

WHEN AI CANNOT HEAR APHASIA: DIAGNOSING AND IMPROVING AUTOMATED SPEECH RECOGNITION IN A THERAPY ROBOT FOR CANTONESE APHASIA PATIENTS



Enqiang (Yutong) Yuan

Duke University | Laidlaw Scholars Leadership and Research Program

Supervisor: Prof. Anthony Kong, The University of Hong Kong



1. BACKGROUND & RESEARCH QUESTIONS

- **Aphasia** impairs speaking, understanding, reading, or writing, most often after a stroke. Access to consistent speech-language therapy is limited by clinician availability.
- Joey, an **AI companion robot**, delivers structured naming and social-communication exercises to extend therapy beyond the clinic.
- **Aphasic speech** shows reduced rate, false starts, repeated/revised material... + **Cantonese's six contrastive lexical tones**
- Clinicians report a **high error rate** in word & error identification results in Joey's automatic speech recognition (ASR) system.

RQ1: Are Joey's failures mainly caused by inaccurate speech transcription or downstream response evaluation?

RQ2: What features of Cantonese aphasic speech make it difficult for general-purpose ASR systems?

RQ3: Can prompt-level changes or small-scale fine-tuning improve performance?



2. METHODOLOGY

1 Characterize aphasic speech

Automated **CHAT-annotation parsing** in **CanAB corpus**: (repetitions, false starts, substitutions, unintelligible speech) + **Independent Praat acoustic analysis** of rate, pausing, and duration.

2 Isolate the ASR bottleneck

Benchmark **Whisper** and a transcription-only configuration of **Gemini** on held-out aphasic vs. control recordings, scored with **Character Error Rate (CER)**

3 Prompt-based interventions

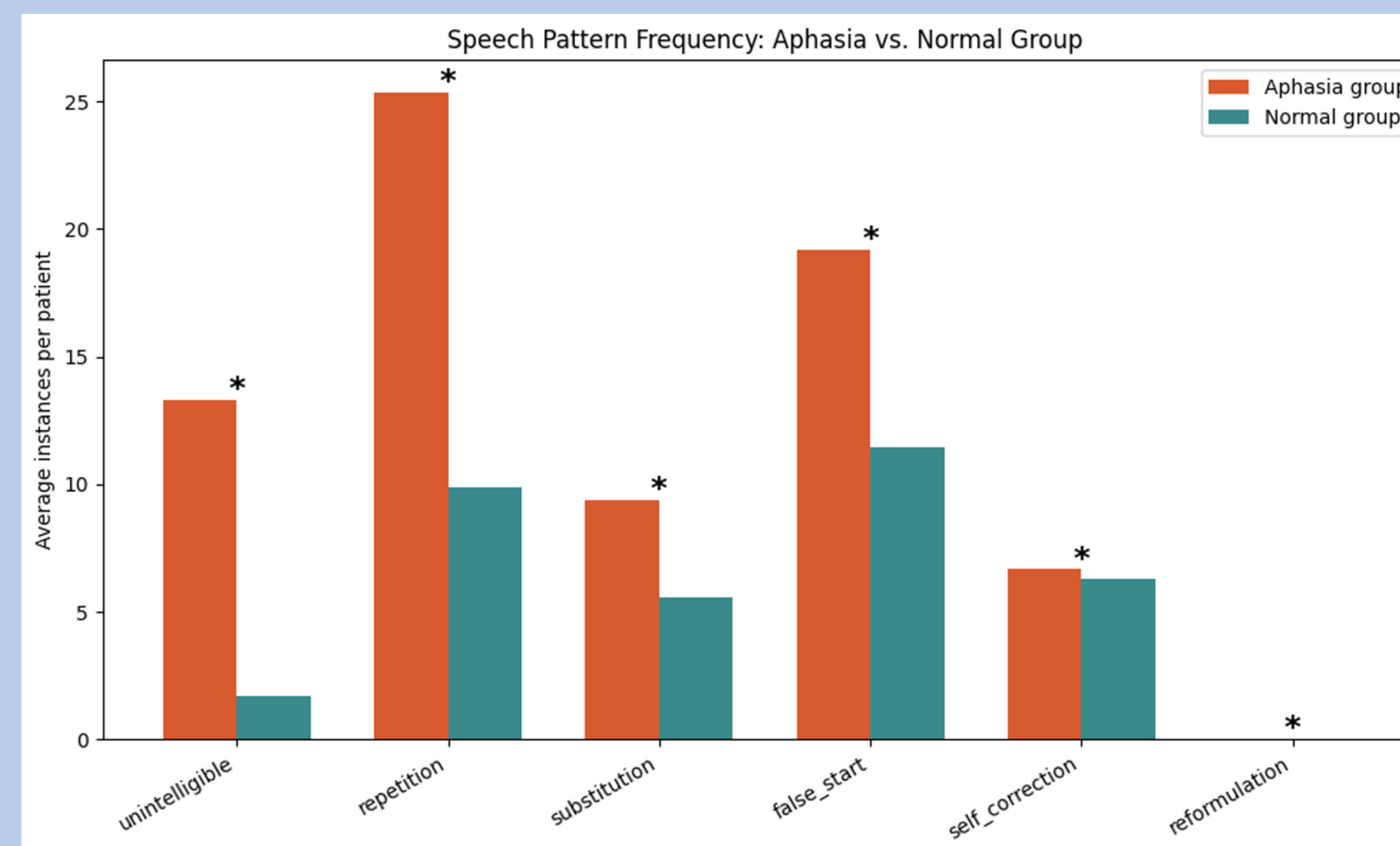
Design and test **three rule-based patches** targeting low-confidence input, repetition parsing, and aphasia-specific sound substitutions.

