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 **surged** in popularity for voice assistants, meeting transcription, and multimodal systems. Yet they train on **text** that **rarely** contains conversational disfluencies.

DRES: A Framework for Measuring Structural Robustness

✅ Observable L_θ behavior, hence we can measure **robustness** to linguistic structure:

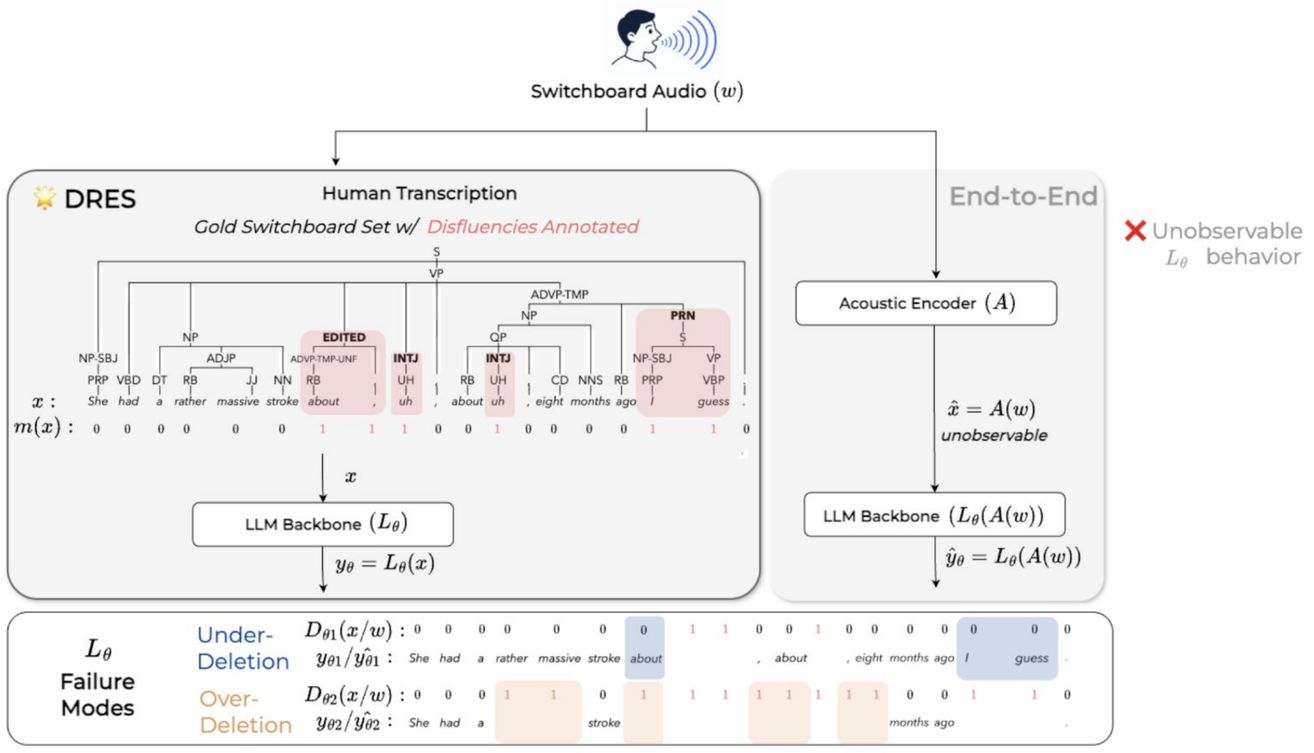
$$R_\theta(x) = R(D_\theta(x), m(x))$$

in terms of under- and over-deletion **failure modes** using

$$R_\theta(x) = R(D_\theta(x), m(x)) = \{\mathcal{E}_P, \mathcal{E}_R\}(D_\theta(x), m(x))$$

