Conversational Speech Reveals Structural Robustness Failures in SpeechLLM Backbones

Maria Teleki[1], Sai Tejas Janjur[1], Haoran Liu[1], Oliver Grabner[1], Ketan Verma[1], Thomas Docog[1], Xiangjue Dong[1], Lingfeng Shi[1], Cong Wang[1], Stephanie Birkelbach[1], Jason Kim[1], Yin Zhang[1], Éva Székely[2], James Caverlee[1]

[1]Texas A&M University, USA
[2]KTH Royal Institute of Technology, Sweden

Keywords: speech recognition, human-computer interaction, computational paralinguistics

Abstract

LLMs serve as the backbone in SpeechLLMs, yet their behavior on spontaneous conversational input remains poorly understood. Conversational speech contains pervasive disfluencies – interjections, edits, and parentheticals – that are rare in the written corpora used for pre-training. Because gold disfluency removal is a deletion-only task, it serves as a controlled probe to determine whether a model performs faithful structural repair or biased reinterpretation. Using the DRES evaluation framework, we evaluate proprietary and open-source LLMs across architectures and scales. We show that model performance clusters into stable precision-recall regimes reflecting distinct "editing policies." Notably, reasoning models systematically over-delete fluent content, revealing a bias toward semantic abstraction over structural fidelity. While fine-tuning achieves SOTA results, it harms generalization. Our findings demonstrate that robustness to speech is shaped by specific training objectives.

1 Introduction

As SpeechLLMs become central to voice assistants, meeting transcription, and multimodal conversational systems, a key assumption has emerged: increasing model scale and reasoning capability improves robustness to real-world speech [1, 2, 3, 4]. We show that this assumption is incomplete. Conversational speech reveals structural failure patterns in current LLM backbones that are not clearly captured by standard semantic benchmarks or aggregate end-to-end metrics.

Spontaneous conversational speech contains pervasive disfluencies – interjections (*uh*, *um*), repetitions, false starts, and parentheticals (*you know*, *I mean*) – that are intrinsic to incremental human production [5, 6, 7, 8, 9, 10, 11]. Consider the utterance:

> *"I uh I mean the other driver was — was going through the red light when the crash happened."*

A fluent representation of this utterance is:

> *"The other driver was going through the red light when the crash happened."*

The transformation required here is subtractive: disfluent spans are removed while the fluent content is otherwise preserved. In other words, the fluent transcript corresponds to a monotonic subsequence of the original input. This constraint makes disfluency removal a tightly controlled transformation: there is no paraphrastic freedom, and additional deletion or rewriting constitutes a structural error.

Yet large generative models are optimized for abstraction, compression, and semantic reinterpretation. These objectives conflict with deletion-constrained repair, which requires strict preservation of the original token sequence while removing only disfluent spans. As a result, models may rewrite or reinterpret conversational structure rather than faithfully repairing it. However, conversational disfluencies carry important paralinguistic signals – for example, filled pauses can convey speaker uncertainty or cognitive state [12, 13, 14]. Misinterpreting these

[0]Under review at INTERSPEECH '26.

structures can therefore have downstream consequences in high-stakes settings, including linguistic forensics and judicial decision-making [15, 16, 17, 18, 19], medical documentation and decision-making [20], and social reasoning tasks [21] including personality assessment [22] and deception detection [23, 24]. We therefore ask:

> ***Are robustness failures in SpeechLLMs driven by limitations in how their LLM backbones handle conversational speech?***

To investigate this question, we introduce a controlled probe: conversational disfluency removal. Given gold annotations, the correct output is uniquely defined by a deletion-only transformation. The resulting fluent transcript must preserve all fluent tokens from the input while removing only annotated disfluencies. Under this constraint, robustness reduces to token-level agreement with a gold deletion mask. Any over-deletion or under-deletion directly reveals structural editing errors.

We operationalize this idea through the **Disfluency Removal Evaluation Suite (DRES)**, a structural evaluation framework designed to isolate language-level editing behavior in SpeechLLM backbones. Unlike end-to-end speech benchmarks, which conflate acoustic transcription errors with language-level editing decisions, DRES evaluates models using fixed gold conversational transcripts. This factorization allows **structural editing behavior** to be measured independently of acoustic suppression effects.

Using DRES, we evaluate a diverse set of proprietary and open-source LLM backbones spanning architectures, parameter scales, prompting regimes, and reasoning variants (Table 1). Our analysis reveals that models exhibit stable editing policies characterized by distinct precision–recall trade-offs in deletion behavior. Notably, reasoning-oriented models tend to over-delete fluent content, reflecting a bias toward semantic abstraction rather than structural fidelity. While lightweight fine-tuning can substantially improve structural repair performance, it introduces measurable degradation on unrelated reasoning and knowledge benchmarks.

We contribute:

- **DRES: a structural evaluation framework for SpeechLLM backbones (Figure 1, §3).** We introduce a factorized evaluation protocol that isolates language-level editing behavior by providing models with gold conversational transcripts and enforcing deletion-only constraints. We will open-source the code for DRES.
- **A structural definition of conversational robustness (§3.3).** We formalize robustness as deletion-constrained repair and measure it through token-level agreement with gold deletion masks, enabling direct analysis of over-deletion and under-deletion errors.
- **Empirical identification of editing policies in LLM backbones (Figure 2, §3, §4).** Across diverse proprietary and open-source models, we show that LLMs cluster into stable precision–recall regimes corresponding to **under-deletion**, **over-deletion**, **balanced**, and **poor** editing policies shaped by training objectives.
- **Evidence of a robustness–generalization trade-off under adaptation (§4.6).** Fine-tuning significantly improves structural fidelity on disfluency removal but degrades performance on reasoning and knowledge benchmarks, indicating a specialization cost.
- **Practical deployment recommendations (§4.7).** We translate these empirical findings into practical recommendations, including transcript segmentation for stability, editing-policy-aware model selection, and monitoring generalization degradation under fine-tuning.

Conversational speech therefore provides a controlled stress test for structural alignment in language models. As SpeechLLMs become increasingly integrated into real-world systems, evaluating robustness requires not only measuring semantic accuracy but also auditing how faithfully models preserve the structure of conversational language.