4 Supervised fine-tuning

Fine-tune Gemini via Vertex AI on 54 → 135, paired CanAB audio/expert-transcript examples; re-score with CER

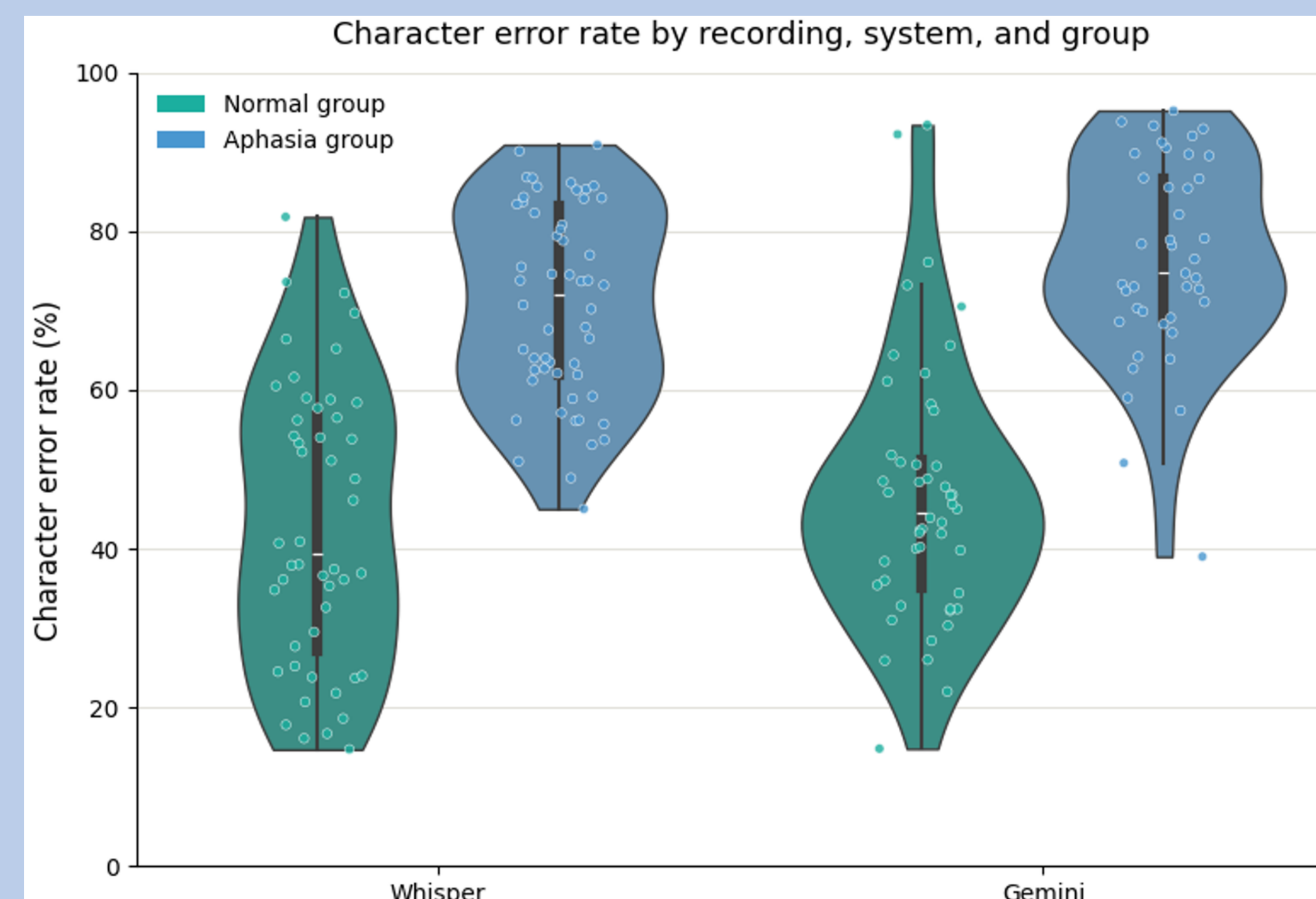
3. RESULTS

1. Aphasic speech's patterns quantified



- Unintelligible speech occurred ~7.8× more often, and repetitions ~2.6× more often, among speakers with aphasia (all $p < .001$); false starts and sound/word substitutions were also significantly more frequent.
- An independent Praat acoustic analysis found longer syllable duration, slower articulation and speech rate, and more pauses in aphasic speech (all $p < .001$)

2. Transcription as the main bottleneck



- Whisper CER: 42.9% (control) → 70.8% (aphasia);
- Gemini CER: 46.5% (control) → 76.4% (aphasia)
- The similar degradation across two different systems indicates that the difficulty lies in transcribing aphasic Cantonese itself, not only in Joey's downstream evaluation logic.

