Z-Scores: A Metric For Linguistically Assessing 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

Accepted to ICASSP '26!

Why disfluency?

- Spontaneous speech contains disfluencies:
  - **INTJ**: interjections like uh, um
  - **PRN**: parentheticals like you know, I mean
  - **EDITED**: false starts / repairs / restarts
- These disfluencies hurt downstream systems like transcription, translation, and conversational recommendation.
- Existing evaluation **(ε-Scores)** mainly uses **token-level precision**, **recall**, and **F1** tell us how well a model performs overall, but not what types of disfluencies it fails on.

Metaprompting Case Study

		gpt-4o-mini					
M		$\mathcal{E}_F$	$\mathcal{E}_P$	$\mathcal{E}_R$	$\mathcal{Z}_E$	$\mathcal{Z}_I$	$\mathcal{Z}_P$
s	P_0	$72.69_{5.79}$	$75.61_{7.05}$	$70.48_{7.35}$	$85.20_{8.23}$	$61.89_{11.08}$	$65.02_{20.99}$
s	P_1	$81.94_{3.75}$	$84.47_{4.92}$	$79.90_{5.65}$	$83.67_{9.27}$	$78.28_{8.10}$	$74.86_{22.06}$
s	P_2	$79.86_{5.42}$	$76.88_{7.02}$	$83.52_{6.12}$	$87.45_{7.48}$	$79.60_{8.89}$	$87.09_{15.46}$

- Evaluated on the Switchboard dataset
- Compare a baseline prompt (P0) against metaprompts (P1, P2) containing explicit disfluency examples
- Traditional E-Scores indicate modest aggregate improvement
- Z-Scores reveal gains are concentrated on INTJ and PRN
- Performance on EDITED spans is already strong and remains stable
- Z-Scores provide insight that is not visible from token-level metrics alone

Z-Score Framework

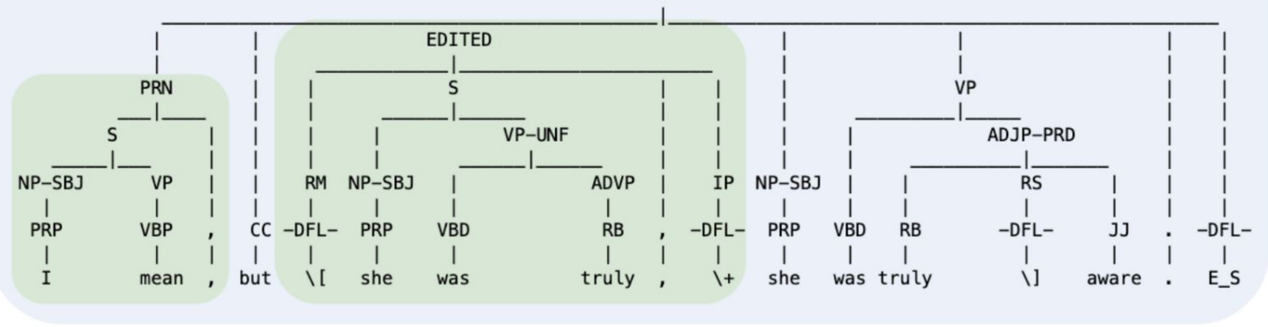
Z-Scores measure performance on disfluencies: EDITED, INTJ, and PRN.

Data Pre-Processing: $T = \left\{ \left(t^{(i)}_{tree}, t^{(i)}_{fluent}, t^{(i)}_{tag}, t^{(i)}_{disfluent} \right) \right\}_{i=1}^{N}$ from $t^{(i)}_{tree}$ using disfluent node types: $\{EDITED, PRN, INTJ\}$.

t_{tree}

```
((S(PRN(S (NP-SBJ (PRP I))(VP (VBP mean)))(, ,))(CC but)(EDITED(RM(-DFL- \[))(S(NP-SBJ(PRP she))(VP-UNF(VBD wa
s)(ADVP(RB truly))))(, ,)(IP(-DFL- \+)))(NP-SBJ(PRP she))(VP(VBD was)(ADJP-PRD(RB truly)(RS(-DFL -\]))(JJ awar
e)))(. .)(-DFL- E_S)))
```

Equivalently:

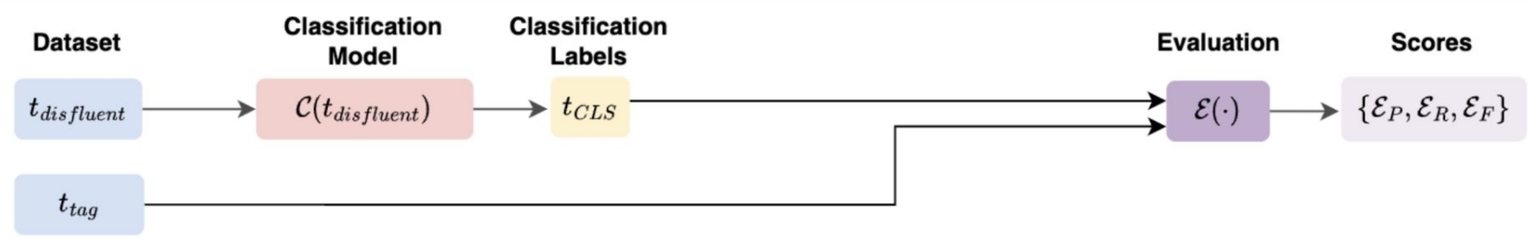


$t_{\text{disfluent}}$
"I mean, but she was truly, she was truly aware."

