2025 ICASSP HYDERABAD, INDIA
Celebrating Signal Processing

Speech Recognition with LLMs Adapted to Disordered Speech Using Reinforcement Learning

https://arxiv.org/abs/2501.00039

Chirag Nagpal, **Subhashini Venugopalan**, Jimmy Tobin, Marilyn Ladewig, Katherine Heller, Katrin Tomanek

Project Euphonia

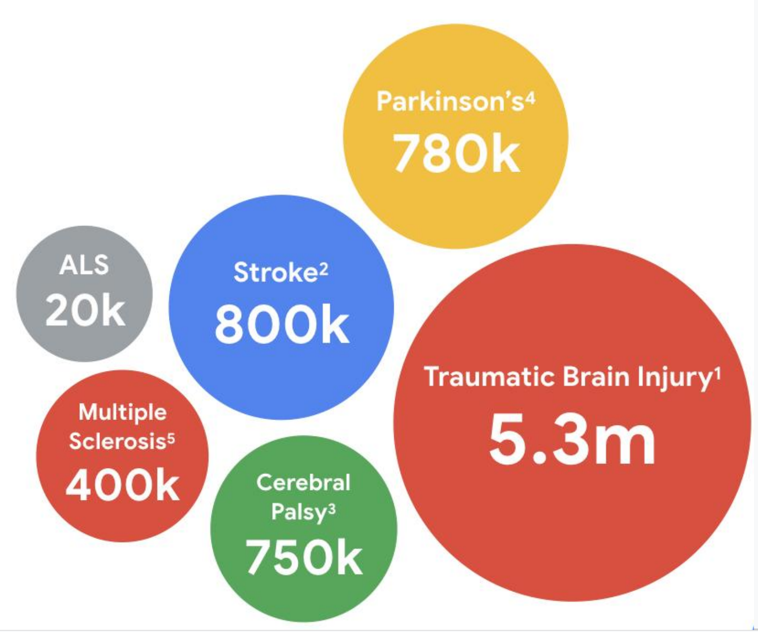
Improve ASR to help people with **speech disorders** who have difficulty being understood by other people and technology.

Our goal is to help these users **communicate** and **gain independence**.

https://sites.research.google/euphonia/about/

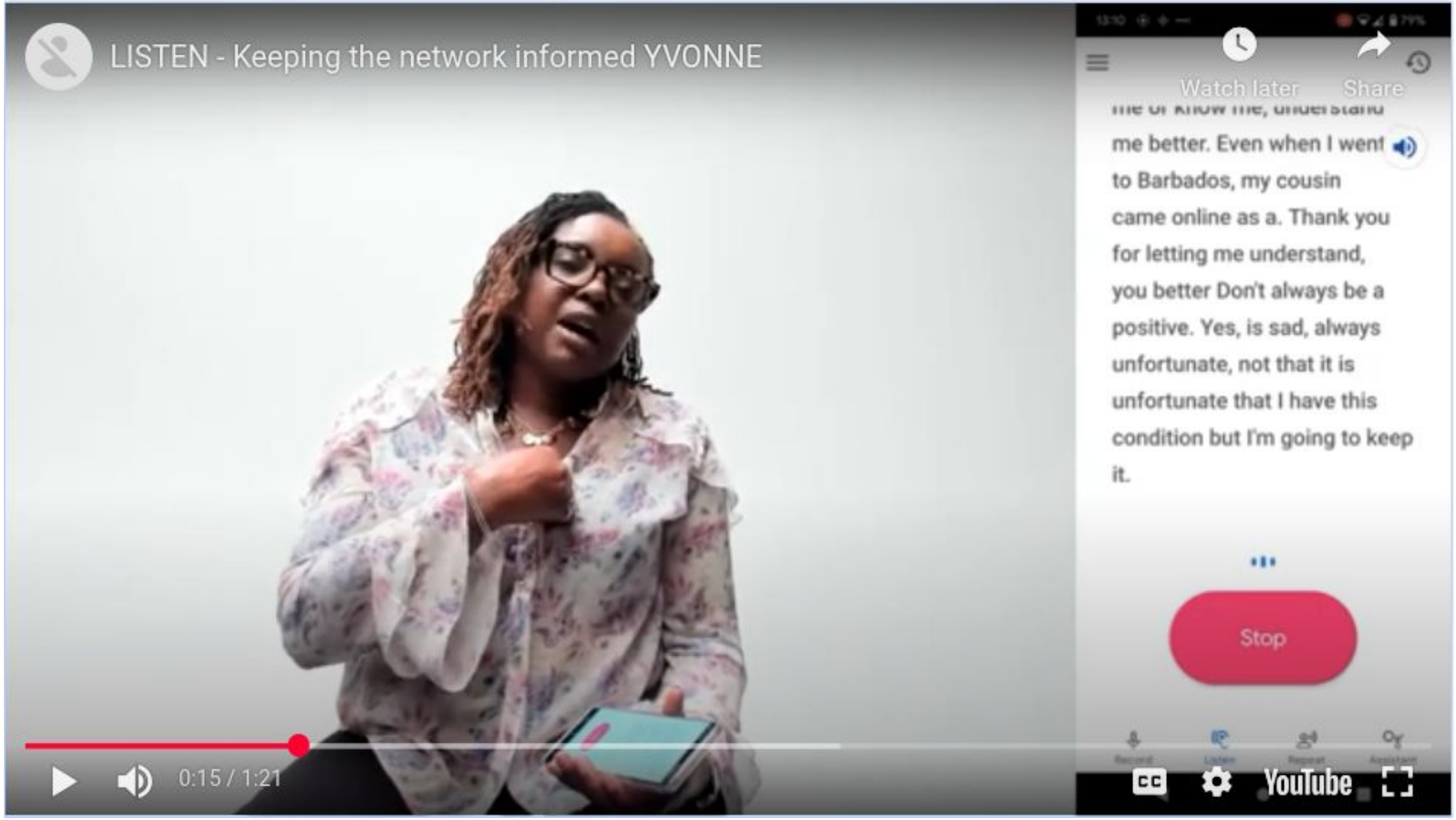
Condition prevalence (US)

Millions of users have neurological conditions that cause speech impairments, in the US and around the world.