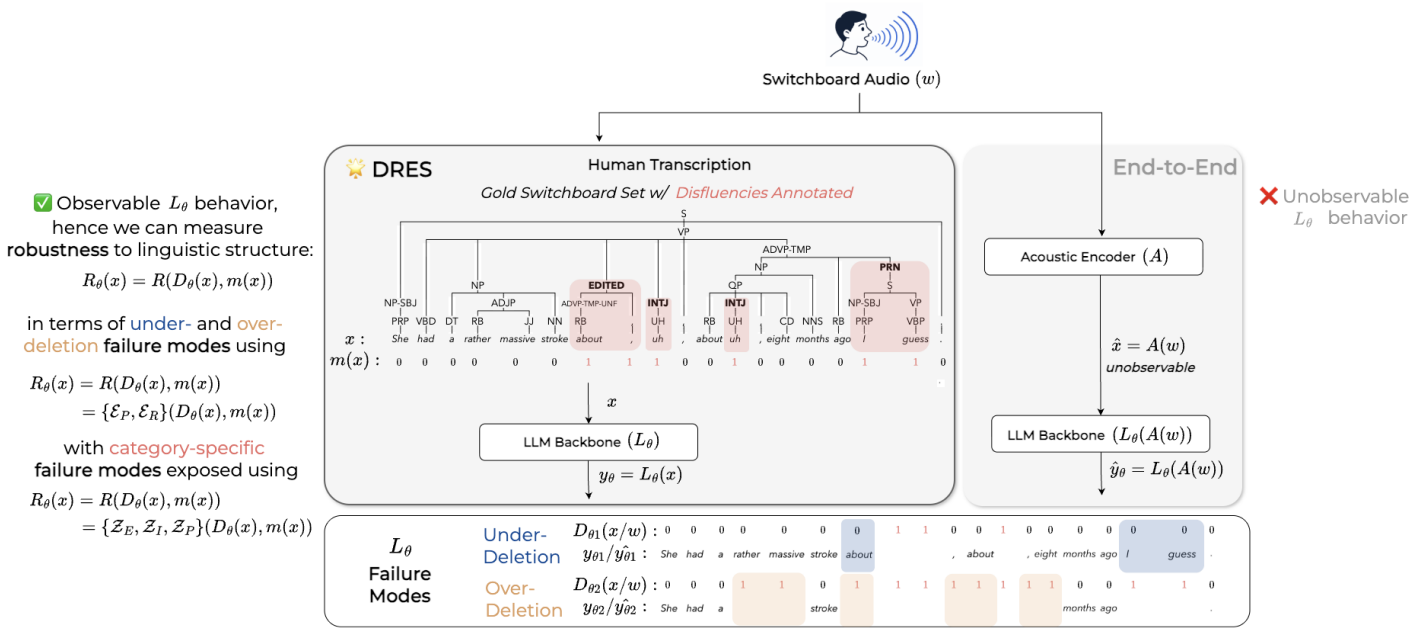

Figure 1: **Overview of DRES, a factorized structural evaluation framework for SpeechLLM backbones:** Gold Switchboard transcripts with annotated disfluencies are provided directly to the LLM backbone to isolate language-level deletion behavior ($D_\theta(x)$) from acoustic suppression effects (unobservable $A(w)$) in end-to-end SpeechLLMs. Structural robustness is computed by comparing the model-implied deletion mask to the gold mask, decomposing errors into (i) over-deletion and under-deletion failure modes in the precision–recall space (Figure 2), and (ii) category-specific failure modes (Figure 3).

2 Related Work

2.1 SpeechLLMs and End-to-End Evaluation

Recent SpeechLLM architectures combine acoustic encoders with pretrained LLM backbones to produce end-to-end systems mapping speech to text outputs [25, 26, 27, 28]. Benchmarks such as VocalBench-DF [29], WildSpeech-Bench [30], and VoxEval [31] evaluate system-level robustness to disfluent speech under realistic acoustic conditions. These efforts provide important insights into deployment behavior.

However, end-to-end evaluation entangles acoustic transcription and language-level editing policies. Modern ASR systems are known to under-transcribe certain disfluencies, especially short interjections, which can influence downstream task performance [32, 33, 34]. Further, ASR systems are known to error specifically on disfluent speech, with harmful hallucinations [35, 20]. As a result, improvements in system-level robustness may reflect acoustic suppression or expansion, language-level editing behavior, or their interaction.

DRES complements end-to-end benchmarks by operating on fixed gold transcripts and enforcing deletion-only evaluation, *isolating how backbone LLMs understand highly disfluent conversational speech.* By removing acoustic variability, it isolates backbone editing behavior and enables direct measurement of structural fidelity in conversational repair.

2.2 Disfluency Detection and Removal

Research on disfluency has largely focused on detection within conversational speech, particularly using the Switchboard corpus and its treebank annotations. Early approaches modeled repairs using noisy-channel and parsing-based methods that treated EDITED regions as structured syntactic objects [36, 37]. Subsequent work reframed disfluency detection as a sequence labeling problem, employing BiLSTM and Semi-CRF [38] and EGBC [39] and more. Recently, pretrained contextual encoders have been used to identify disfluent spans [40]. The primary metrics for disfluency removal are word-level precision ($\mathcal{E}_P$), recall ($\mathcal{E}_R$), and F1 scores ($\mathcal{E}_F$). Recent work introduces a method to perform alignment between generative model outputs and gold parse trees, enabling this scoring for backbone LLMs [41]. Previous systems [36, 37, 38, 39, 40] substantially improved token- and span-level F1, par-

LLM	*Citation*	*Open-Source*	*Instruct*	*Sizes*	*Architecture*	*Context Length*	*Features*
gpt-4o	[49]	✗	✓	Nx200B*	MoE*	128k	Multimodal, Instruct
gpt-4o-mini	[49]	✗	✓	Nx8B*	MoE*	128k	Multimodal, Instruct
o4-mini	[49]	✗	✓	100B	Dense	200k	Multimodal, Instruct, Reasoning
Llama-3.1	[50]	✓	✓	8B	Dense	128k	Instruct
Llama-3.2	[50]	✓	✓	1B,3B	Dense	128k	Instruct
Llama-3.3	[50]	✓	✓	70B	Dense	128k	Multimodal, Instruct
MobileLLM	[51]	✓	✗	125M, 350M, 600M, 1B	Dense	2048+	Small for edge devices
Qwen3	[52]	✓	✓	0.6B,1.7B, 4B, 8B	Dense + MoE	32,768+	Instruct
Phi-4-mini	[53]	✓	✓	3.8B	Dense	128k	Instruct, Reasoning

Table 1: **Backbone LLMs Evaluated**: * indicates rumored sizes [54, 55, 56]. o4-mini is available in high/medium reasoning variants, we extrapolate the size from rumored o1-mini sizes. While some include multimodal capabilities, they are primarily text-based language models.

ticularly for **INTJ** and **PRN** structures. A new standardized metric, $\mathcal{Z}$-Scores, measures performance on these structures at the span level [41].

2.3 Robustness in Large Language Models

Robustness in large language models has largely been studied through adversarial perturbations, distribution shift, and reasoning benchmarks [42, 43, 44, 45]. These evaluations test whether models maintain semantic consistency under lexical variation, domain transfer, or multi-step inference stress tests. Performance is typically measured using task accuracy or semantic similarity metrics that explicitly tolerate paraphrase and abstraction, such as BERTScore [46], BLEURT [47], and COMET [48].

Conversational speech poses a different challenge. Disfluency removal is a deletion-constrained task requiring the fluent output to preserve the original token sequence except for annotated disfluencies. Unlike semantic benchmarks that permit paraphrasing, this setting demands strict structural fidelity, where additional deletions or rewrites are errors. Robustness must therefore be measured by agreement with the transcript's structural constraints rather than semantic equivalence.

3 Robustness to Conversational Structure: DRES as a Factorized Structural Evaluation Framework

We formalize robustness to conversational structure as fidelity to a uniquely specified deletion-only transformation over conversational transcripts, as shown in **Figure 1**.

Let x denote a tokenized transcript containing disfluencies, and let $m \in \{0,1\}^n$ denote its gold disfluency mask. Because disfluency removal is deletion-constrained and uniquely determined under gold annotation, robustness can be defined in terms of agreement with this mask. DRES operationalizes this definition by isolating language-level deletion behavior from acoustic variability and evaluating agreement with m directly.

3.1 SpeechLLMs as Composed Systems

To operationalize structural robustness, we first characterize how deletion behavior arises within a SpeechLLM pipeline. A SpeechLLM composes an acoustic encoder A with a language backbone L_θ [25, 26, 27, 28]:

$$A : W \to X, \qquad L_\theta : X \to X,$$

yielding the end-to-end mapping

$$y = L_\theta(A(w)).$$

Deletion behavior therefore arises from two operators: acoustic suppression in A and structural editing in L_θ.

