DRES: BENCHMARKING LLMS FOR DISFLUENCY REMOVAL

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

Texas A&M University

ABSTRACT

Disfluencies – such as "um," "uh," interjections, parentheticals, and edited statements – remain a persistent challenge for speech-driven systems, degrading accuracy in command interpretation, summarization, and conversational agents. We introduce DRES (Disfluency Removal Evaluation Suite), a controlled text-level benchmark that establishes a reproducible semantic upper bound for this task. DRES builds on human-annotated Switchboard transcripts, isolating disfluency removal from ASR errors and acoustic variability. We systematically evaluate proprietary and open-source LLMs across scales, prompting strategies, and architectures. Our results reveal that (i) simple segmentation consistently improves performance, even for long-context models; (ii) reasoning-oriented models tend to over-delete fluent tokens; and (iii) fine-tuning achieves near state-of-the-art precision and recall but harms generalization abilities. We further present a set of LLM-specific error modes and offer nine practical recommendations (R1-R9) for deploying disfluency removal in speech-driven pipelines. DRES provides a reproducible, model-agnostic foundation for advancing robust spoken-language systems.

***Index Terms*—** Disfluency removal, LLMs, benchmark, Switchboard, speech applications

1. INTRODUCTION

Disfluencies – *um, uh, interjections, parentheticals, and edited statements* – are pervasive in conversational speech, yet absent from most written text [1, 2, 3]. As speech-driven interfaces proliferate (e.g., Siri, Alexa, ChatGPT and Gemini voice modes, smart speakers, and emerging smart glasses) robust handling of disfluencies has become essential.

However, large language models (LLMs) trained primarily on written text often degrade in performance when processing spoken input with disfluencies. Prior studies show drops in voice command accuracy [4], conversational recommendation quality [5], and summarization fidelity [6, 7]. This mismatch is further amplified by ASR pipelines: while modern systems such as Whisper achieve strong recognition, they frequently omit disfluencies [8], preventing end-to-end models from being trained on realistic distributions. Consequently, explicit benchmarking of disfluency handling remains critical.

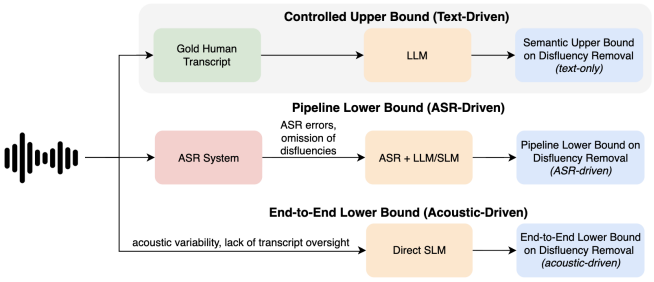

Fig. 1. *Speech pipelines introduce ASR and acoustic errors that mask true disfluency removal ability (practical lower bounds). In contrast,* ***controlled evaluation on gold transcripts provides a semantic upper bound****, isolating disfluency removal performance without recognition confounds.*

In this work, we deliberately focus on LLMs rather than speech language models (SLMs). LLMs provide a controlled platform for isolating the disfluency removal task, establishing a semantic upper bound on performance that is not confounded by acoustic variability or ASR errors as shown in Figure 1. Their flexible scaling – from edge-ready small models to very large proprietary systems – makes them particularly suitable for systematic comparison across architectures, scales, and prompting conditions. While the framework naturally extends to SLMs in future audio-based or hybrid evaluations, LLMs allow us to first define a reproducible baseline and clarify the limits of disfluency removal.

In this paper, we introduce **DRES: the Disfluency Removal Evaluation Suite**, the first large-scale, systematic benchmark of LLMs for this task. DRES builds on the Switchboard corpus, enabling controlled and reproducible evaluation. We make the following contributions:

- First, we introduce **DRES, a reproducible benchmark for disfluency removal based on Switchboard**, enabling systematic comparison of models across scales and modalities. We make the code and additional analysis available at `https://github.com/mariateleki/dres`.
- Second, we conduct a **comprehensive evaluation of open-source and proprietary LLMs**, including instruction-

LLM	Citation	*Open-Source*	*Instruct*	*Sizes*	*Architecture*	*Context Length*	*Features*
`gpt-4o`	[9]	✗	✓	Nx200B*	MoE*	128k	Multimodal, Instruct
`gpt-4o-mini`	[9]	✗	✓	Nx8B*	MoE*	128k	Multimodal, Instruct
`o4-mini`	[9]	✗	✓	Not disclosed.	MoE*	200k	Multimodal, Instruct, Reasoning
`Llama-3.1`	[10]	✓	✓	8B	Dense	128k	Instruct
`Llama-3.2`	[10]	✓	✓	1B,3B	Dense	128k	Instruct
`Llama-3.3`	[10]	✓	✓	70B	Dense	128k	Multimodal, Instruct
`MobileLLM`	[11]	✓	✗	125M, 350M, 600M, 1B	Dense	2048+	Small for edge devices
`Qwen3`	[12]	✓	✓	0.6B,1.7B, 4B, 8B	Dense + MoE	32,768+	Instruct
`Phi-4-mini`	[13]	✓	✓	3.8B	Dense	128k	Instruct, Reasoning

Table 1. **LLMs**: * indicates rumored sizes. o4-mini is available in high/medium reasoning variants. While some models include multimodal features, they are primarily text-based models (e.g., gpt models use a pipeline approach with Whisper).

tuned versus base, dense versus MoE, and small to very large models, revealing new trends in context handling, segmentation, and few-shot learning.