with category-specific **failure modes** exposed using

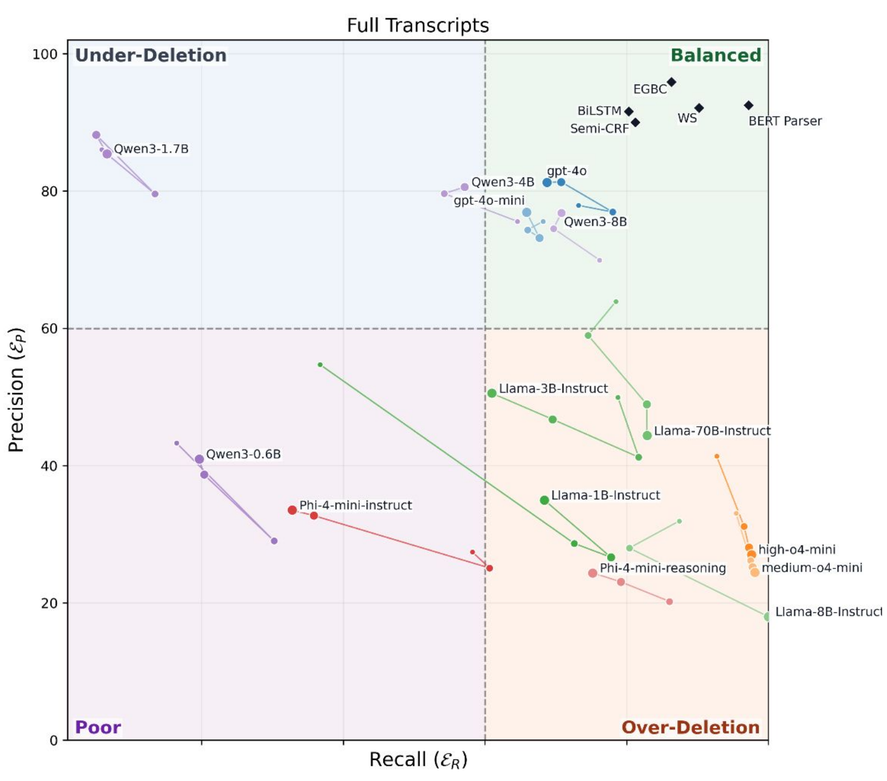
$$R_\theta(x) = R(D_\theta(x), m(x)) = \{\mathcal{Z}_E, \mathcal{Z}_I, \mathcal{Z}_P\}(D_\theta(x), m(x))$$