3.2 Acoustic Suppression and Editing Bias

This decomposition highlights a key measurement challenge: deletions in the final transcript may originate either from the acoustic encoder or from the language model's editing policy. Without separating these effects, structural failures cannot be attributed to the backbone model itself.

Modern ASR systems under-transcribe certain disfluencies, especially short interjections [32, 33]. We model acoustic omission as a deletion mask

$$S_A : x \mapsto \hat{x}, \qquad \hat{x} = (x_i : s_i = 0), \quad s \in \{0,1\}^n.$$

Observed end-to-end deletions reflect the composition

$$x \xrightarrow{S_A} \hat{x} \xrightarrow{D_\theta} y.$$

Without controlling S_A, deletions cannot be attributed to acoustic omission versus backbone editing. Even when $S_A = 0$ (perfect transcription), the backbone model may fall into one of the four editing policies; thus systematic over- or under-deletion arises from the backbone model D_θ itself, reflecting training objectives that favor abstraction or compression rather than structural fidelity. DRES fixes $S_A = 0$ and evaluates

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

thereby isolating backbone-level editing policies.

3.3 Robustness and Editing Policy Definitions

Having isolated backbone behavior, we can now define robustness directly in terms of the model's deletion decisions. Let x be a transcript, $m \in \{0,1\}^n$ its gold deletion mask, and $D_\theta(x) \in \{0,1\}^n$ the model-implied deletion mask recovered via alignment [41]. We define:

$$TP_\theta(x) := \sum_i \mathbb{1}[m_i = 1 \wedge D_{\theta,i}(x) = 1],$$

$$O_\theta(x) := \sum_i \mathbb{1}[m_i = 0 \wedge D_{\theta,i}(x) = 1],$$

$$U_\theta(x) := \sum_i \mathbb{1}[m_i = 1 \wedge D_{\theta,i}(x) = 0].$$

Here $TP_\theta(x)$ denotes true positive deletions (i.e., the token should be deleted in reference to the gold transcript and it was deleted by L_θ), $O_\theta(x)$ denotes over-deletions and $U_\theta(x)$ under-deletions. Word-level precision and recall ($\mathcal{E}$-Scores) can be written directly as

$$\mathcal{E}_P(x;\theta) = \frac{TP_\theta(x)}{TP_\theta(x) + O_\theta(x)}, \quad \mathcal{E}_R(x;\theta) = \frac{TP_\theta(x)}{TP_\theta(x) + U_\theta(x)}.$$

Definition 1: Robustness via Editing Policies. Thus robustness is defined in terms of $\mathcal{E}$-Scores as:

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

and it is fully determined by the pair $(O_\theta(x), U_\theta(x))$. This definition quantifies how robust the backbone model L_θ is to over-deletion and under-deletion.

Editing Policies. These quantities naturally induce a geometric interpretation of model behavior in precision–recall space. Different relative magnitudes of over- and under-deletions correspond to distinct editing regimes, which we refer to as editing policies. There are four editing policies in $(\mathcal{E}_P, \mathcal{E}_R)$-space:

Under-Deletion	$U_\theta(x) \gg O_\theta(x) \Rightarrow (\mathcal{E}_P \uparrow, \mathcal{E}_R \downarrow)$
Over-Deletion	$O_\theta(x) \gg U_\theta(x) \Rightarrow (\mathcal{E}_P \downarrow, \mathcal{E}_R \uparrow)$
Balanced	$O_\theta(x), U_\theta(x)$ both small $\Rightarrow (\mathcal{E}_P \uparrow, \mathcal{E}_R \uparrow)$
Poor	$O_\theta(x), U_\theta(x)$ both large $\Rightarrow (\mathcal{E}_P \downarrow, \mathcal{E}_R \downarrow)$

We show these policies in Figure 2. As discussed in §4.1, clustering modeling in this space recovers groups that align with these regions (Figure 3), indicating that editing policies emerge as natural geometric structure in robustness space.

Definition 2: Robustness via Disfluency Categories. We also define robustness in terms of **category-specific** $\mathcal{Z}$-Scores [41]:

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

Which quantify how robust the backbone model L_θ is to the three types of disfluencies in the Shriberg definition [6]: **EDITED**, **INTJ**, and **PRN**, and all other parts-of-speech are considered 'fluent,' as shown in Figure 1.

3.4 Switchboard Dataset

DRES is built on the Switchboard Treebank [57, 58], which provides gold parse trees with paired fluent and disfluent realizations. Disfluencies are defined using the Shriberg scheme [6]: **EDITED**, **INTJ**, and **PRN** are labeled disfluent; all other parts-of-speech are fluent [40, 36]. Fluent and disfluent transcripts are constructed recursively from gold trees [36], retaining partial words and punctuation to match modern ASR outputs [37]. Because Switchboard transcripts have been manually produced and, importantly, they have been iteratively corrected [59], they provide reliable structural targets. Gold transcripts are essential, as ASR systems systematically under-transcribe disfluencies [32, 33].

3.5 Evaluation Protocol

Deletion Alignment. Token-level deletion decisions are recovered by alignment, using the methodology of [41].

Context Conditions. Models are evaluated under: full transcripts, and segmented transcripts (approximately 4 sentences each). Segmentation shortens context while preserving local structure, enabling separation of intrinsic editing policy from long-context instability.

Fine-Tuning and Generalization. To measure specialization effects, models are evaluated before and after adaptation on GSM8K [60], MMLU [61], and CoQA [62].

In-Context Learning (k). We evaluate models under in-context learning [63], where the model is conditioned on a small number of task demonstrations provided directly in the prompt rather than through parameter updates. Example pairs are drawn from the development portion of the dataset and formatted as disfluent—fluent pairs, illustrating the deletion-only transformation the model should perform. The ordering of demonstrations is fixed across models to ensure consistent evaluation. Here k denotes the number of demonstration pairs included in the prompt prior to the evaluation instance. We evaluate models with $k \in \{0, 1, 3, 5\}$, shown in Figure 2.

Comparison to Previous SOTA. Prior work on conversational disfluency primarily treats the problem as supervised span detection using sequence labeling models such as BiLSTM and Semi-CRF [38], EGBC [39], weight-sharing approaches [64], and BERT-based parsers [40] (performance shown in Figure 2). In contrast, SpeechLLMs rely on generative LLM backbones that perform disfluency removal through open-ended editing rather than explicit span prediction. DRES evaluates these generative repairs under the same structural criteria used in prior work, enabling direct comparison between sequence-labeling and generative approaches.

3.6 Model Axes

We evaluate proprietary and open-source LLMs (Table 1) spanning: **model size** (125M to frontier scale); **architecture** (dense vs. mixture-of-experts); **instruction-tuned vs. base vs. reasoning** variants; **context window length**. This controlled design allows comparison of how scale, objective, architecture, and context management shape deletion-constrained repair.

4 Empirical Evidence of Policy-Level Robustness Patterns

Figure 2 plots model performance in the precision–recall (E_P, E_R) space under full and segmented transcript conditions.