- Third, we provide the **first taxonomy of LLM-specific failure modes in disfluency removal**, such as over-deletion collapse, under-deletion, and reasoning-driven misinterpretation.
- Finally, we offer **practical recommendations for deployment (R1-R9) by connecting benchmark results to downstream tasks** where disfluencies are known to impair performance, including voice commands, conversational recommendation, and summarization, thereby providing actionable guidance for practitioners.

2. RELATION TO PRIOR WORK

2.1. Switchboard Dataset

We define the dataset T as:

$$T = \left\{(t_{tree}, t_{fluent}, t_{tag}, t_{disfluent})_i\right\}_{i=1}^{N}$$

where each element is a 4-tuple consisting of a fluent and disfluent version of the same utterance. We construct T from the Switchboard dataset [14, 15] following [16], where t_{fluent}, $t_{\text{disfluent}}$ are constructed from the tree, t_{tree}, using a top-down recursive approach. However, we do not remove partial words and punctuation, since ASR systems have since improved [17]. **Disfluency Definition.** We adopt the Shriberg [18] annotation scheme, where INTJ, PRN, and EDITED nodes are marked as disfluent [19, 16]; all others are fluent. **Syntactic Ambiguity** English syntax is often ambiguous, with sentences admitting multiple valid parse trees. In Switchboard, human annotators resolve this ambiguity using domain and discourse context, ensuring accurate identification of disfluencies [14]. As a result, Switchboard parse trees are more reliable than model-generated ones, particularly in disfluent contexts. **Correctly Transcribed Conversations.** Another important aspect of the Switchboard dataset [14] is that these parse trees are built on correctly-transcribed conversations [18]. State-of-the-art ASR models are known to significantly under-transcribe disfluencies [20] – even simple disfluencies, such as *uh* and *um*. Hence, it is not possible to automatically obtain transcriptions of this quality.

Previous Models	$\mathcal{E}_F$	$\mathcal{E}_P$	$\mathcal{E}_R$
BERT Parser [19]	**94.8**	92.5	**97.2**
EGBC [21]	90.9	**95.9**	86.3
Noisy BiLSTM [21]	92.2	94.7	89.8
Weight Sharing [22]	91.1	92.1	90.2
BiLSTM [23]	85.9	91.6	80.3
Semi-CRF [23]	85.4	90.0	81.2

Table 2. **Related Work**: Previous models evaluated in terms of word-based precision, recall, and F-score.

2.2. LLMs for the Disfluency Removal Task

We evaluate a broad set of LLMs (Table 1) that vary along several key dimensions. We compare *open-source vs. proprietary* models, *instruct vs. base* variants, and three *size categories* (small $\leq$1B parameters for edge use, mid-sized 1B–8B, and large $\geq$8B). We also study different *architectures* (dense vs. mixture-of-experts), *context lengths* (full vs. segmented transcripts to address long-context limitations), and special *features* such as multimodality and reasoning.

While specialized disfluency removal methods have been shown in Table 2 to perform well, LLMs offer a few key potential advantages for the disfluency removal task looking forward: (1) LLMs can leverage internet-scale pretraining for richer semantic understanding of domain terms, and (2) LLMs can use continual learning to track evolving language, i.e., slang, new terms, and shifting usage patterns.

2.3. Evaluation

All evaluations are performed on gold-standard Switchboard transcripts, which explicitly mark disfluent segments. This design isolates disfluency handling from ASR errors and ensures a controlled, reproducible benchmark. Model outputs are aligned with $t_{disfluent}$ using a variation of Gestalt pattern matching. This text-level setup provides a clean **upper bound for disfluency removal performance** and is complementary to audio-based SLM evaluation, which necessarily conflates disfluency handling with ASR quality.

$\mathcal{E}$-Scores. We report word-level precision ($\mathcal{E}_P$), recall ($\mathcal{E}_R$), and F1 ($\mathcal{E}_F$), following prior work on disfluency detec-

