

# Z-Scores: A Metric for Linguistically Assessing Disfluency Removal

OPEN TO POSTDOGS!

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

## Disfluencies



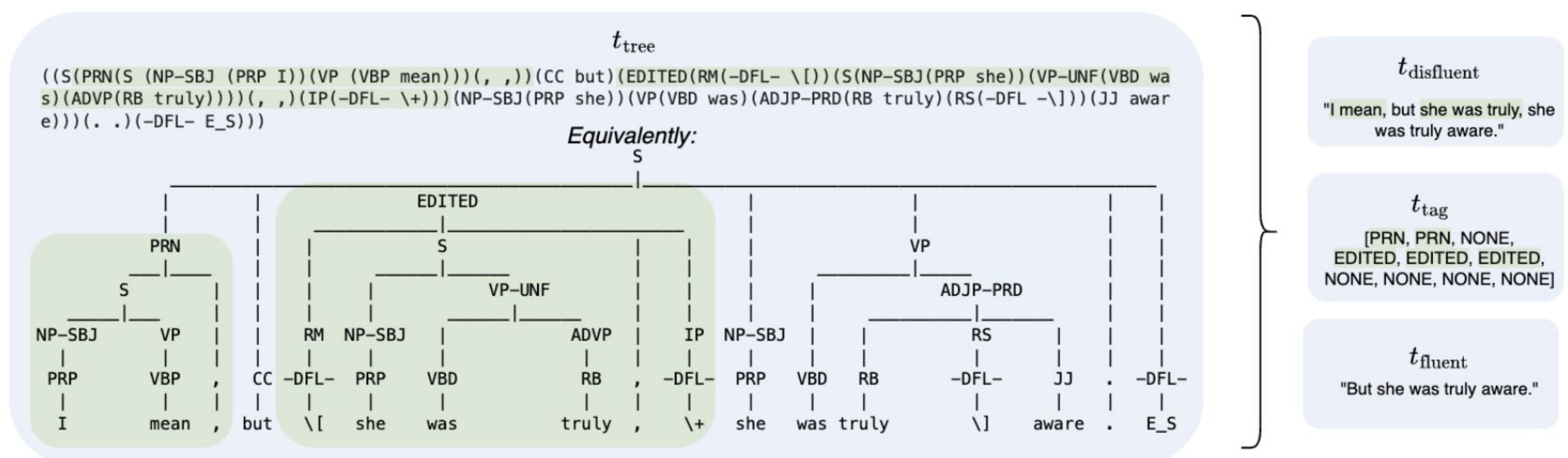
## Results

		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 <sub>5.79</sub>	75.61 <sub>7.05</sub>	70.48 <sub>7.35</sub>	85.20 <sub>8.23</sub>	61.89 <sub>11.08</sub>	65.02 <sub>20.99</sub>	
$s$	$P_1$	81.94 <sub>3.75</sub>	84.47 <sub>4.92</sub>	79.90 <sub>5.65</sub>	83.67 <sub>9.27</sub>	78.28 <sub>8.10</sub>	74.86 <sub>22.06</sub>	
$s$	$P_2$	79.86 <sub>5.42</sub>	76.88 <sub>7.02</sub>	83.52 <sub>6.12</sub>	87.45 <sub>7.48</sub>	79.60 <sub>8.89</sub>	87.09 <sub>15.46</sub>	

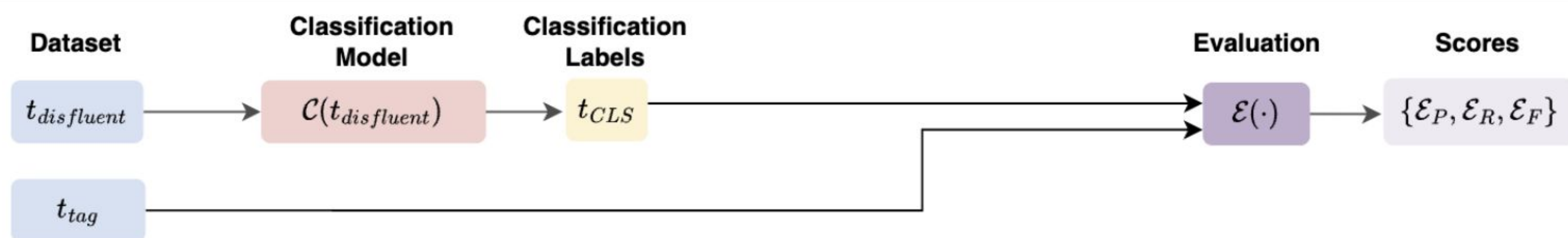
**Metaprompting Experiment.** We provide examples of the most-failed-on categories of disfluencies, and see that performance improves.

## Z-Score Framework

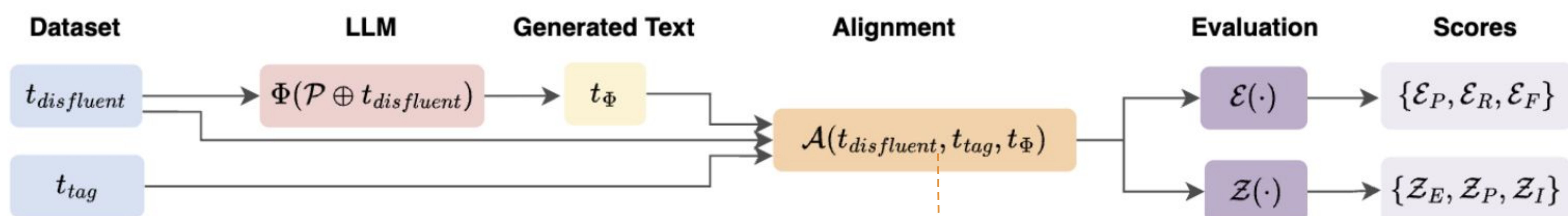
**Data Pre-Processing:**  $T = \left\{ \left( t_{tree}^{(i)}, t_{fluent}^{(i)}, t_{tag}^{(i)}, t_{disfluent}^{(i)} \right) \right\}_{i=1}^N$  from  $t_{tree}^{(i)}$  using disfluent node types:  $\{EDITED, PRN, INTJ\}$ .



**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 Module**, enables scoring of generative models, and complements previous E-Score metrics.

$t_{disfluent}$	$t_{tag}$	$t_{\phi}$	$t_{CLS}$	$\mathbb{1}_{gt}$	$\mathbb{1}_{pred}$	$\mathbb{1}_{tp}$	$\mathbb{1}_{tn}$	$\mathbb{1}_{fp}$	$\mathbb{1}_{fn}$
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

PAPER HERE