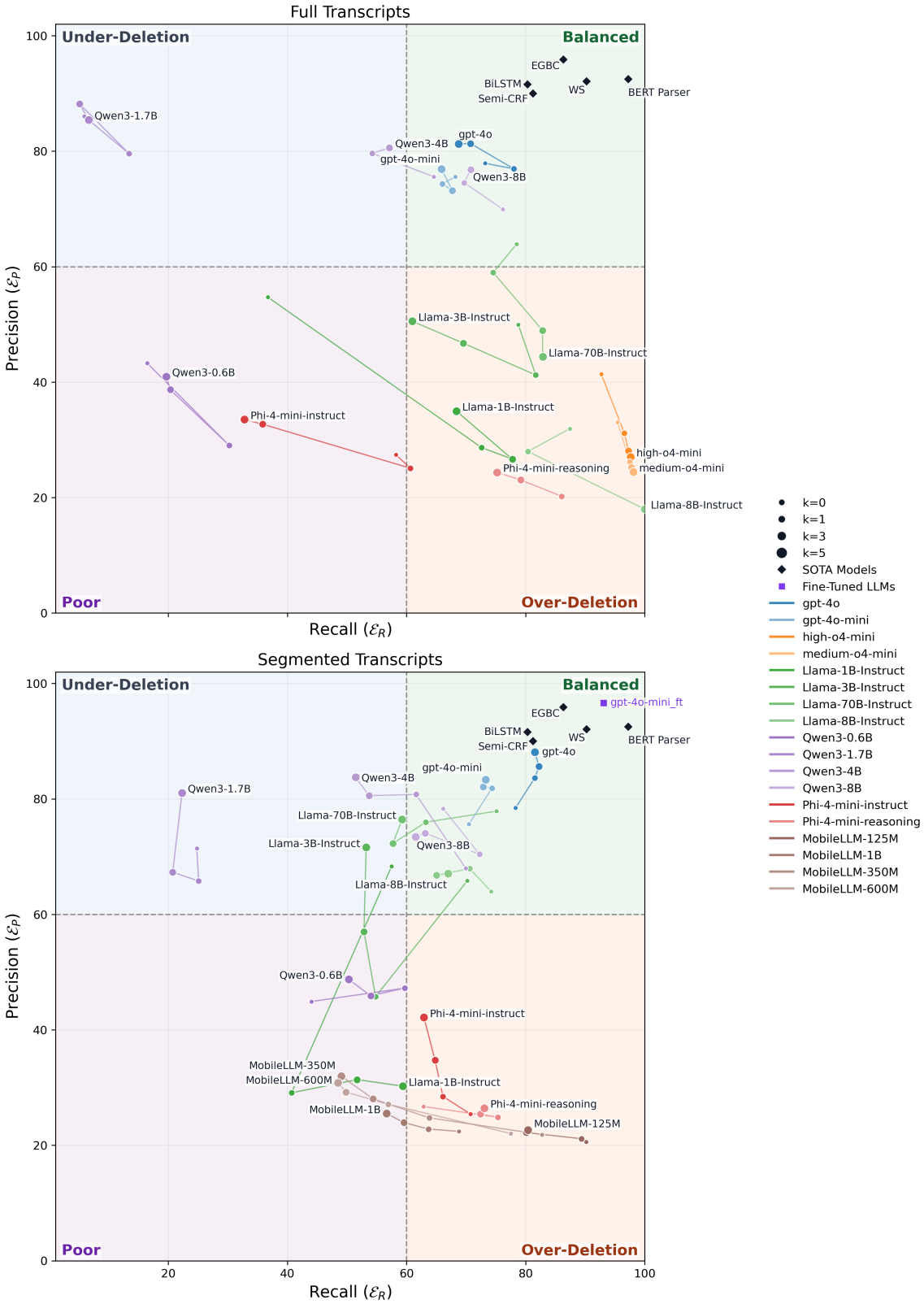

Figure 2: Precision–recall trade-offs across models under full (top) and segmented (bottom) transcripts. Each point corresponds to a model evaluated at varying in-context learning levels ($k = \{0, 1, 3, 5\}$). Shaded quadrants reveal four **editing policies**:
Under-Deletion ($\mathcal{E}_P \uparrow, \mathcal{E}_R \downarrow$) occurs when models fail to recognize conversational structure and leave many disfluencies untouched.
Over-Deletion ($\mathcal{E}_P \downarrow, \mathcal{E}_R \uparrow$) reflects a rewriting bias: models treat the task as paraphrasing and delete fluent words.
Balanced ($\mathcal{E}_P \uparrow, \mathcal{E}_R \uparrow$) represents the desired behavior, combining accurate disfluency identification with preservation of fluent content.
Poor ($\mathcal{E}_P \downarrow, \mathcal{E}_R \downarrow$) inherits the worst of both behaviors, missing disfluencies while also deleting fluent tokens.
Proprietary models cluster in the Balanced region, while reasoning models cluster in the Over-Deletion region; small models more frequently occupy the Over-Deletion or Poor regimes. Qualitative structure remains consistent for thresholds in the range 0.55–0.70; quantitative clustering analysis appears in §4 (▷ Findings 1, 3, 4, 5).

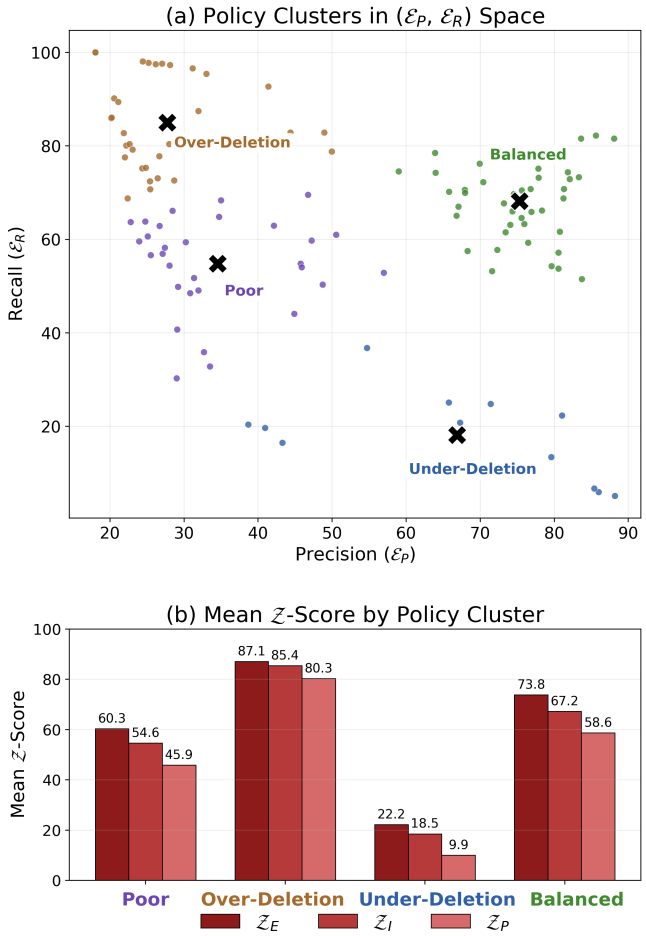

Figure 3: Policy structure in robustness space. (a) Clustering in $(\mathcal{E}_P, \mathcal{E}_R)$ space recovers groups that align with the quadrant regimes in Figure 2, indicating that **editing policies emerge as geometric structure in robustness space**. Black markers denote cluster centers (▷ Finding 1). (b) Mean category-level $\mathcal{Z}$ scores for each policy cluster ($\mathcal{Z}_E$, $\mathcal{Z}_I$, $\mathcal{Z}_P$) show consistent behavioral differences across disfluency types (▷ Finding 2).