M	k	$\mathcal{E}_F$	$\mathcal{E}_P$	$\mathcal{E}_R$	$\mathcal{Z}_E$	$\mathcal{Z}_I$	$\mathcal{Z}_P$	$\mathcal{E}_F$	$\mathcal{E}_P$	$\mathcal{E}_R$	$\mathcal{Z}_E$	$\mathcal{Z}_I$	$\mathcal{Z}_P$
		gpt-4o						**gpt-4o-mini**					
f	0	$74.19_{7.87}$	$77.91_{10.63}$	$73.18_{12.98}$	$83.86_{13.15}$	$71.25_{16.57}$	$52.52_{31.91}$	$70.52_{9.22}$	$75.56_{9.89}$	$68.17_{13.90}$	$86.13_{10.30}$	$60.38_{18.22}$	$55.26_{26.17}$
f	1	$76.13_{7.63}$	$76.94_{10.48}$	$78.02_{13.05}$	$85.11_{11.73}$	$77.37_{15.64}$	$62.71_{29.59}$	$68.85_{9.53}$	$74.35_{9.65}$	$66.02_{14.04}$	$86.05_{10.37}$	$57.62_{19.88}$	$52.04_{24.06}$
f	3	$73.99_{9.33}$	$81.33_{10.10}$	$70.75_{15.11}$	$77.83_{17.30}$	$72.93_{15.26}$	$49.35_{29.73}$	$68.70_{10.64}$	$73.20_{11.59}$	$67.68_{15.21}$	$86.64_{10.10}$	$61.42_{20.83}$	$51.58_{25.42}$
f	5	$73.06_{8.04}$	$81.24_{10.69}$	$68.75_{13.39}$	$76.92_{15.55}$	$71.49_{13.97}$	$42.39_{30.38}$	$69.03_{11.17}$	$76.93_{13.18}$	$65.86_{14.69}$	$84.77_{11.10}$	$59.46_{19.58}$	$48.78_{24.87}$
s	0	$78.17_{4.04}$	$78.44_{5.83}$	$78.30_{5.27}$	$82.92_{8.67}$	$77.55_{7.16}$	$69.80_{20.56}$	$72.69_{5.79}$	$75.61_{7.05}$	$70.48_{7.35}$	$85.20_{8.23}$	$61.89_{11.08}$	$65.02_{20.99}$
s	1	$82.38_{4.18}$	$83.61_{6.03}$	$81.53_{5.17}$	$83.77_{9.41}$	$79.74_{6.94}$	$79.65_{19.40}$	$77.69_{5.08}$	$81.84_{6.00}$	$74.34_{6.81}$	$84.56_{9.47}$	$65.46_{10.36}$	$77.59_{20.30}$
s	3	$83.72_{3.85}$	$85.64_{5.28}$	$82.22_{5.41}$	$83.42_{8.90}$	$81.94_{6.66}$	$78.26_{19.29}$	$77.01_{4.82}$	$82.10_{5.73}$	$72.88_{6.56}$	$83.50_{9.32}$	$64.80_{9.94}$	$73.83_{17.24}$
s	5	$84.52_{3.35}$	$88.09_{4.84}$	$81.56_{5.15}$	$81.15_{9.70}$	$82.97_{6.39}$	$77.14_{19.05}$	$77.76_{4.69}$	$83.31_{5.42}$	$73.29_{6.61}$	$83.84_{9.17}$	$65.12_{10.14}$	$74.69_{18.04}$
		medium-o4-mini						**high-o4-mini**					
f	0	$48.18_{9.41}$	$33.04_{9.81}$	$95.43_{4.92}$	$96.90_{6.46}$	$93.66_{7.38}$	$96.51_{7.17}$	$55.81_{11.28}$	$41.40_{12.59}$	$92.69_{7.82}$	$95.49_{7.10}$	$90.49_{10.52}$	$93.48_{12.34}$
f	1	$40.71_{8.79}$	$26.19_{7.57}$	$97.49_{2.44}$	$98.19_{2.50}$	$96.40_{4.27}$	$98.68_{3.44}$	$46.03_{11.09}$	$31.17_{10.83}$	$96.57_{4.03}$	$97.42_{5.44}$	$95.39_{6.38}$	$97.77_{5.13}$
f	3	$39.65_{7.74}$	$25.20_{6.42}$	$97.74_{2.21}$	$98.67_{1.94}$	$96.84_{3.51}$	$98.40_{4.24}$	$42.91_{9.50}$	$28.11_{8.49}$	$97.27_{2.69}$	$98.02_{2.93}$	$96.35_{4.17}$	$97.64_{5.85}$
f	5	$38.68_{8.02}$	$24.43_{6.63}$	$98.07_{2.02}$	$98.89_{1.64}$	$97.45_{3.03}$	$98.00_{5.26}$	$41.71_{9.26}$	$27.03_{8.01}$	$97.62_{2.15}$	$98.15_{2.51}$	$97.04_{3.03}$	$97.92_{5.04}$
		Llama-1B-Instruct						**Llama-3B-Instruct**					
f	0	$35.27_{17.26}$	$54.71_{22.71}$	$36.72_{26.13}$	$44.18_{26.76}$	$35.97_{27.80}$	$23.35_{27.59}$	$58.45_{10.66}$	$49.96_{15.09}$	$78.76_{14.57}$	$84.46_{13.29}$	$77.40_{15.91}$	$69.96_{25.35}$
f	1	$33.35_{11.37}$	$28.67_{17.55}$	$72.58_{29.09}$	$75.84_{26.84}$	$72.69_{30.26}$	$66.53_{35.43}$	$50.45_{13.19}$	$41.22_{17.23}$	$81.70_{18.33}$	$86.01_{17.16}$	$82.32_{18.14}$	$71.11_{29.67}$
f	3	$32.20_{11.45}$	$26.65_{17.91}$	$77.77_{28.82}$	$80.83_{26.69}$	$77.41_{29.70}$	$73.35_{33.29}$	$49.24_{15.20}$	$46.74_{19.42}$	$69.55_{24.60}$	$76.19_{23.61}$	$69.10_{25.99}$	$56.35_{33.18}$
f	5	$35.60_{13.72}$	$34.99_{21.79}$	$68.37_{31.82}$	$73.13_{29.46}$	$68.51_{32.72}$	$59.15_{39.18}$	$48.74_{14.78}$	$50.54_{17.93}$	$60.98_{26.24}$	$67.52_{26.10}$	$62.42_{27.51}$	$45.47_{32.88}$
s	0	$61.88_{5.82}$	$68.31_{8.65}$	$57.49_{7.39}$	$56.38_{10.73}$	$70.81_{7.12}$	$27.84_{16.32}$	$67.52_{5.90}$	$65.81_{8.48}$	$70.20_{6.80}$	$77.96_{8.24}$	$69.57_{9.17}$	$57.13_{19.47}$
s	1	$32.98_{6.22}$	$29.09_{8.45}$	$40.70_{8.26}$	$43.15_{13.15}$	$41.44_{8.28}$	$34.47_{18.00}$	$48.69_{8.88}$	$45.74_{13.31}$	$54.80_{7.07}$	$57.20_{10.45}$	$53.30_{8.21}$	$53.65_{19.15}$
s	3	$38.26_{6.12}$	$31.37_{7.59}$	$51.71_{8.19}$	$53.39_{11.33}$	$56.34_{8.26}$	$38.24_{18.11}$	$53.78_{8.89}$	$57.00_{15.41}$	$52.86_{7.06}$	$56.94_{11.02}$	$54.27_{9.37}$	$40.14_{17.67}$