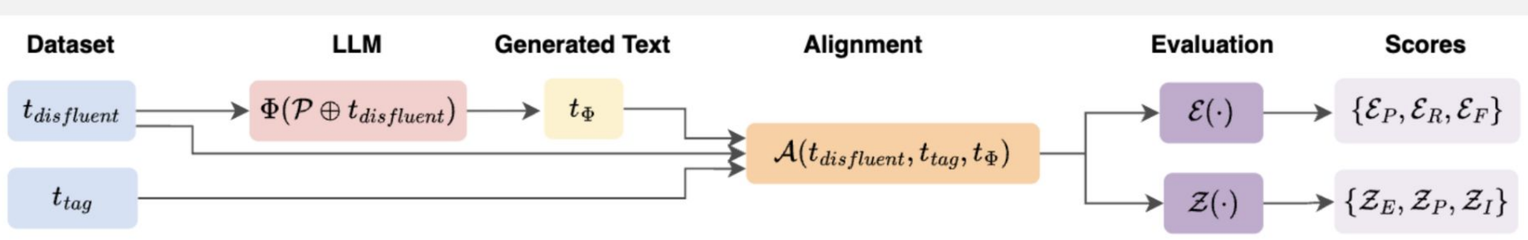
t_{tag}
[PRN, PRN, NONE, EDITED, EDITED, EDITED, NONE, NONE, NONE, NONE]

t_{fluent}
"But she was truly aware."

Previous Work: Classify each word in the original input as fluent (O) or disfluent (I) to evaluate disfluency removal performance in terms of $\mathcal{E}$-Scores.



Our Method: Align LLM output with original input to evaluate disfluency removal performance in terms of $\mathcal{E}$-Scores & $\mathcal{Z}$-Scores.



Alignment

We contribute an alignment module, so we can evaluate generative models:

- Previous work used span classification; instead, the alignment module allows models to be evaluated directly on disfluency removal.

$\mathbb{1}_{gt}$: *Based on the ground truth parse tree, should this word be removed by* Φ*?*
$\mathbb{1}_{pred}$: *Was this word actually removed by* Φ*?*

$t_{disfluent}$	t_{tag}	t_{Φ}	t_{CLS}	$\mathbb{1}_{gt}$	$\mathbb{1}_{pred}$	$\mathbb{1}_{tp}$	$\mathbb{1}_{tn}$	$\mathbb{1}_{fp}$	$\mathbb{1}_{fn}$
i	PRN	i	I	1	0	0	0	0	1
mean	PRN	mean	I	1	0	0	0	0	1
but	NONE	but	O	0	0	0	1	0	0
she	EDITED		I	1	1	1	0	0	0
was	EDITED		I	1	1	1	0	0	0
truly	EDITED		I	1	1	1	0	0	0
she	NONE		I	0	1	0	0	1	0
-	-	Luna	O	*	*	*	*	*	*
was	NONE	was	O	0	0	0	1	0	0
truly	NONE	truly	O	0	0	0	1	0	0
aware	NONE	aware	O	0	0	0	1	0	0

Calculation

We calculate E-Scores, enabling comparison to previous work:

$$\mathcal{E}_P = \frac{\Sigma_{tp}}{\Sigma_{tp}+\Sigma_{fp}} = \frac{3}{3+1} \rightarrow 75.0$$
$$\mathcal{E}_R = \frac{\Sigma_{tp}}{\Sigma_{tp}+\Sigma_{fn}} = \frac{3}{3+2} \rightarrow 60.0$$
$$\mathcal{E}_F = \frac{2 \cdot \mathcal{E}_P \cdot \mathcal{E}_R}{\mathcal{E}_P+\mathcal{E}_R} = \frac{2 \cdot 0.75 \cdot 0.60}{0.75+0.60} \rightarrow 66.0$$

- εP penalizes over-deletion
- εR penalizes under-deletion
- εF is the harmonic mean

We calculate Z-Scores, revealing structural performance deficits:

$$\mathcal{Z}_E = \frac{\mathbb{1}_{gt} \wedge (w_{tag}=EDITED) \wedge \mathbb{1}_{pred}}{\mathbb{1}_{gt} \wedge (w_{tag}=EDITED)} = \frac{3}{3} \rightarrow 100\%$$
$$\mathcal{Z}_I = \frac{\mathbb{1}_{gt} \wedge (w_{tag}=INTJ) \wedge \mathbb{1}_{pred}}{\mathbb{1}_{gt} \wedge (w_{tag}=INTJ)} = \frac{0}{0} \rightarrow NaN$$
$$\mathcal{Z}_P = \frac{\mathbb{1}_{gt} \wedge (w_{tag}=PRN) \wedge \mathbb{1}_{pred}}{\mathbb{1}_{gt} \wedge (w_{tag}=PRN)} = \frac{0}{2} \rightarrow 0\%$$

In this example, Z-Scores reveals that the model is good at removing EDITED, but poor at PRN

See Also →

Teleki, Maria, et al. "Conversational Speech Reveals Structural Robustness Failures in SpeechLLM Backbones."

DRES builds on Z-Scores to investigate how reliably LLMs handle the structural complexity of conversational speech.

Texas A&M University
Department of Computer Science & Engineering

Contact: thomasdocog@tamu.edu
Applying to PhD programs in Fall 2026!