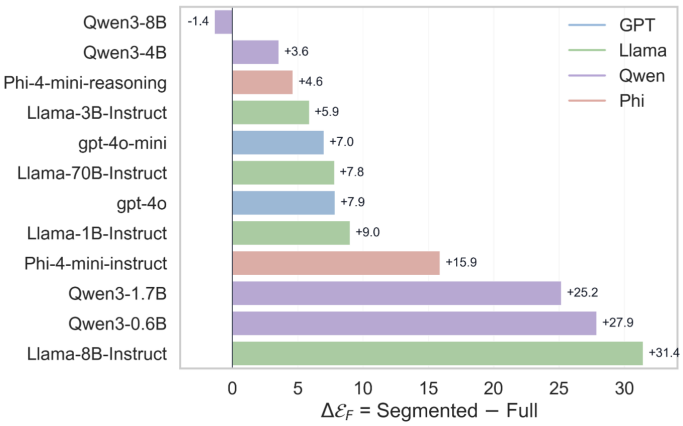

Figure 4: **Segmentation improves robustness to conversational structure**. Segmenting long conversational transcripts consistently improves structural fidelity across models. Performance gains in $\Delta\mathcal{E}_F$ indicate that robustness failures largely arise from long-context instability rather than knowledge limitations. Averaged across k (▷ Finding 3).

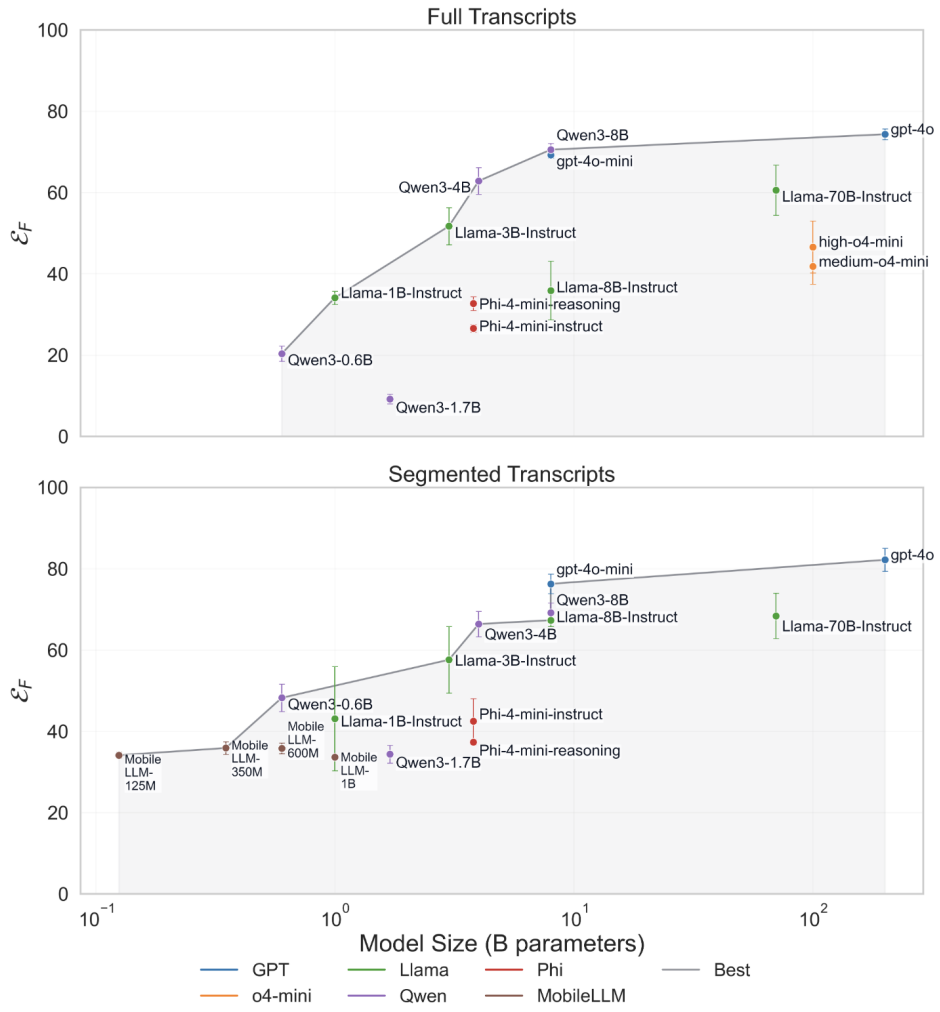

Figure 5: **Scale improves performance, but does not change editing policy.** Larger models within each family achieve higher $\mathcal{E}_F$ but reasoning-oriented variants consistently underperform relative to the best performance curve. This shows that scale improves execution of a policy but does not change the underlying editing behavior.(▷ Finding 4)

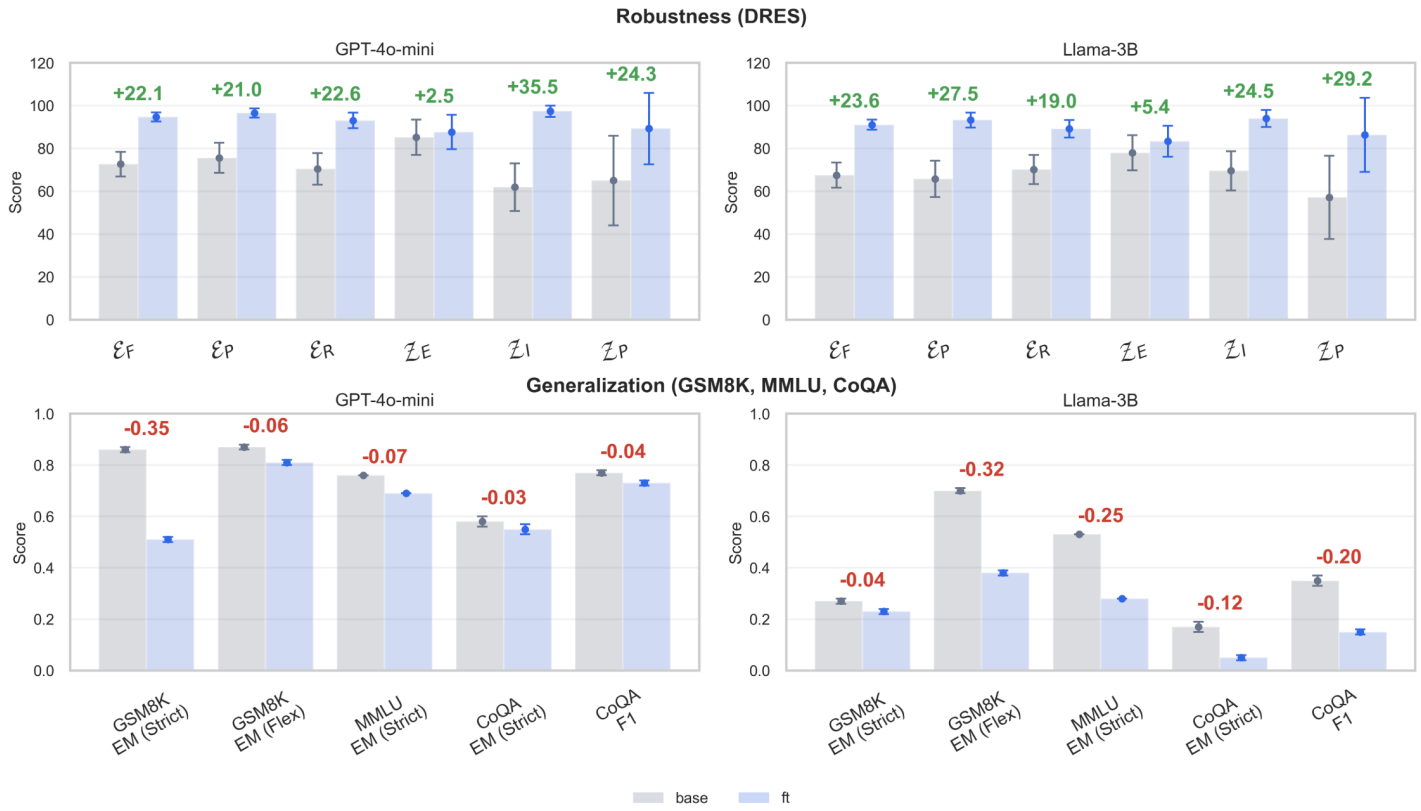

Figure 6: **Fine-Tuning Results** (▷ Finding 5).