s	5	$39.34_{6.97}$	$30.24_{7.93}$	$59.38_{8.02}$	$62.51_{10.85}$	$63.46_{8.22}$	$44.34_{17.61}$	$60.43_{7.84}$	$71.60_{13.85}$	$53.22_{6.82}$	$60.13_{11.05}$	$55.63_{9.75}$	$34.58_{16.03}$
		Llama-8B-Instruct						**Llama-70B-Instruct**					
f	0	$45.48_{9.37}$	$31.93_{9.82}$	$87.43_{12.08}$	$88.57_{13.03}$	$88.83_{11.67}$	$81.12_{22.77}$	$67.83_{9.90}$	$63.90_{16.75}$	$78.48_{13.30}$	$81.14_{13.92}$	$79.48_{14.76}$	$69.48_{26.23}$
f	1	$37.46_{11.94}$	$27.99_{15.68}$	$80.38_{22.39}$	$81.35_{22.81}$	$81.58_{22.97}$	$75.18_{27.90}$	$62.85_{12.61}$	$58.99_{17.71}$	$74.55_{15.59}$	$78.90_{16.01}$	$74.84_{16.51}$	$66.90_{26.51}$
f	3	$30.32_{5.90}$	$18.02_{4.19}$	$99.94_{0.60}$	$100.00_{0.00}$	$99.94_{0.59}$	$99.85_{2.03}$	$58.37_{12.91}$	$48.95_{17.06}$	$82.83_{14.57}$	$85.32_{12.83}$	$84.17_{16.30}$	$75.74_{23.83}$
f	5	$30.33_{5.91}$	$18.02_{4.19}$	$100.00_{0.00}$	$100.00_{0.00}$	$100.00_{0.00}$	$100.00_{0.00}$	$53.37_{14.54}$	$44.39_{17.41}$	$82.87_{18.75}$	$83.42_{18.37}$	$85.37_{18.61}$	$76.18_{27.59}$
s	0	$68.30_{5.95}$	$63.97_{8.94}$	$74.22_{5.58}$	$80.84_{8.19}$	$74.66_{7.23}$	$60.62_{20.33}$	$76.14_{4.84}$	$77.87_{7.82}$	$75.10_{5.53}$	$73.61_{10.81}$	$76.36_{5.50}$	$71.56_{21.92}$
s	1	$68.90_{6.98}$	$67.91_{9.82}$	$70.60_{6.10}$	$75.98_{9.23}$	$66.07_{8.72}$	$69.04_{19.92}$	$68.31_{8.47}$	$75.97_{14.31}$	$63.25_{6.43}$	$61.65_{11.28}$	$60.78_{8.25}$	$68.39_{20.89}$
s	3	$65.50_{6.72}$	$66.80_{9.56}$	$65.06_{7.17}$	$71.32_{9.56}$	$60.78_{8.39}$	$63.12_{20.91}$	$63.20_{8.53}$	$72.30_{15.56}$	$57.72_{7.17}$	$53.18_{11.18}$	$56.67_{9.35}$	$64.54_{21.82}$
s	5	$66.65_{6.50}$	$67.05_{9.37}$	$66.98_{6.41}$	$74.17_{9.45}$	$63.13_{9.31}$	$62.23_{21.11}$	$65.99_{7.41}$	$76.46_{13.85}$	$59.26_{6.86}$	$54.21_{11.52}$	$59.15_{9.26}$	$64.88_{20.04}$
		Qwen3-0.6B						**Qwen3-1.7B**					
f	0	$18.04_{9.30}$	$43.29_{26.59}$	$16.46_{14.61}$	$22.51_{17.29}$	$14.26_{14.25}$	$11.56_{15.87}$	$10.25_{9.49}$	$86.02_{20.54}$	$5.91_{7.31}$	$10.36_{9.49}$	$4.10_{7.96}$	$2.75_{6.48}$
f	1	$22.36_{9.02}$	$29.02_{20.01}$	$30.24_{26.06}$	$35.09_{26.05}$	$28.20_{26.43}$	$27.18_{29.81}$	$10.12_{9.34}$	$79.60_{29.76}$	$13.39_{28.55}$	$16.82_{27.74}$	$11.84_{29.26}$	$10.70_{28.97}$
f	3	$21.17_{9.28}$	$38.69_{23.36}$	$20.34_{14.32}$	$24.92_{17.04}$	$19.29_{13.91}$	$14.89_{17.93}$	$7.78_{6.27}$	$88.20_{21.59}$	$5.09_{10.74}$	$8.82_{12.02}$	$3.24_{10.26}$	$2.98_{11.34}$
f	5	$19.87_{9.99}$	$40.95_{26.37}$	$19.67_{15.96}$	$24.09_{18.15}$	$17.87_{15.44}$	$16.06_{19.42}$	$8.64_{6.49}$	$85.42_{24.87}$	$6.66_{15.30}$	$10.71_{15.52}$	$4.88_{15.55}$	$4.17_{15.88}$
s	0	$43.74_{6.31}$	$44.90_{9.86}$	$44.05_{7.56}$	$60.78_{11.78}$	$33.34_{8.34}$	$41.94_{19.35}$	$36.00_{8.24}$	$71.41_{10.04}$	$24.79_{7.60}$	$39.13_{11.40}$	$18.96_{7.61}$	$14.56_{15.02}$
s	1	$51.97_{6.39}$	$47.23_{9.62}$	$59.72_{6.79}$	$70.39_{9.05}$	$55.32_{8.05}$	$53.93_{17.38}$	$35.78_{7.46}$	$65.77_{11.56}$	$25.10_{6.39}$	$23.24_{7.90}$	$33.64_{8.64}$	$8.18_{9.65}$
s	3	$48.90_{6.31}$	$45.89_{8.84}$	$54.00_{8.14}$	$62.76_{10.75}$	$52.82_{9.24}$	$42.36_{18.94}$	$31.26_{7.05}$	$67.30_{10.40}$	$20.75_{5.80}$	$18.55_{7.01}$	$28.84_{8.57}$	$5.09_{7.21}$
s	5	$48.38_{6.33}$	$48.74_{10.75}$	$50.29_{8.46}$	$58.66_{11.18}$	$50.05_{9.90}$	$36.34_{18.73}$	$34.46_{7.97}$	$81.04_{11.03}$	$22.31_{6.50}$	$23.00_{8.94}$	$29.00_{8.67}$	$5.08_{6.61}$
		Qwen3-4B						**Qwen3-8B**					
f	0	$66.39_{16.56}$	$75.58_{16.33}$	$64.58_{20.96}$	$62.47_{23.52}$	$70.30_{22.66}$	$49.66_{31.16}$	$71.04_{10.77}$	$69.94_{12.12}$	$76.17_{15.49}$	$75.24_{18.17}$	$79.32_{14.48}$	$67.37_{29.01}$
f	1	$59.89_{20.20}$	$79.63_{14.52}$	$54.23_{23.08}$	$55.44_{26.35}$	$59.91_{24.54}$	$35.45_{28.72}$	$68.86_{13.90}$	$74.54_{15.50}$	$69.69_{17.63}$	$67.92_{20.37}$	$76.12_{16.61}$	$54.80_{30.83}$
f	3	$62.29_{20.97}$	$80.58_{14.28}$	$57.13_{22.87}$	$54.85_{23.60}$	$65.16_{24.09}$	$39.32_{28.43}$	$71.77_{11.73}$	$76.82_{10.36}$	$70.78_{16.75}$	$68.89_{18.34}$	$77.84_{15.80}$	$54.49_{30.81}$
