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
*Equal contribution

Making mainstream ASR more accessible to speakers with aphasia through lightweight LLM-based correction.

1) MOTIVATION

- ASR performs well on typical speech but remains weak on *disordered speech*.
- Conversational word error rates (WER) can exceed 50%, creating accessibility barriers.
- Goal: test whether LLMs can correct ASR transcripts for aphasic speech without retraining acoustic models.

2) DATASET & METRICS

- Dataset: APROCSA corpus
- 6 conversational samples from speakers with aphasia
- ~20 minutes per sample
- Ground truth removes fillers and stutters while preserving meaningful repetitions.
- Metrics: WER ↓ and SBERT semantic similarity ↑

3) METHODS

A. MULTI-ASR FUSION



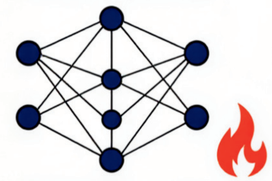
LLMs consolidate hypotheses from 10 state-of-the-art ASR services.

B. FEW-SHOT CORRECTION



GPT-4.1 corrects individual ASR transcripts using a small set of examples.

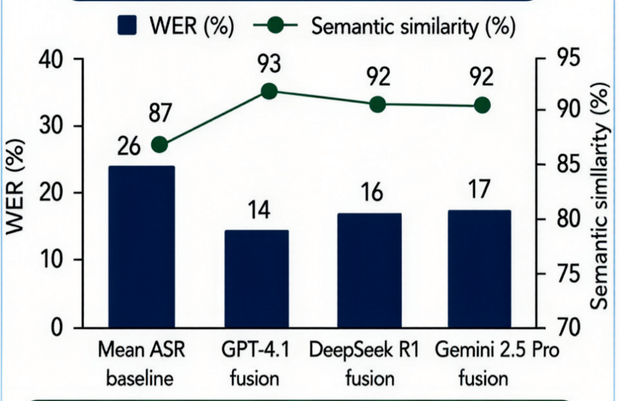
C. SUPERVISED FINE-TUNING



Qwen2.5-14B is adapted using LoRA on paired ASR/reference transcripts.

4) RESULTS

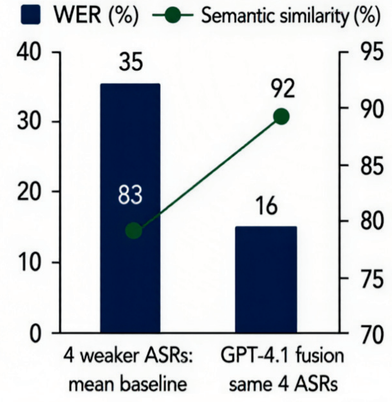
A. Fusion beats the ASR baseline



GPT-4.1 fusion: 14% WER, 93% similarity
46% relative WER reduction

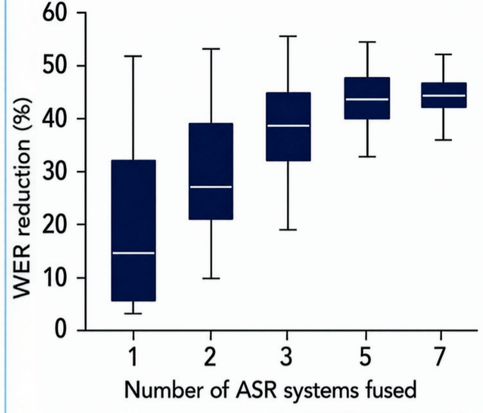
Fusion leverages complementary ASR errors and yields the strongest overall performance.

B. Even weak ASR outputs can be substantially improved



54% relative WER reduction (35% → 16%)

C. Diversity matters



Reduction rises from ~15% with 1 ASR to ~45% with 7 ASRs fused.

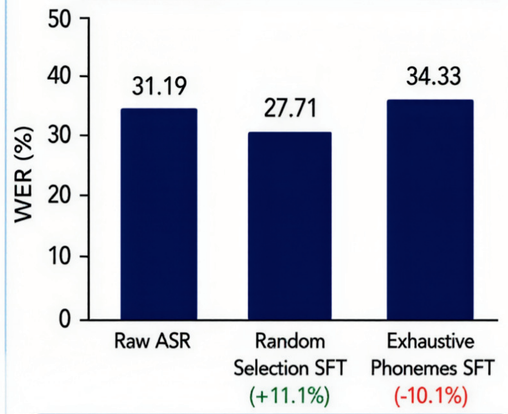
More diverse inputs generally produce larger gains.

D. Few-shot correction is lightweight and effective

- Relative reductions exceed 53%.
- Best performance usually occurs with 4–8 examples.
- Example quality matters more than quantity.
- Strong correlation between baseline ASR WER and gain: $R^2 = 0.90$.

Random selection performed surprisingly well; the data-driven strategy was consistently effective.

E. Fine-tuning gives mixed results with small data



With only 26 training utterances, LoRA fine-tuning likely overfits; few-shot prompting is more robust under limited data.

MAIN RESULT

Fusion achieved **14%** WER and **93%** semantic similarity.

Few-shot prompting achieved over **53%** relative WER reduction.

Small-data fine-tuning was less reliable.

5) CONCLUSION

- LLM-based post-ASR correction is a practical adaptation layer for disordered speech.
- Multi-ASR fusion gives the strongest gains.
- Few-shot prompting is lightweight and highly effective.
- Fine-tuning requires more data to be reliable.
- Future work: Larger datasets and broader coverage of speech disorders.

6) CONTRIBUTIONS

- First systematic study of LLM-based post-ASR correction for disordered speech.
- Comparison of fusion, prompting, and parameter-efficient fine-tuning.
- Analysis of ASR quality, input diversity, and example selection.

7) REFERENCES / RESOURCES

[1] Tobin et al., JSLHR 2024.
[2] Nagpal et al., ICASSP 2025.
[3] Ma et al., IEEE/ACM TASLP 2025.

Code & materials

github.com/cmu-impactlab/LLM-Corrector-for-Aphasic-ASR