# Conversational Speech Reveals Structural Robustness Failures in LLMs

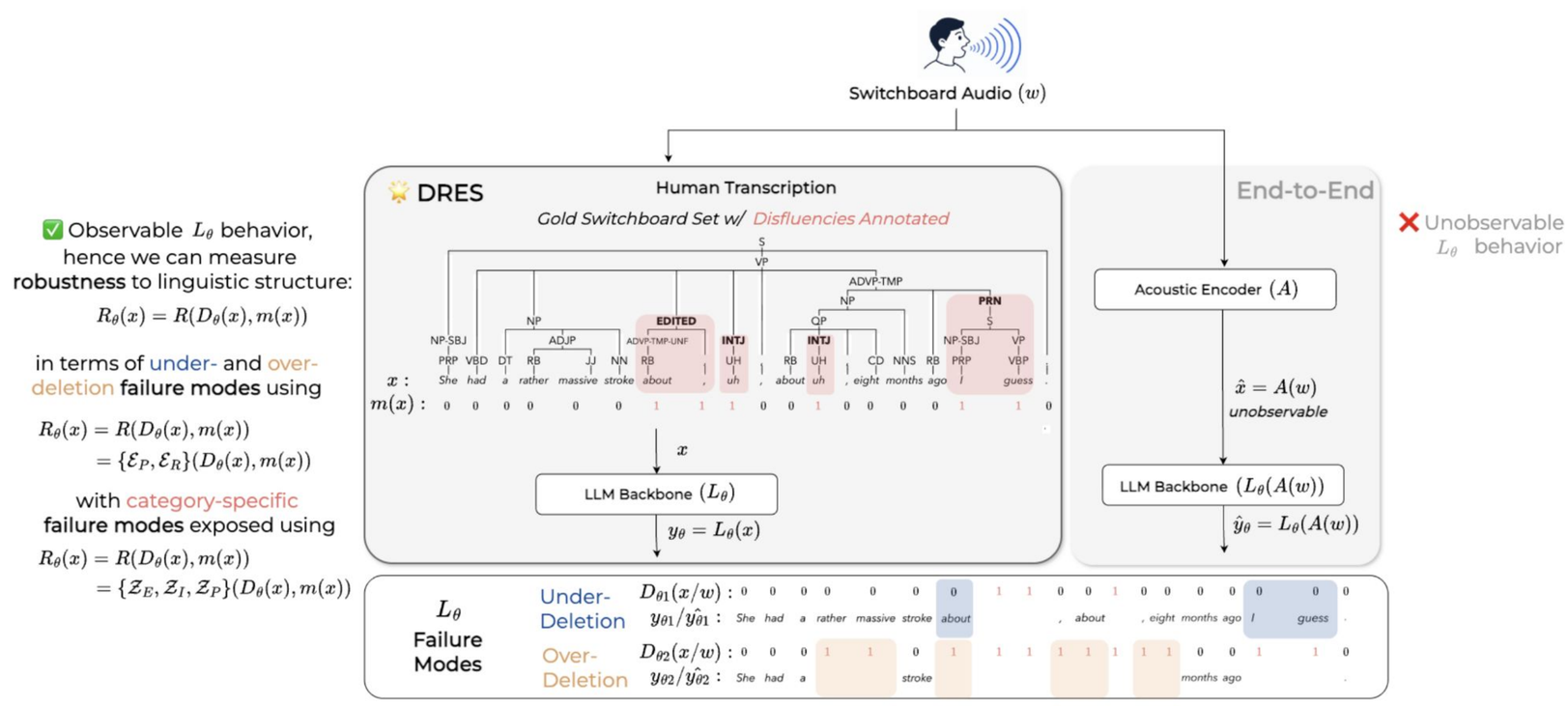
**Maria Teleki, Sai Tejas Janjur, Haoran Liu, Oliver Grabner, Ketan Verma, Thomas Docog, Xiangjue Dong, Lingfeng Shi, Cong Wang, Stephanie Birkelbach, Jason Kim, Yin Zhang, Éva Székely, James Caverlee**

***Texas A&M University, KTH Royal Institute of Technology, Sweden***

## Introduction

SpeechLLMs have recently ballooned in popularity from being utilized in voice assistants, meeting transcriptions and multimodal systems but the **text** they train on **rarely** contain conversational disfluencies.