s	0	$68.48_{5.55}$	$67.96_{8.94}$	$69.95_{6.29}$	$70.68_{10.41}$	$72.43_{7.71}$	$62.02_{19.69}$	$71.46_{6.01}$	$78.32_{7.62}$	$66.17_{7.12}$	$73.77_{10.23}$	$63.29_{10.24}$	$57.96_{22.96}$
s	1	$69.66_{5.59}$	$80.77_{7.32}$	$61.60_{6.43}$	$72.82_{10.90}$	$54.30_{9.35}$	$57.76_{22.39}$	$70.91_{5.30}$	$70.41_{8.29}$	$72.25_{6.22}$	$73.13_{10.83}$	$74.45_{8.67}$	$62.03_{22.19}$
s	3	$64.14_{6.23}$	$80.56_{7.25}$	$53.72_{7.15}$	$66.76_{10.58}$	$47.47_{10.41}$	$44.88_{22.09}$	$67.78_{5.87}$	$74.06_{7.97}$	$63.12_{7.34}$	$68.64_{11.34}$	$63.81_{9.49}$	$50.23_{20.53}$
s	5	$63.37_{6.55}$	$83.74_{7.48}$	$51.46_{7.47}$	$64.45_{11.10}$	$44.56_{11.12}$	$44.35_{22.42}$	$66.64_{5.87}$	$73.42_{7.91}$	$61.50_{6.73}$	$65.22_{10.97}$	$63.24_{9.64}$	$50.63_{21.16}$
		Phi-4-mini-instruct						**Phi-4-mini-reasoning**					
f	0	$27.65_{10.63}$	$27.41_{16.18}$	$58.22_{36.39}$	$63.15_{35.84}$	$54.33_{37.21}$	$59.29_{39.26}$	$31.72_{6.26}$	$20.17_{5.30}$	$85.91_{18.36}$	$88.66_{16.58}$	$84.47_{20.18}$	$85.25_{21.16}$
f	1	$26.03_{9.22}$	$25.09_{15.37}$	$60.63_{36.65}$	$65.80_{37.01}$	$55.60_{36.94}$	$63.76_{39.45}$	$30.84_{6.30}$	$20.22_{6.15}$	$86.03_{24.10}$	$88.99_{20.86}$	$84.17_{26.94}$	$86.54_{24.74}$
f	3	$27.00_{10.51}$	$32.71_{15.59}$	$35.84_{25.59}$	$42.49_{26.41}$	$32.25_{25.59}$	$33.76_{29.60}$	$33.79_{7.53}$	$23.06_{6.88}$	$79.18_{22.99}$	$82.80_{21.91}$	$77.84_{24.47}$	$77.43_{24.64}$
f	5	$25.64_{9.46}$	$33.52_{16.85}$	$32.80_{23.46}$	$39.64_{24.78}$	$28.62_{23.47}$	$31.76_{29.64}$	$34.42_{7.32}$	$24.37_{8.47}$	$75.20_{22.62}$	$79.17_{21.61}$	$73.97_{23.50}$	$71.52_{27.27}$
s	0	$36.91_{6.51}$	$25.42_{6.02}$	$70.73_{7.71}$	$75.57_{10.01}$	$72.50_{7.88}$	$58.36_{16.58}$	$36.83_{6.72}$	$26.71_{7.11}$	$62.88_{8.07}$	$66.11_{11.09}$	$64.61_{8.33}$	$51.63_{19.00}$
s	1	$39.14_{6.56}$	$28.45_{6.76}$	$66.12_{7.60}$	$67.84_{10.63}$	$71.55_{6.54}$	$48.62_{16.75}$	$36.94_{6.40}$	$24.86_{5.83}$	$75.31_{6.83}$	$78.67_{9.96}$	$76.94_{6.47}$	$65.10_{17.50}$
s	3	$44.50_{7.70}$	$34.74_{8.99}$	$64.79_{7.32}$	$67.76_{10.90}$	$69.29_{7.01}$	$48.68_{18.34}$	$37.15_{6.03}$	$25.41_{5.79}$	$72.40_{6.56}$	$74.39_{10.97}$	$76.00_{6.41}$	$59.47_{17.63}$
s	5	$49.37_{8.37}$	$42.15_{11.78}$	$62.92_{7.08}$	$66.12_{10.61}$	$67.10_{7.91}$	$47.13_{16.23}$	$38.44_{6.03}$	$26.44_{5.67}$	$73.07_{6.47}$	$75.58_{8.86}$	$75.58_{6.43}$	$62.29_{14.12}$
		MobileLLM-125M						**MobileLLM-350M**					
s	0	$33.23_{5.19}$	$20.58_{4.79}$	$90.18_{4.51}$	$91.42_{6.51}$	$91.95_{4.72}$	$83.76_{12.10}$	$34.25_{6.38}$	$21.85_{5.15}$	$82.74_{5.47}$	$82.78_{8.54}$	$85.27_{5.10}$	$76.78_{14.04}$
s	1	$33.90_{6.30}$	$21.13_{4.85}$	$89.38_{4.80}$	$91.02_{6.24}$	$90.79_{4.63}$	$83.87_{11.54}$	$35.16_{6.37}$	$24.75_{6.22}$	$63.83_{7.27}$	$68.40_{10.90}$	$64.54_{7.66}$	$55.39_{17.50}$
s	3	$34.40_{6.36}$	$22.20_{5.25}$	$80.11_{6.58}$	$84.82_{8.07}$	$80.91_{6.72}$	$70.45_{15.53}$	$36.35_{6.33}$	$28.03_{7.01}$	$54.38_{7.73}$	$60.63_{11.27}$	$54.72_{7.83}$	$44.07_{17.69}$
s	5	$34.99_{6.17}$	$22.64_{5.11}$	$80.40_{5.77}$	$83.64_{7.59}$	$84.12_{5.00}$	$66.45_{15.72}$	$37.84_{6.73}$	$31.96_{8.82}$	$49.07_{7.26}$	$57.45_{9.64}$	$48.28_{7.38}$	$38.26_{18.38}$
		MobileLLM-600M						**MobileLLM-1B**					
s	0	$33.96_{6.45}$	$22.03_{5.30}$	$77.53_{7.00}$	$79.65_{10.03}$	$78.33_{6.82}$	$71.91_{16.25}$	$33.38_{6.05}$	$22.40_{5.36}$	$68.78_{7.75}$	$69.21_{11.51}$	$71.55_{7.65}$	$62.87_{17.81}$
s	1	$36.13_{6.33}$	$27.10_{6.48}$	$56.91_{8.40}$	$61.45_{12.22}$	$58.28_{8.32}$	$45.65_{17.68}$	$33.12_{5.92}$	$22.81_{5.45}$	$63.68_{8.23}$	$66.74_{11.72}$	$63.90_{7.93}$	$57.95_{18.52}$
s	3	$36.18_{6.44}$	$29.21_{7.69}$	$49.84_{7.57}$	$57.54_{11.21}$	$48.52_{7.63}$	$40.83_{17.95}$	$33.55_{5.76}$	$23.94_{5.71}$	$59.55_{8.39}$	$61.97_{12.41}$	$61.06_{8.38}$	$52.16_{18.97}$
s	5	$36.98_{6.48}$	$30.83_{7.65}$	$48.51_{7.86}$	$57.93_{11.09}$	$46.94_{7.56}$	$36.70_{17.85}$	$34.57_{6.05}$	$25.51_{6.31}$	$56.64_{7.68}$	$59.28_{12.09}$	$58.67_{7.50}$	$47.52_{18.44}$