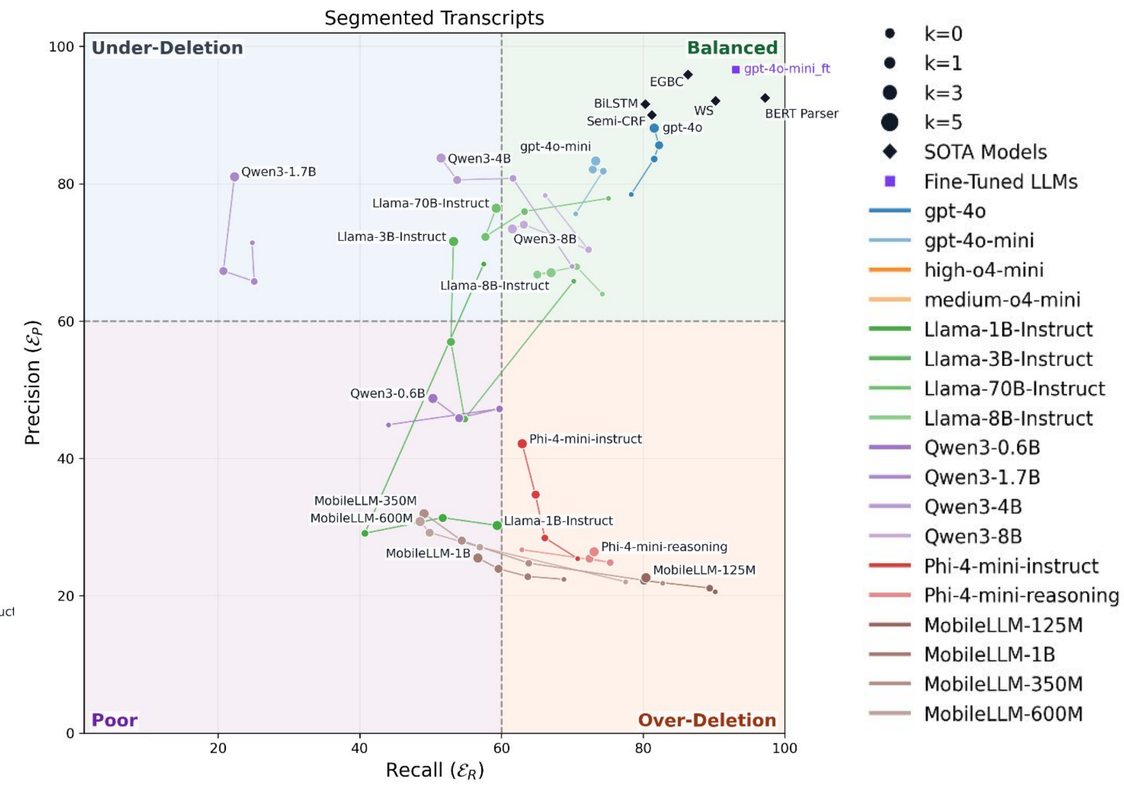


Editing Policy in Precision–Recall Space

Under-Deletion | Balanced | Poor | Over-Deletion

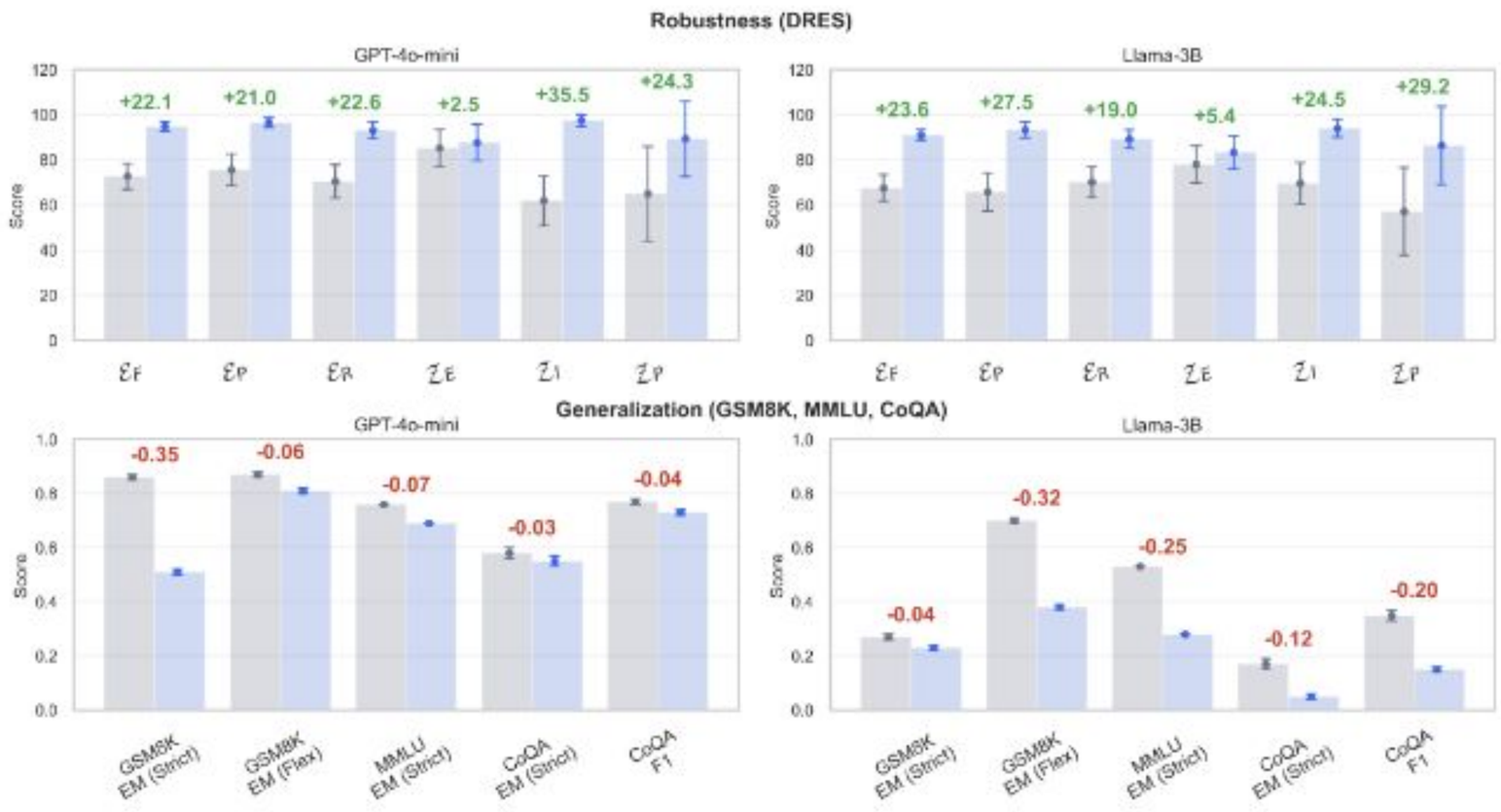
- **OpenAI's frontier GPT models have strong robustness**, following the Balanced policy
- **Reasoning models have poor robustness** (o4-mini, Phi-4 reasoning), following the Over- Deletion policy
- **Small models have poor robustness**, following the Under-Deletion / Poor policies





Robustness–Generalization Tradeoff

- **Scale improves robustness** within a model family but does NOT change the **editing policy**
- **Fine-tuning improves robustness** by +22 pts (GPT-4o mini and Llama-3B) but **degrades generalization** (GSM8K, MMLU, and CoQA)
- **Segmenting transcripts improves robustness** by +1.4 to +31.4 across all models, even those with 128k context windows



See Also →

Z-Scores: A Metric for Linguistically Assessing Disfluency Removal (Teleki, Janjur, Liu, Grabner, et al., ***ICASSP 2026***)

Powers the category-level Z-Score diagnostics used in this work.

Texas A&M University
Department of Computer Science & Engineering