4.1 Quantitative Validation of Editing Policies

To verify that the *editing policies* described in §3.3 (**Under-Deletion**, **Over-Deletion**, **Balanced**, and **Poor**) reflect intrinsic structure, we analyze model behavior in the (E_P, E_R) space defined in §3.3. We perform a k-means clustering analysis in the (E_P, E_R) plane. To select the number of clusters, we compute the silhouette score (S) across $k \in \{2, \ldots, 7\}$ (Figure 3). The score peaks at $k = 4$ and declines for larger k. The silhouette score ($S = 0.54$) and Davies-Bouldin Index ($DBI = 0.664$) indicate moderate-to-strong cluster separation and internal cohesion, confirming that four clusters provide the best separation in the (E_P, E_R) space.

A permutation test with 200 randomizations produced no silhouette scores exceeding the observed value ($p \approx 0.005$), indicating that the cluster structure is unlikely under a random (E_P, E_R) pairing. This indicates that the editing polices indeed correspond to natural groupings in this space.

We find that cluster assignments are largely stable within model families. Across prompting levels, models retain their policy classification in 70–100% of cases, indicating that editing behavior is a backbone-intrinsic property. However, stability differs based on whether we compare performance over full transcripts or segmented transcripts: while GPT models exhibit perfect policy stability across both settings ($p = 1.00$), the Llama family shows significant volatility ($p = 0.25$), with Phi ($p = 0.75$) and Qwen ($p = 0.79$) maintaining moderate consistency in their cluster assignments across both settings.

4.2 ▷ Finding 1: Editing policies align with training objectives (Figures 2, 3).

Models consistently occupy distinct precision–recall regimes corresponding to conservative, aggressive, balanced, and poor editing behavior. GPT models cluster in the balanced region across prompting conditions, while reasoning-oriented models (e.g., o4-mini and Phi-4 reasoning variants) systematically shift toward aggressive over-deletion. Smaller or base models frequently adopt conservative under-deletion policies, preserving fluent tokens while failing to remove many disfluencies.

These patterns remain stable across prompting levels and scale within families, indicating that editing behavior is primarily determined by training objectives rather than parameter count.

Model	ΔDRES (95% CI)	ΔMMLU (95% CI)	ΔGSM8K (95% CI)	ΔCoQA (95% CI)
GPT-4o-mini	+22.05 [+9.83, +34.10]	−0.07 [−0.07, −0.07]	−0.35 [−0.38, −0.32]	−0.03 [−0.09, +0.02]
Llama-3B	+23.46 [+10.90, +35.91]	−0.25 [−0.25, −0.25]	−0.04 [−0.07, −0.01]	−0.12 [−0.16, −0.08]

Table 2: Paired Base → Fine-Tuned Performance Changes. ΔDRES reflects improvement in structural fidelity (segmented condition, $\mathcal{E}_F$). Negative Δ values indicate degradation on generalization benchmarks. Confidence intervals are 95%, (▷ Finding 5).

4.3 ▷ Finding 2: Models handle overt repairs well (EDITED) but struggle with short conversational markers (INTJ, PRN) (Figure 3b).

Category-level $\mathcal{Z}$-scores reveal systematic differences across disfluency types (Figure 3b). The models perform well on **EDITED** structures, commonly studied in prior disfluency detection work [40, 37].

In contrast, performance drops substantially on **INTJ** (interjections such as "uh" and "um") and **PRN** (parenthetical insertions such as "you know" and "I mean"). These markers are short, frequent, and embedded within otherwise fluent structure. Notably, this result contrasts with prior work suggesting that these categories are among the easiest to detect and remove [40, 37]. This discrepancy indicates that generative models encounter *different failure modes* than traditional sequence-tagging approaches (§2.2).

The pattern largely persists across model families and scales, suggesting a training distribution mismatch: conversational markers common in spontaneous speech appear underrepresented in pretraining corpora.

4.4 ▷ Finding 3: Long transcripts expose context instability rather than knowledge gaps (Figures 2, 4).

Performance differences between full transcripts and segmented inputs reveal a context-management failure mode.[1] Under full transcripts, several models exhibit unstable precision–recall behavior, frequently collapsing into **over-deletion** regimes. Segmenting transcripts into shorter contexts consistently shifts models toward the **balanced** region and reduces variance in performance. This improvement occurs even for models with large context windows, indicating that instability arises from context handling rather than insufficient capacity. Thus, robustness failures in conversational speech partly reflect architectural sensitivity to long-context structure.

4.5 ▷ Finding 4: Scale generally improves performance within model families but does not change the editing policy (Figures 2, 5).

Structural performance generally improves with model size (Figures 2, 5). Within each model family, larger variants consistently achieve higher $\mathcal{E}_F$, forming an improvement curve as parameter count increases.

However, *scale does not alter the underlying editing policy.* Models tend to remain within the same precision–recall policy as they grow. Conservative models become more precise but continue to **under-delete** disfluencies, while aggressive models improve recall yet still **over-delete** fluent content.

This pattern indicates that scale primarily refines performance along a family-specific trajectory rather than changing the qualitative editing behavior itself. The editing policy therefore appears to be determined by training objectives and alignment choices, while scale controls how well the model executes that policy.

4.6 ▷ Finding 5: Fine-tuning introduces a robustness–generalization trade-off, improving structural fidelity but reducing generalization (Figures 2, 6, Table 2).

Fine-tuning substantially improves structural fidelity on the disfluency removal task. As shown in Figures 2 and 6, both models move close to the upper bound of the DRES metrics after adaptation, with $\mathcal{E}_F$ rising from the mid–70s to the mid–90s for GPT-4o-mini and from the high–60s to above 90 for Llama-3B. Table 2 summarizes these gains as improvements of more than **+22 points** in overall DRES performance for both models.

[1] This pattern is consistent with prior observations that long-context language models often handle information appearing in the middle of a sequence less reliably than information near its boundaries [65].

This improvement comes at a cost. Across GSM8K, MMLU, and CoQA, both models exhibit lower post–fine-tuning scores (Figure 6), with the largest decline appearing on *reasoning-heavy tasks* such as GSM8K. Recent studies have also shown this same phenomenon: fine-tuning LLMs on domain-specific datasets can substantially impair their generalization capabilities [66, 67, 68, 69].

Together, these results reveal a robustness–generalization trade-off: fine-tuning aligns models with deletion-constrained structural repair, but this specialization narrows their broader reasoning and knowledge capabilities.