Table 3. **k-Shot Results**: We highlight the 1st and 2nd highest and 1st and 2nd lowest $\mathcal{E}$-Scores ($mean_{std.\ dev}$), which are then annotated with the highest and lowest $\mathcal{Z}$-Scores ($mean_{std.\ dev}$), cases with low $\mathcal{E}_P$ and high $\mathcal{E}_R$ (an over-deletion failure mode), and cases with high $\mathcal{E}_P$ and low $\mathcal{E}_R$ (an under-deletion failure mode).

	$\mathcal{E}_F$	$\mathcal{E}_P$	$\mathcal{E}_R$	$\mathcal{Z}_E$	$\mathcal{Z}_I$	$\mathcal{Z}_P$
gpt-4o-mini	$72.69_{4.79}$	$75.61_{7.05}$	$70.48_{7.35}$	$85.20_{8.23}$	$61.89_{11.08}$	$65.02_{20.99}$
gpt-4o-mini$_{ft}$	$94.77_{2.09}$	$96.63_{2.11}$	$93.08_{3.61}$	$87.71_{8.03}$	$97.39_{2.64}$	$89.31_{16.66}$
Llama-3B	$67.52_{5.90}$	$65.81_{8.48}$	$70.20_{6.89}$	$77.96_{8.24}$	$69.57_{9.17}$	$57.13_{19.47}$
Llama-3B$_{ft}$	$91.10_{2.39}$	$93.26_{3.49}$	$89.21_{4.05}$	$83.39_{7.22}$	$94.06_{3.96}$	$86.36_{17.26}$

