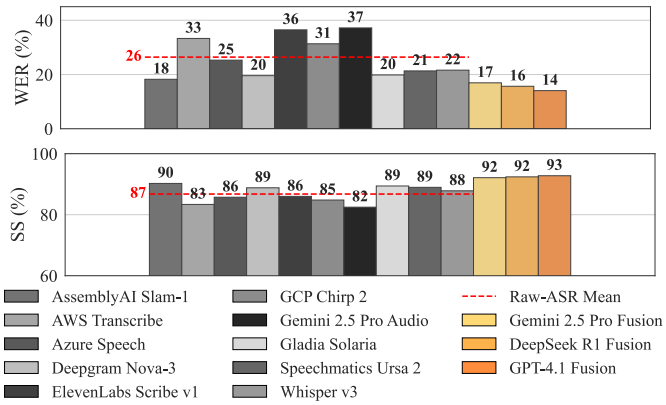

Fig. 1: Performance of ASR services and LLM-based fusion on disordered speech. SS denotes semantic similarity. Lower WER and higher SS are better.

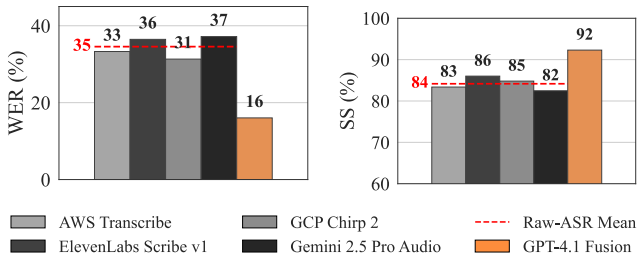

Fig. 2: GPT-4.1 fusion reapplied to four weaker ASR services. SS denotes semantic similarity.

Notably, when GPT-4.1 fusion was applied to four weaker ASRs with a mean WER of 35%, it still achieved 16% WER and 92% semantic similarity, the former a 54% relative improvement (Figure 2). This motivated a systematic analysis of input diversity (Fig. 3), which showed monotonic gains as more ASR services were fused, with median relative improvement rising from 15% for a single ASR to 45% for seven ASRs. Improvements were strongly correlated with the baseline WER of the inputs, indicating that lower-quality transcripts benefit the most from fusion. Overall, these findings demonstrate that LLM-based multi-ASR fusion provides consistent and scalable improvements for disordered speech.

While fusion provides strong gains, it requires access to multiple ASR services, which may not always be feasible in practice. Few-shot single-hypothesis correction offers a lighter-weight alternative and proved to be both practical and computationally efficient. However, fully on-device deployment may be unrealistic in many accessibility settings today due to LLM size. A strong correlation was observed between baseline ASR WER and relative improvement ($R^2 = 0.90$; Fig. 4), with the greatest gains for weaker ASRs and relative improvements exceeding 53%.

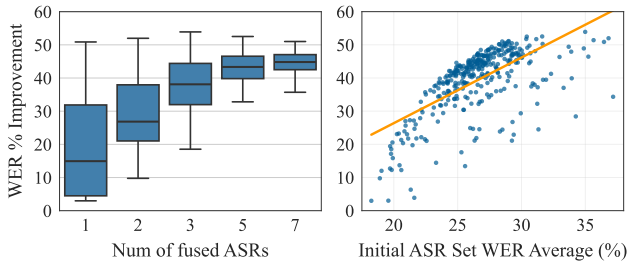

Fig. 3: Effect of input diversity and baseline ASR quality on WER improvement after GPT-4.1 fusion.

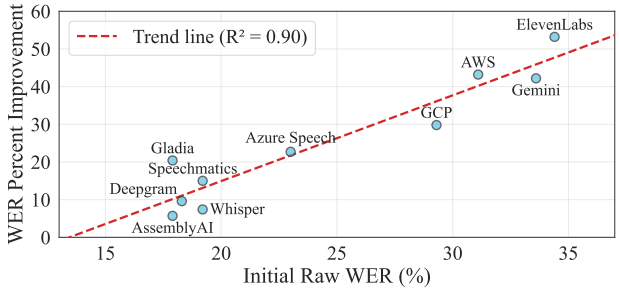

Fig. 4: Correlation between baseline ASR WER and relative improvement after few-shot correction (data-driven strategy, 6 examples).