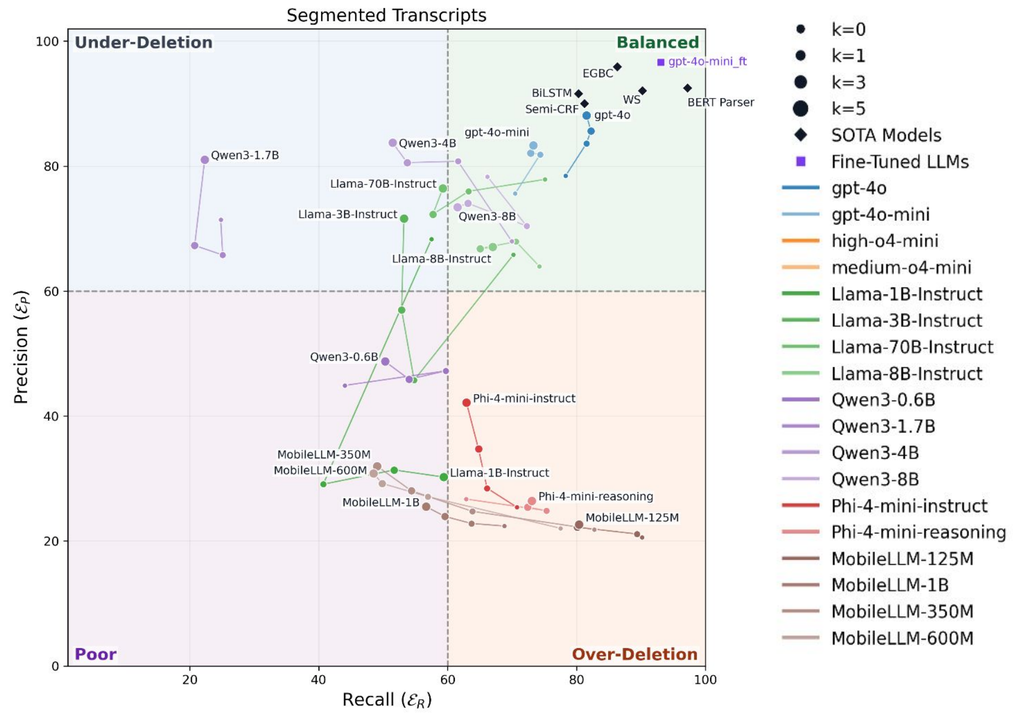

## DRES Framework



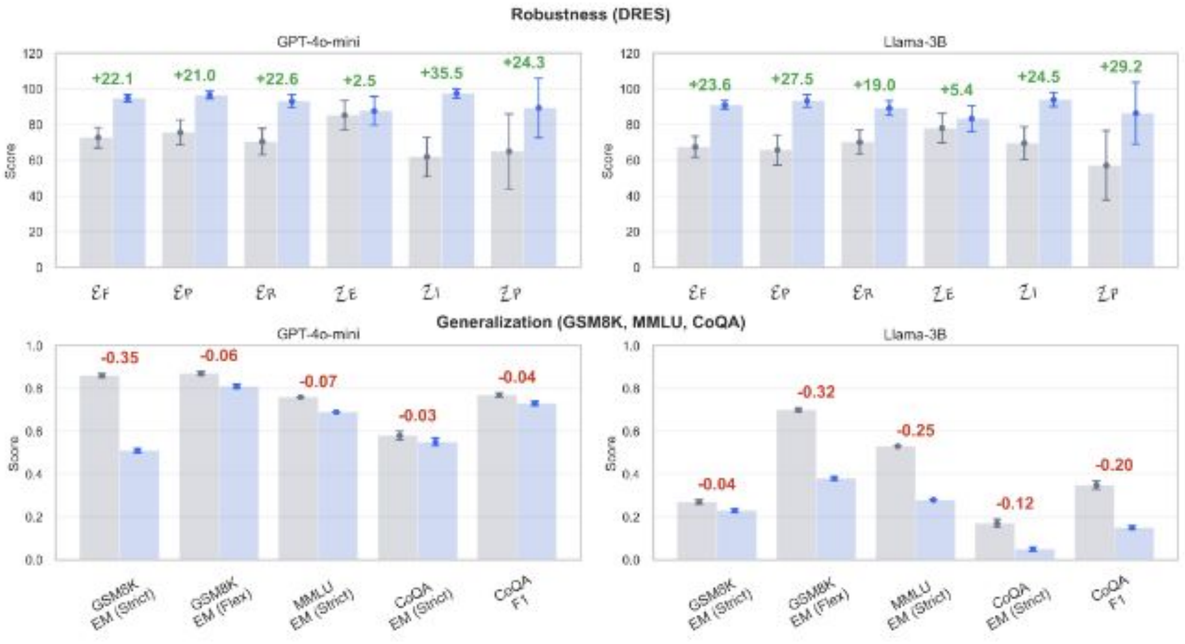

## Editing Policy in Precision Recall Space

Under-Deletion | Balanced | Poor | Over-Deletion

- OpenAI's GPT models are generally balanced
- **Reasoning models** (o4-mini, Phi-4 reasoning) shift to over deletion as they delete fluent content
- **Small models** fall under Under Deletion / Poor


## Generalization




## Key Findings

- Fine Tuning **improves** DRES $\mathcal{E}_F$ by +22 pts (GPT-4o mini and Llama-3B) but **degrades** in generalized evaluations (GSM8K, MMLU, and CoQA)
- Segmenting transcripts **improves** $\mathcal{E}_F$ by +1.4 to +31.4 across all models, even those with 128k context windows. Failures come from context instability
- Scale improves performance within a model family but does NOT change **editing policy**

## See Also

**Z-Scores:** A Metric for Linguistically Assessing Disfluency Removal (Teleki, Janjur, Liu, Grabner, et al., *ICASSP 2025*)
Powers the category-level Z-Score diagnostics used in this work.

Texas A&M University
Department of Computer Science & Engineering