Table 4. **Fine-Tuning** (s).

	GSM8K		*MMLU*	*CoQA*	
	EM$_{Strict}$ ↑	EM$_{Flexible}$ ↑	EM$_{Strict}$ ↑	EM$_{Strict}$ ↑	F1 ↑
gpt-4o-mini	$0.86_{0.01}$	$0.87_{0.01}$	$0.76_{0.00}$	$0.58_{0.02}$	$0.77_{0.01}$
gpt-4o-mini$_{ft}$	$0.51_{0.01}$	$0.81_{0.01}$	$0.69_{0.00}$	$0.55_{0.02}$	$0.73_{0.01}$
Llama-3B	$0.27_{0.01}$	$0.70_{0.01}$	$0.53_{0.00}$	$0.17_{0.02}$	$0.35_{0.02}$
Llama-3B$_{ft}$	$0.23_{0.01}$	$0.38_{0.01}$	$0.28_{0.00}$	$0.05_{0.01}$	$0.15_{0.01}$

Table 5. **Generalization of Fine-Tuned Models.**

tion. $\mathcal{Z}$**-Scores.** We further exploit parse tree annotations to measure the proportion of disfluent nodes removed: edited-type nodes ($\mathcal{Z}_E$), interjection nodes ($\mathcal{Z}_I$), and parenthetical nodes ($\mathcal{Z}_P$). These fine-grained scores capture structural differences in model behavior and highlight failure modes such as over-deletion and under-deletion.

3. RESULTS & RECOMMENDATIONS

We analyze the disfluency removal behavior of LLMs and provide recommendations (R1-R9).

Open-Source vs. Proprietary. Looking to Table 3, proprietary models (gpt-4o, gpt-4o-mini) achieve the highest scores, with margins of 10–15 points over the best open-source alternatives. We attribute this to training exposure to Whisper-transcribed speech data [24]. **(R1) Proprietary models are currently the most reliable for production systems, while open-source models require targeted augmentation with spoken data.**

Segmentation (s) vs. Full Input (f). Segmenting transcripts consistently improves both mean performance and stability, e.g., gpt-4o improves from $\mathcal{E}_F$=76.13 (f) to 82.38 (s) at k=1. This supports prior evidence of long-context degradation in LLMs [25, 26]. **(R2) Segmentation is an effective preprocessing step that should be applied.**

Few-Shot Sensitivity (k). Increasing k does not uniformly improve results. Small models (e.g., MobileLLM) gain slightly, but others show degradation (e.g. Llama-3B/8B/70B) when more examples are provided. **(R3) Few-shot prompting should be used with caution, as some model families misinterpret exemplars and over-edit fluent text.**

Disfluency Category Performance. $\mathcal{Z}$-Scores show that EDITED nodes are handled well, but INTJ and PRN nodes are frequently missed, despite prior work suggesting these are the easiest to detect [19, 17]. **(R4) Future modeling should focus on under-served categories (INTJ, PRN) to improve robustness across all disfluency types.**

Over-Deletion Failures. Several models (e.g., Llama-8B, o4-mini) achieve near perfect recall but at the cost of very low precision, deleting fluent tokens. Segmentation often mitigates this collapse mode. **(R5) Segment-level evaluation helps reduce over-deletion risk.**

Under-Deletion Failures. Some models (e.g., Qwen series) exhibit the opposite trend of over-deletion, achieving high precision but low recall (purple). These models preserve most fluent tokens but fail to remove many true disfluencies, especially in INTJ and PRN categories. This reflects conservative editing strategies and limited exposure to conversational disfluency distributions. **(R6) Models prone to under-deletion require additional filtering or targeted fine-tuning to ensure sufficient disfluency coverage.**