Category		Guideline for Deployment and Development
Deployment	R1	Use small, **under-deletion** models (e.g., Qwen3-1.7B) for edge-based, real-time intent preservation
	R2	Prioritize segmented input over full transcripts to ensure stability, regardless of context window size
	R3	Leverage 8B-class models with few-shot prompting for a reasonable latency-accuracy trade-off
Objective	R4	Avoid reasoning-oriented models for literal repair to prevent **over-deletion**
	R5	In command-driven interfaces, select models with a **under-deletion** policy to protect fluent nouns/verbs
	R6	Account for stable family-wide **editing policies** when scaling; larger models retain similar error biases
Training	R7	Use domain-specific fine-tuning for archival-grade transcription where SOTA precision is required
	R8	Monitor the generalization tax; evaluate MMLU/GSM8K scores when fine-tuning for speech tasks
	R9	Utilize DRES as a diagnostic structural probe to audit new LLMs before integration into SpeechLLMs

Table 3: Practical Recommendations (R1–R9) for Conversational Robustness in Speech Pipelines

4.7 Practical Recommendations (Table 3)

The empirical results from our DRES evaluation translate into several deployment guidelines for SpeechLLM pipelines (Table 3). Segmenting transcripts improves robustness even for long-context models, while small high-precision models suit edge settings and mid-scale models with light prompting provide strong latency–accuracy trade-offs (R1–R3). Editing behavior is driven primarily by training objectives rather than scale – reasoning-oriented models often over-delete fluent content—while fine-tuning improves structural fidelity but introduces a measurable generalization cost, motivating structural diagnostics such as DRES before deployment (R4–R9).

5 Limitations

This study focuses on structural identifiability rather than full end-to-end deployment behavior. DRES fixes acoustic suppression and measures deletion decisions against gold masks, enabling structural editing policies to be analyzed independently of ASR errors, though real systems combine both effects. Experiments are conducted on English Switchboard transcripts; while the deletion-constrained framework is language-agnostic, conversational repair patterns may vary across languages and remain to be validated. Finally, the deletion-only task captures minimal structural fidelity but does not model broader normalization behaviors.

6 Future Directions

Beyond typical conversational disfluency, clinical speech conditions such as aphasia, dysarthria, and other language production disorders introduce different structural phenomena [70, 71]. Future work should develop condition-

aware structural auditing frameworks and incorporate domain-specific safeguards when deploying SpeechLLMs in healthcare environments.

7 Conclusion

With DRES, we show that conversational speech reveals systematic, objective-induced limits in current LLMs. Robustness to conversational structure does not improve monotonically with scale. Instead, models adopt stable editing biases shaped by training objectives and context handling. Deletion-only structural tasks expose these tendencies with minimal ambiguity. Aggressive semantic abstraction conflicts with constrained repair; conservative preservation fails to remove true disfluencies. Fine-tuning can improve local structural fidelity, but at measurable cost to broader generalization. Structural diagnostics such as DRES therefore provide a complementary evaluation axis for SpeechLLM development.

8 Generative AI Use Disclosure

AI tools supported parts of the coding process and helped structure and clarify written content. All generated material was carefully reviewed, validated, and revised by the authors.

References

[1] Z. Ma, Z. Chen, Y. Wang, E. S. Chng, and X. Chen, "Audio-cot: Exploring chain-of-thought reasoning in large audio language model," 2025. [Online]. Available: https://arxiv.org/abs/2501.07246

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[3] S. Hu, L. Zhou, S. Liu, S. Chen, L. Meng, H. Hao, J. Pan, X. Liu, J. Li, S. Sivasankaran, L. Liu, and F. Wei, "Wavllm: Towards robust and adaptive speech large language model," 2024. [Online]. Available: https://arxiv.org/abs/2404.00656

[4] J. Fan, R. Ren, J. Li, R. Pandey, P. G. Shivakumar, Y. Gu, A. Gandhe, G. Liu, and I. Bulyko, "Incentivizing consistent, effective and scalable reasoning capability in audio LLMs via reasoning process rewards," in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=DUr48hxO2h

[5] A. L. Alter, "The benefits of cognitive disfluency," *Current Directions in Psychological Science*, vol. 22, no. 6, pp. 437–442, 2013.

[6] E. Shriberg, "Preliminaries to a Theory of Speech Disfluencies," Ph.D. dissertation, University of California at Berkeley, 1994.

[7] ——, "Disfluencies in switchboard," in *International Conference on Spoken Language Processing*, vol. 96, 1996, pp. 11–14.

[8] H. Bortfeld, S. D. Leon, J. E. Bloom, M. F. Schober, and S. E. Brennan, "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender," *Language and speech*, vol. 44, no. 2, pp. 123–147, 2001.

[9] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, p. 73–111, 2002.

[10] M. Meteer *et al.*, "Dysfluency annotation stylebook for the savitchboard corpus," Technical report, Tech. Rep., 1995.

[11] E. Diachek and S. Brown-Schmidt, "The effect of disfluency on memory for what was said." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 49, no. 8, p. 1306, 2023.

[12] S. E. Brennan and M. Williams, "The feeling of another s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers," *Journal of memory and language*, vol. 34, no. 3, pp. 383–398, 1995.

[13] A. Kirkland, H. Lameris, E. Székely, and J. Gustafson, "Where's the uh, hesitation? the interplay between filled pause location, speech rate and fundamental frequency in perception of confidence." in *INTERSPEECH*, 2022, pp. 4990–4994.

[14] T. Dinkar, C. Clavel, and I. Vasilescu, "Fillers in spoken language understanding: Computational and psycholinguistic perspectives," 2023. [Online]. Available: https://arxiv.org/abs/2301.10761

[15] N. Torstenston and B. Gawrońska, "Discourse disfluencies in bilingual court hearings," 2009.

[16] L. Harrington, R. Rhodes, and V. Hughes, "Style variability in disfluency analysis for forensic speaker comparison," *The International Journal of Speech, Language and the Law*, vol. 28, no. 1, pp. 31–58, 2021.

[17] F. Schiel and C. Heinrich, "Disfluencies in the speech of intoxicated speakers," *The International Journal of Speech, Language and the Law*, vol. 22, no. 1, pp. 19–33, 2015.

[18] M. Larson, M. A. Britt, and A. A. Larson, "Disfluencies in comprehending argumentative texts," *Reading Psychology*, vol. 25, no. 3, pp. 205–224, 2004.

[19] I. Hernandez and J. L. Preston, "Disfluency disrupts the confirmation bias," *Journal of Experimental Social Psychology*, vol. 49, no. 1, pp. 178–182, 2013.

[20] A. Koenecke, A. S. G. Choi, K. X. Mei, H. Schellmann, and M. Sloane, "Careless whisper: Speech-to-text hallucination harms," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1672–1681. [Online]. Available: https://doi.org/10.1145/3630106.3658996

[21] J. E. Loy, H. Rohde, and M. Corley, "Real-time social reasoning: the effect of disfluency on the meaning of some," *Journal of Cultural Cognitive Science*, vol. 3, no. 2, pp. 159–173, 2019.

[22] M. Wester, M. Aylett, M. Tomalin, and R. Dall, "Artificial personality and disfluency," in *INTERSPEECH 2015 16th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2015, pp. 3365–3369.

[23] B. De Keersmaecker, R. J. Hartsuiker, and A. Pistono, "(don't) believe me, i'm telling the truth! speech disfluency and eye contact as cues to veracity, intention, and truth judgement," *Language, Cognition and Neuroscience*, vol. 39, no. 10, pp. 1263–1277, 2024.

[24] J. Loy, H. Rohde, and M. Corley, "Lying, in a manner of speaking," in *Proc. SpeechProsody 2016*, 2016, pp. 984–988.

[25] Z. Yang, S. Shimizu, Y. Yu, and C. Chu, "When large language models meet speech: A survey on integration approaches," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 20 298–20 315. [Online]. Available: https://aclanthology.org/2025.findings-acl.1041/

[26] W. Cui, D. Yu, X. Jiao, Z. Meng, G. Zhang, Q. Wang, S. Y. Guo, and I. King, "Recent advances in speech language models: A survey," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 13 943–13 970.