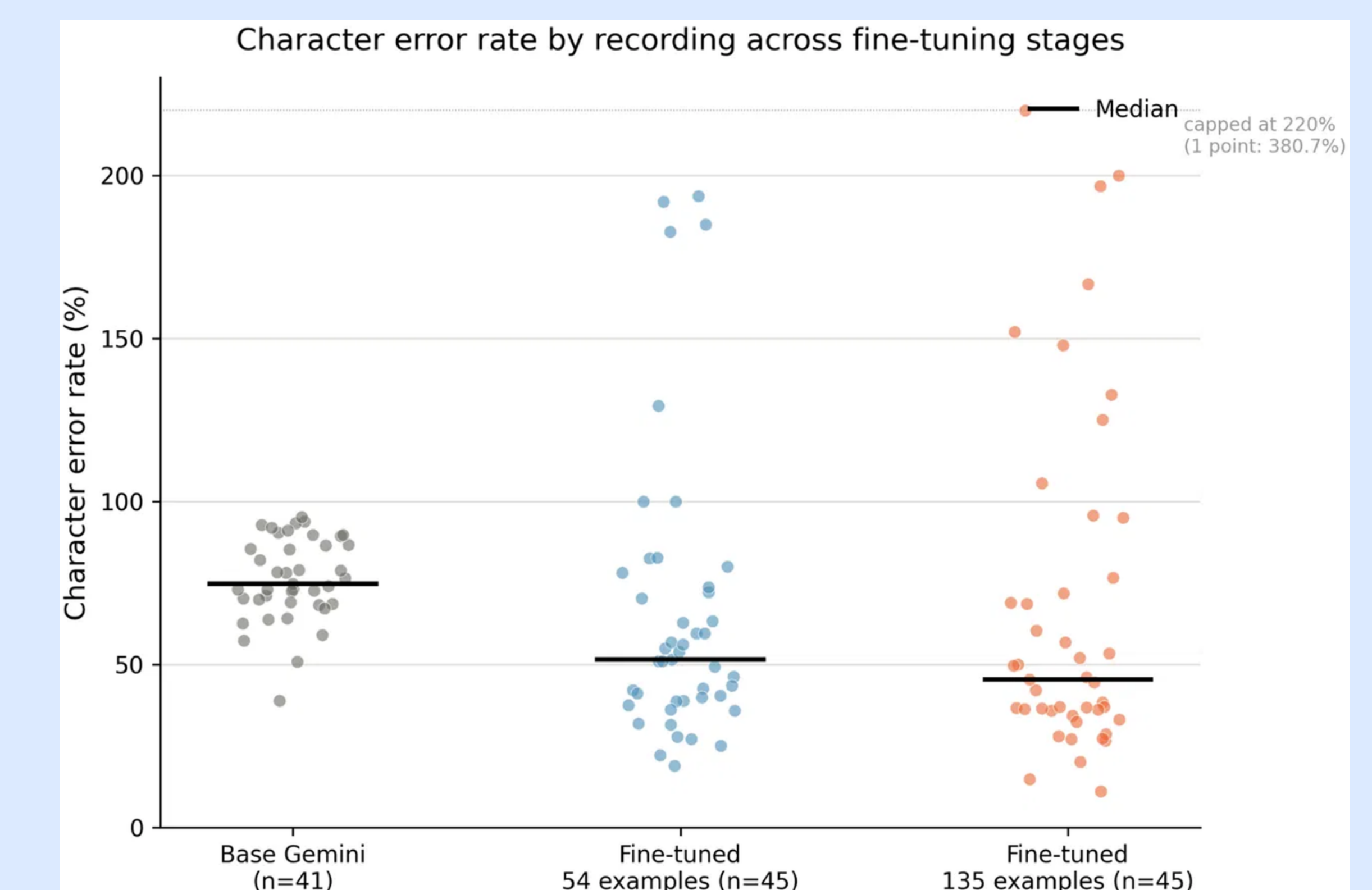
3. Effectiveness of Prompt Patches vs. Fine-Tuning

Prompt-based intervention:

Patch	Intended behavior	Patch excerpt	Evaluation outcome
Confidence handling	Avoid guessing on unclear input	"If the audio is unclear, explicitly indicate uncertainty rather than infer..."	Mixed; 20/45 improved, 22/45 worsened
Repetition handling	Prevent repeated fragments from dominating output	"Resolve repeated or incomplete material to the final intended production..."	Improved targeted examples; did not eliminate looping generally
Candidate substitution	Avoid premature correct/incorrect classification	"Flag plausible aphasia-related substitutions for review..."	Altered judgment in intended direction; exploratory only

Takeaway: individual patches successfully altered recognition results, yet the effectiveness remains mixed and unstable across specific tasks and cases.

Fine-tuning intervention



- With 54 training pairs, mean CER fell from 76.4% → 65.8% on held-out speakers (~50.0% excluding one difficult patient).
- Expanding to 135 pairs improved CER on 25/45 recordings; excluding known outliers, mean CER fell further to 43.3% (a few severe failures persisted).

4. CONCLUSION & DESIGN IMPLICATIONS

- Joey's failures begin at the transcription stage, not only at response evaluation.
- **Domain-specific fine-tuning is the most promising strategy for improving transcription of aphasic Cantonese.**
- **A layered solution is needed: fine-tuning + targeted prompt rules + system-level adjustments for slower, more disfluent speech.**

Acknowledgments: Prof. Anthony Pak-Hin Kong; Hannah Wu; Prof. Qin Ying; Duke Laidlaw Scholars Program.