Reasoning-Oriented Models. Models tuned for reasoning (o4-mini, Phi-4) perform poorly, showing high recall but extreme over-deletion (blue). **(R7) Reasoning capability does not translate to disfluency removal; specialized evaluation remains necessary.**

Impact of Model Size. Model scaling generally improves disfluency removal, with Qwen, GPT, and Llama families showing upward trends. However, gains are nonlinear – e.g., Qwen3-1.7B underperforms both smaller and larger variants – likely due to training data or optimization differences rather than capacity limits. **(R8) Model choice should be guided by empirical benchmarks on target domains and disfluency categories rather than size alone.**

Fine-Tuning and Generalization. Looking to Tables 4 and 5, fine-tuning improves performance to near SOTA levels (e.g., gpt-4o-mini$_{ft}$ achieves $\mathcal{E}_P$=96.6), but evaluation on GSM8K, MMLU, and CoQA shows degraded performance on unrelated tasks. **(R9) Fine-tuning is suitable for dedicated disfluency pipelines, but not for general-purpose conversational models.**

4. DISCUSSION & FUTURE WORK

DRES establishes a controlled, reproducible upper bound for disfluency removal, revealing consistent gaps between proprietary and open-source LLMs that are driven more by training exposure than model size. Our results highlight open challenges: closing the precision-recall trade-off (especially over-deletion in reasoning-tuned models), improving coverage for under-served categories (INTJ, PRN), and mitigating generalization loss from fine-tuning.

We see value in a modular approach where specialized disfluency removal components preprocess ASR output before downstream reasoning, helping maintain the flexibility of general purpose LLMs. Promising directions include exploring lightweight adapters, multi-task setups, and continual learning to approach state-of-the-art accuracy while mitigating catastrophic forgetting. As speech language models mature, incorporating disfluency handling into end-to-end architectures will require careful design to maintain generalization. Extending DRES to multilingual and multimodal settings and systematically evaluating its downstream impacts (e.g., command success rates, summarization fidelity) can help the community build more robust and deployable systems.

5. REFERENCES

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

[2] Heather Bortfeld, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, and Susan 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.

[3] Sharon Oviatt, "Predicting spoken disfluencies during human-computer interaction," *Computer Speech and Language*, 1995.

[4] Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li, "Voicebench: Benchmarking llm-based voice assistants," *arXiv preprint arXiv:2410.17196*, 2024.

[5] Maria Teleki, Lingfeng Shi, Chengkai Liu, and James Caverlee, "I want a horror – comedy – movie: Slips-of-the-Tongue Impact Conversational Recommender System Performance," in *INTERSPEECH*, 2025.

[6] Varun Nathan, Ayush Kumar, and Jithendra Vepa, "Investigating the role and impact of disfluency on summarization," in *EMNLP: Industry Track*, 2023, pp. 541–551.

[7] Fabian Retkowski, Maike Züfle, Andreas Sudmann, Dinah Pfau, Shinji Watanabe, Jan Niehues, and Alexander Waibel, "Summarizing speech: A comprehensive survey," 2025.

[8] Maria Teleki, Xiangjue Dong, Soohwan Kim, and James Caverlee, "Comparing asr systems in the context of speech disfluencies," in *Interspeech 2024*, 2024, pp. 4548–4552.

[9] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[10] Hugo Touvron, Thibault Sellam, Yacine Jernite, Jade Copet, Loubna Ben Allal, et al., "The llama 3 herd of models," *Computing Research Repository*, vol. arXiv:2407.21783, 2024.

[11] Zechun Liu, Changsheng Zhao, et al., "MobileLLM: Optimizing sub-billion parameter language models for on-device use cases," 2024.

[12] Qwen Team, "Qwen3 technical report," 2025.

[13] Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, et al., "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," 2025.

[14] Marcus Mitchell, Beatrice Santorini, M Marcinkiewicz, and Ann Taylor, "Treebank-3 LDC99T42 Web Download," 1999, [Online]. Available:https://catalog.ldc.upenn.edu/LDC99T42.

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

[16] Eugene Charniak and Mark Johnson, "Edit Detection and Parsing for Transcribed Speech," *NAACL*, 2001.

[17] Mark Johnson and Eugene Charniak, "A TAG-based noisy-channel model of speech repairs," in *ACL*, 2004, pp. 33–39.

[18] Elizabeth Shriberg, *Preliminaries to a Theory of Speech Disfluencies*, Ph.D. thesis, 1994.

[19] Paria Jamshid Lou and Mark Johnson, "Improving disfluency detection by self-training a self-attentive model," in *ACL*, 2020, pp. 3754–3763.

[20] Dena Mujtaba, Nihar 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.

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

[22] Feng Wang, Wei Chen, Zhen Yang, Qianqian Dong, Shuang Xu, and Bo Xu, "Semi-supervised disfluency detection," in *International Conference on Computational Linguistics*, 2018, pp. 3529–3538.

[23] Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi, "Disfluency detection using a bidirectional LSTM," in *INTERSPEECH*, 2016.

[24] Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson, and Nico Grant, "How tech giants cut corners to harvest data for a.i.," *The New York Times*, 2024.

[25] Jiaheng Liu, Dawei Zhu, et al., "A comprehensive survey on long context language modeling," *arXiv*, 2025.

[26] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen, "Long-context llms struggle with long in-context learning," *arXiv*, 2024.