Overall Performance Summary

Our fine-tuning of Whisper models for non-standard speech has yielded promising but uneven results across languages, model sizes, and—critically—across different speech impairment etiologies and severity levels. The English models demonstrate strong overall performance with the whisper-large model achieving 9.0% WER, 91.0% word accuracy, and 81% LATTEScore. However, these aggregate figures conceal significant disparities when examining specific medical conditions and impairment levels. Speakers with Parkinson's Disease achieve 75-80% meaning preservation, representing our strongest results, while those with Multiple Sclerosis see only 55-60% meaning preservation, indicating that not all impairments are equally supported by our current models. The severity-based analysis reveals equally important patterns, with mild impairment cases achieving 94.0% word accuracy compared to 68.3% for moderate cases and 72.7% for severe cases. This clear performance gradient underscores the challenge of serving users with more significant speech impairments.

Detailed English Model Performance

The English models demonstrate a clear correlation between model size and performance quality, but with substantial variation across both impairment severity and underlying etiology. Our analysis reveals that the large model preserves meaning in 85-90% of mild cases, 75-80% of moderate cases, and 65-70% of severe cases, showing a consistent 20-25 point performance drop from mild to severe impairments. The etiology-based analysis reveals even more pronounced disparities, with Parkinson's Disease speakers achieving 75-80% meaning preservation, neurodevelopmental disorder speakers reaching 65-75%, Cerebral Palsy speakers attaining 60-65%, and Multiple Sclerosis speakers struggling at only 55-60%. These gaps represent a 30-point performance difference between our best and worst supported conditions, highlighting significant equity concerns that must be addressed in future development cycles.

Detailed Swahili Model Performance

The Swahili models demonstrate concerning performance patterns that differ fundamentally from their English counterparts. The whisper-small model achieves a WER of 34.1%, word accuracy of 65.9%, and LATTEScore of 39.4% on the test set. Surprisingly, the whisper-large model performs worse across all metrics, with a WER of 41.6%, word accuracy of 58.4%, and LATTEScore of 22.0%. This inverse relationship between model size and performance suggests fundamental architectural mismatches or training data

inadequacies specific to Swahili non-standard speech patterns. The performance breakdown by severity level reveals uniformly poor results across all impairment categories, with mild cases achieving only 35-45% meaning preservation, moderate cases 25-35%, and severe cases 20-30%. This flat severity gradient indicates the models struggle with basic Swahili speech recognition regardless of impairment severity. Most alarmingly, the 26-36 percentage point gap between word accuracy and meaning preservation signals critical failures in semantic comprehension, where models recognize individual words but fail to capture overall meaning.

Swahili Model Challenges

The Swahili models present fundamental challenges that extend across all severity levels and etiologies, performing poorly regardless of impairment characteristics. Unlike the English models where performance varies systematically by condition, Swahili models achieve only 20-45% meaning preservation and 58-66% word accuracy across all user groups. This consistent underperformance suggests fundamental architectural or data-related issues rather than etiology-specific limitations. The particularly concerning performance of the large Swahili model, which shows worse results than its smaller counterpart across all metrics, indicates potential overfitting or fundamental misalignment with Swahili linguistic structures. The massive 26-36% gap between word accuracy and meaning preservation persists across all impairment types, signaling critical semantic comprehension failures that require complete architectural reconsideration.

Technical Insights and Metric Validation

Our multi-metric evaluation framework has proven particularly valuable for understanding performance patterns across different etiologies and severity levels. The consistency we observe across metrics within each etiology group validates our findings—for example, Multiple Sclerosis speakers show consistently poor performance across WER, word accuracy, and LATTEScore simultaneously. The severity gradient remains evident across all English etiologies but is most pronounced for Cerebral Palsy and Multiple Sclerosis speakers, where the combination of specific speech characteristics and severe impairment creates the most challenging recognition scenarios. This pattern holds true across all four evaluation metrics, confirming these user groups as priority areas for targeted improvement efforts in our development roadmap.

Practical Implications and Deployment Recommendations

For practical deployment, the English large model demonstrates sufficient performance to support users with mild to moderate speech impairments from most etiologies. However, the significant performance drop for severe impairments and for speakers with Cerebral Palsy or Multiple Sclerosis indicates these groups will require continued model refinement and potentially supplemental communication strategies. The etiology-based disparities raise critical equity concerns that cannot be overlooked—a user with mild Parkinson's may experience 85% meaning preservation and near-seamless interaction, while a user with severe Multiple Sclerosis might receive only 55% meaning preservation and find the system only marginally useful. This 30-point performance gap based purely on medical condition necessitates targeted, condition-specific improvements rather than continuing with one-size-fits-all solutions.

Future Directions

Moving forward, we must prioritize condition-specific and severity-aware optimization with immediate focus on improving performance for Cerebral Palsy and Multiple Sclerosis speakers across all severity levels. Simultaneously, we need to address the foundational issues affecting Swahili recognition through resource augmentation, transfer learning from better-performing English models, and exploration of language-specific acoustic models. The consistent patterns observed across metrics and speaker groups provide a clear, equity-informed roadmap for future work—one that prioritizes not only word-level accuracy but true meaning preservation for all users, regardless of language, impairment type, or severity level. Our specific targets include reducing the etiology-based performance gap from 30 points to under 15 points and ensuring minimum 60% meaning preservation for all severity-etiology combinations within the next development cycle.