[27] S. Arora, K.-W. Chang, C.-M. Chien, Y. Peng, H. Wu, Y. Adi, E. Dupoux, H.-Y. Lee, K. Livescu, and S. Watanabe, "On the landscape of spoken language models: A comprehensive survey," *arXiv preprint arXiv:2504.08528*, 2025.

[28] J. Peng, Y. Wang, B. Li, Y. Guo, H. Wang, Y. Fang, Y. Xi, H. Li, X. Li, K. Zhang, S. Wang, and K. Yu, "A survey on speech large language models for understanding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 20, no. 1, p. 2–31, Jan. 2026. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2025.3640535

[29] H. Liu, Y. Hou, H. Liu, Y. Wang, Y. Wang, and Y. Wang, "Vocalbench-df: A benchmark for evaluating speech llm robustness to disfluency," 2025. [Online]. Available: https://arxiv.org/abs/2510.15406

[30] L. Zhang, J. Zhang, B. Lei, C. Wu, A. Liu, W. Jia, and X. Zhou, "Wildspeech-bench: Benchmarking end-to-end speechllms in the wild," 2025. [Online]. Available: https://arxiv.org/abs/2506.21875

[31] W. Cui, X. Jiao, Z. Meng, and I. King, "Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models," 2025. [Online]. Available: https://arxiv.org/abs/2501.04962

[32] D. Mujtaba, N. R. Mahapatra *et al.*, "Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech," *arXiv preprint arXiv:2405.06150*, 2024.

[33] M. Teleki, X. Dong, S. Kim, and J. Caverlee, "Comparing asr systems in the context of speech disfluencies," in *Interspeech 2024*, 2024, pp. 4548–4552.

[34] F. Retkowski, M. Züfle, A. Sudmann, D. Pfau, S. Watanabe, J. Niehues, and A. Waibel, "Summarizing speech: A comprehensive survey," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 27 275–27 306. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1388/

[35] A. Koenecke, J.-J. Nunez, and I. Y. Chen, "Perspective: Listening to users when auditing medical ai scribes."

[36] E. Charniak and M. Johnson, "Edit Detection and Parsing for Transcribed Speech," *NAACL*, 2001.

[37] M. Johnson and E. Charniak, "A TAG-based noisy-channel model of speech repairs," in *ACL*, 2004, pp. 33–39. [Online]. Available: https://aclanthology.org/P04-1005/

[38] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency detection using a bidirectional LSTM," in *INTERSPEECH*, 2016.

[39] N. Bach and F. Huang, "Noisy bilstm-based models for disfluency detection." in *INTERSPEECH*, 2019, pp. 4230–4234.

[40] P. Jamshid Lou and M. Johnson, "Improving disfluency detection by self-training a self-attentive model," in *ACL*, 2020, pp. 3754–3763. [Online]. Available: https://aclanthology.org/2020.acl-main.346

[41] M. Teleki, S. Janjur, H. Liu, O. Grabner, K. Verma, T. Docog, X. Dong, L. Shi, C. Wang, S. Birkelbach, J. Kim, Y. Zhang, and J. Caverlee, "Z-scores: A metric for linguistically assessing disfluency removal," in *ICASSP*, 2025.

[42] X. Xu, K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli, "An LLM can fool itself: A prompt-based adversarial attack," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=VVgGbB9TNV

[43] K. Zhang, L. Wu, K. Yu, G. Lv, and D. Zhang, "Evaluating and improving robustness in large language models: a survey and future directions," *arXiv preprint arXiv:2506.11111*, 2025.

[44] P.-N. Kung, F. Yin, D. Wu, K.-W. Chang, and N. Peng, "Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 1813–1829.

[45] A. Agrawal, L. Alazraki, S. Honarvar, and M. Rei, "Enhancing llm robustness to perturbed instructions: An empirical study," 2025. [Online]. Available: https://arxiv.org/abs/2504.02733

[46] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[47] T. Sellam, D. Das, and A. Parikh, "Bleurt: Learning robust metrics for text generation," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 7881–7892.

[48] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "Comet: A neural framework for mt evaluation," in *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, 2020, pp. 2685–2702.

[49] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[50] H. Touvron, T. Sellam, Y. Jernite, J. Copet, L. B. Allal *et al.*, "The llama 3 herd of models," *Computing Research Repository*, vol. arXiv:2407.21783, 2024.

[51] Z. Liu, C. Zhao *et al.*, "MobileLLM: Optimizing sub-billion parameter language models for on-device use cases," 2024. [Online]. Available: https://arxiv.org/abs/2402.14905

[52] Q. Team, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388

[53] Microsoft, A. Abouelenin, A. Ashfaq, A. Atkinson *et al.*, "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," 2025. [Online]. Available: https://arxiv.org/abs/2503.01743

[54] M. Zeff, "Openai unveils gpt-4o mini, a smaller and cheaper ai model," *TechCrunch*, Jul 18 2024. [Online]. Available: https://techcrunch.com/2024/07/18/openai-unveils-gpt-4o-mini-a-small-ai-model-powering-chatgpt

[55] E. Erdil, "Frontier language models have become much smaller," 2024. [Online]. Available: https://epoch.ai/gradient-updates/frontier-language-models-have-become-much-smaller

[56] A. B. Abacha, W. wai Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, and T. Lin, "Medec: A benchmark for medical error detection and correction in clinical notes," 2025. [Online]. Available: https://arxiv.org/abs/2412.19260

[57] M. Mitchell, B. Santorini, M. Marcinkiewicz, and A. Taylor, "Treebank-3 LDC99T42 Web Download," p. 2, 1999, [Online]. Available:https://catalog.ldc.upenn.edu/LDC99T42.

[58] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *ICASSP*, vol. 1, 1992, pp. 517–520.

[59] R. J. Lickley and E. G. Bard, "On not recognizing disfluencies in dialogue," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3. IEEE, 1996, pp. 1876–1879.

[60] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[61] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[62] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[63] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[64] F. Wang, W. Chen, Z. Yang, Q. Dong, S. Xu, and B. Xu, "Semi-supervised disfluency detection," in *International Conference on Computational Linguistics*, 2018, pp. 3529–3538.

[65] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the association for computational linguistics*, vol. 12, pp. 157–173, 2024.

[66] T. Chu, Y. Zhai, J. Yang, S. Tong, S. Xie, D. Schuurmans, Q. V. Le, S. Levine, and Y. Ma, "Sft memorizes, rl generalizes: A comparative study of foundation model post-training," *arXiv preprint arXiv:2501.17161*, 2025.

[67] H. Chen, H. Tu, F. Wang, H. Liu, X. Tang, X. Du, Y. Zhou, and C. Xie, "Sft or rl? an early investigation into training r1-like reasoning large vision-language models," *arXiv preprint arXiv:2504.11468*, 2025.

[68] M. Huan, Y. Li, T. Zheng, X. Xu, S. Kim, M. Du, R. Poovendran, G. Neubig, and X. Yue, "Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning," *arXiv preprint arXiv:2507.00432*, 2025.

[69] J. Lin, Z. Wang, K. Qian, T. Wang, A. Srinivasan, H. Zeng, R. Jiao, X. Zhou, J. Gesi, D. Wang *et al.*, "Sft doesn't always hurt general capabilities: Revisiting domain-specific fine-tuning in llms," *arXiv preprint arXiv:2509.20758*, 2025.

[70] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[71] K. X. Mei, A. S. G. Choi, H. Schellmann, M. Sloane, and A. Koenecke, "Addressing pitfalls in auditing practices of automatic speech recognition technologies: A case study of people with aphasia," 2025. [Online]. Available: https://arxiv.org/abs/2506.08846