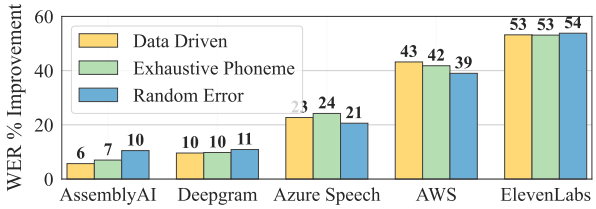

Fig. 5: WER improvement across few-shot example selection strategies with a fixed set of 6 examples.

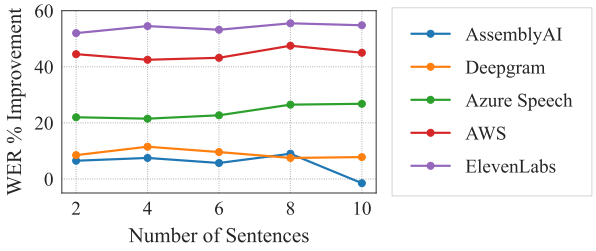

Fig. 6: Effect of the number of few-shot examples on WER improvement (data-driven strategy).

Analysis of example selection strategies (Fig. 5) showed that the strategic data-driven method was consistently effective, but random selection also performed surprisingly well,

Table 1: Supervised fine-tuning (SFT) results on the held-out APROCSA sample. Relative change (Δ) is computed against raw ASR; negative values indicate degradation. Note that the scores cannot be directly compared with previous experiments as they assess only one (not all) of the six samples.

Method	WER (%)	Relative Δ (%)
Raw ASR	31.19	—
Random Selection SFT	27.71	+11.1
Exhaustive Phonemes SFT	34.33	−10.1

achieving a peak improvement of 53.8% and establishing a simple yet strong baseline. Performance did not increase monotonically with the number of examples (Fig. 6); gains typically peaked with 4–8 examples, suggesting that example quality matters more than quantity and reinforcing the lightweight nature of this approach.

While few-shot correction proved effective with minimal supervision, we also examined whether supervised fine-tuning could provide further gains. As shown in Table 1, fine-tuning a Qwen2.5-14B LoRA adapter yielded mixed results depending on the data selection strategy. With the small SFT set built via random selection, WER dropped from 31.19% for raw ASR to 27.71%, an 11.1% relative improvement. In contrast, training on the exhaustive-phoneme set increased WER to 34.33% (10.1% worse than raw ASR), suggesting that phoneme coverage did not translate into better generalization under limited data.

We attribute the underperformance of SFT to the small training set, which likely caused the model to overfit to specific, rare error patterns rather than learning a general correction strategy. By contrast, few-shot prompting performed better because it leverages the model's massive pre-existing knowledge without the risk of distorting it with limited training examples. A limitation of this experiment is the small SFT set; with larger and more diverse data, phoneme-coverage strategies may better balance phonetic diversity with common error patterns and ultimately match or surpass random selection, which we leave to future work.

5. CONCLUSION

This work is, to our knowledge, the first to investigate LLM-based post-ASR error correction for disordered speech. Across multi-ASR fusion, few-shot prompting, and fine-tuning, our results show that LLMs can reduce WER while preserving semantic meaning. Post-ASR correction with LLMs offers a practical and scalable alternative to ASR retraining for disordered speech. More broadly, this study demonstrates that LLM-based post-processing can serve as an adaptation layer for challenging speech conditions without modifying acoustic models. Future work includes using larger datasets and covering a broader range of disorders.

6. REFERENCES

[1] Jimmy Tobin, Phillip Nelson, Bob MacDonald, Rus Heywood, Richard Cave, Katie Seaver, Antoine Desjardins, Pan-Pan Jiang, and Jordan R. Green, "Automatic Speech Recognition of Conversational Speech in Individuals With Disordered Speech," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 11, pp. 4176–4185, 2024.

[2] Chirag Nagpal, Subhashini Venugopalan, Jimmy Tobin, Marilyn Ladewig, Katherine Heller, and Katrin Tomanek, "Speech Recognition with LLMs Adapted to Disordered Speech Using Reinforcement Learning," in *ICASSP 2025*. 2025, pp. 1–5, IEEE.

[3] Rao Ma, Mengjie Qian, Mark Gales, and Kate Knill, "ASR Error Correction Using Large Language Models," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 33, pp. 1389–1401, 2025.

[4] Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill, "Can generative large language models perform asr error correction?," *arXiv preprint arXiv:2307.04172*, 2023.

[5] Sheng Li, Chen Chen, Chin Yuen Kwok, Chenhui Chu, Eng Siong Chng, and Hisashi Kawai, "Investigating ASR Error Correction with Large Language Model and Multilingual 1-best Hypotheses," in *Interspeech 2024*, 2024, pp. 1315–1319.

[6] John Harvill, Rinat Khaziev, Scarlett Li, Randy Cogill, Lidan Wang, Gopinath Chennupati, and Hari Thadakamalla, "Significant ASR Error Detection for Conversational Voice Assistants," in *ICASSP 2024*. 2024, pp. 11606–11610, IEEE.

[7] Ke Hu, Tara N. Sainath, Bo Li, Yu Zhang, Yong Cheng, Tao Wang, Yujing Zhang, and Frederick Liu, "Improving Multilingual and Code-Switching ASR Using Large Language Model Generated Text," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2023, pp. 1–7, IEEE.

[8] Moreno La Quatra, Valerio Mario Salerno, Yu Tsao, and Sabato Marco Siniscalchi, "FlanEC: Exploring Flan-T5 for Post-ASR Error Correction," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. 2024, pp. 608–615, IEEE.

[9] Yingyi Ma, Zhe Liu, and Ozlem Kalinli, "Effective Text Adaptation For LLM-Based ASR Through Soft Prompt Fine-Tuning," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. Dec. 2024, pp. 64–69, IEEE.

[10] Yingyi Ma, Zhe Liu, and Ozlem Kalinli, "Correction Focused Language Model Training For Speech Recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2024, pp. 10856–10860, IEEE.

[11] Saeed S. Alahmari, Lawrence O. Hall, Peter R. Mouton, and Dmitry B. Goldgof, "Repeatability of Fine-Tuning Large Language Models Illustrated Using QLoRA," *IEEE Access*, vol. 12, pp. 153221–153231, 2024.

[12] Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Macro Siniscalchi, Pin-Yu Chen, and Eng Siong Chng, "HyPoradise: An Open Baseline for Generative Speech Recognition with Large Language Models," Oct. 2023, arXiv:2309.15701 [cs].

[13] Nguyen Manh Tien Anh and Thach Ho Sy, "Improving Speech Recognition with Prompt-based Contextualized ASR and LLM-based Re-predictor," in *Interspeech 2024*. Sept. 2024, pp. 737–741, ISCA.

[14] Mo Wang, Minjuan Wang, Xin Xu, Lanqing Yang, Dunbo Cai, and Minghao Yin, "Unleashing ChatGPT's Power: A Case Study on Optimizing Information Retrieval in Flipped Classrooms via Prompt Engineering," *IEEE Trans. on Learning Technologies*, vol. 17, pp. 629–641, 2024.

[15] 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," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2023, pp. 1–8, IEEE.

[16] Yuka Ko, Sheng Li, Chao-Han Huck Yang, and Tatsuya Kawahara, "Benchmarking Japanese Speech Recognition on ASR-LLM Setups with Multi-Pass Augmented Generative Error Correction," 2024.

[17] Zoe Ezzes, Sarah M Schneck, Marianne Casilio, Davida Fromm, Antje S Mefferd, Michael de Riesthal, and Stephen M Wilson, "An open dataset of connected speech in aphasia with consensus ratings of auditory-perceptual features," *Data*, vol. 7, no. 11, pp. 148, 2022.

[18] Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proc. of EMNLP-ICJNLP*. 2019, ACL.

[19] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto, "Alpaca: A strong, replicable instruction-following model," *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, vol. 3, no. 6, pp. 7, 2023.