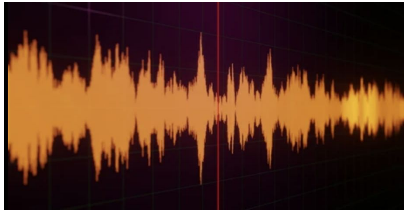
Google

Project Relate - Personalize their on-device ASR model



Can mLLMs help recognize impaired speech?

+



(image+speech)

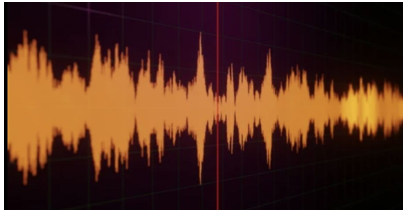
Gemini

→ “I’d like a croissant”

Can start with open source text-only LLMs?

- LLMs already have a lot of world knowledge.
- Can we add speech inputs?
- Small model / on-device

+

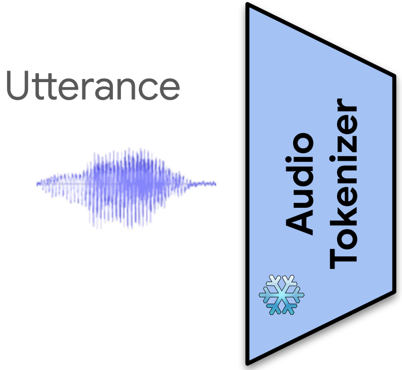


Gemma

→ "I'd like a croissant"

How do you turn an LLM into an ASR model?

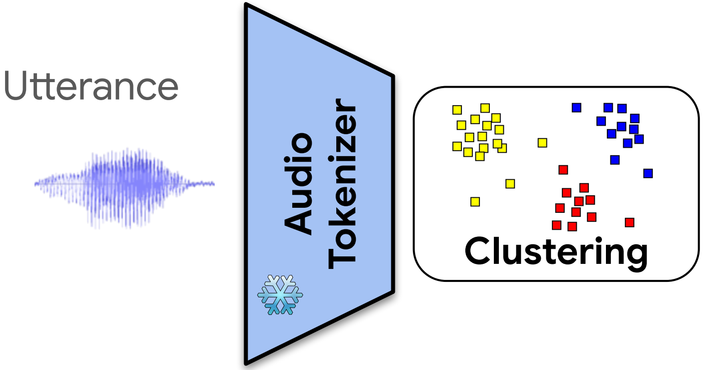
Tokenization of the audio



How do you turn an LLM into an ASR model?

Tokenization of the audio

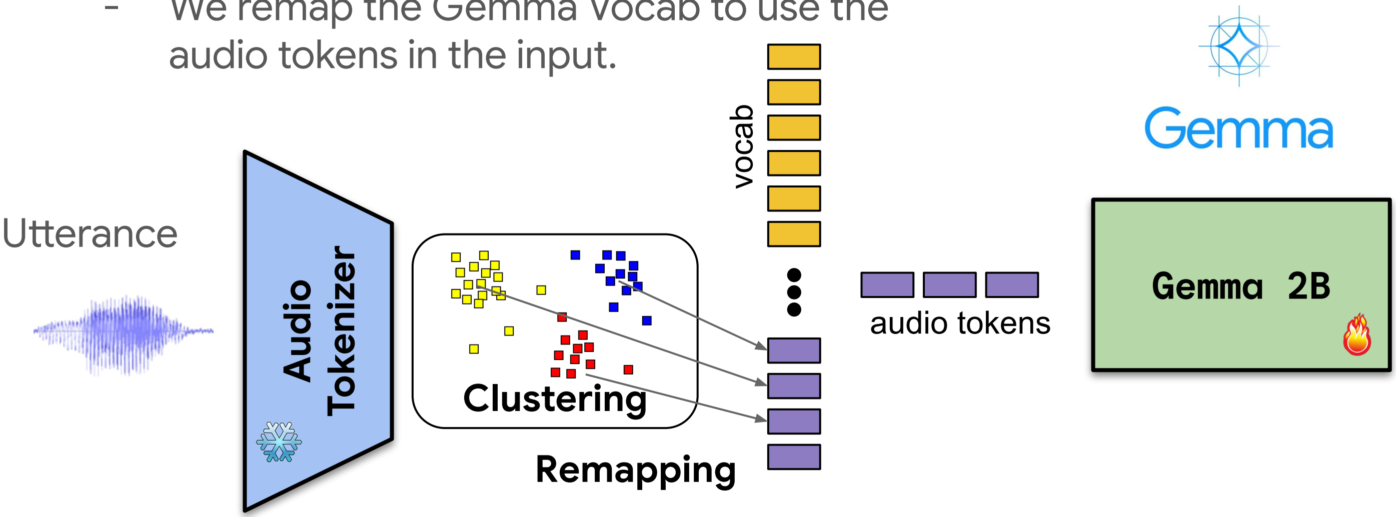
- We cluster embeddings to 1024 tokens from the Librispeech Corpus.



How do you turn an LLM into an ASR model?

Tokenization of the audio

- We cluster embeddings to 1024 tokens from the Librispeech Corpus.
- We remap the Gemma Vocab to use the audio tokens in the input.

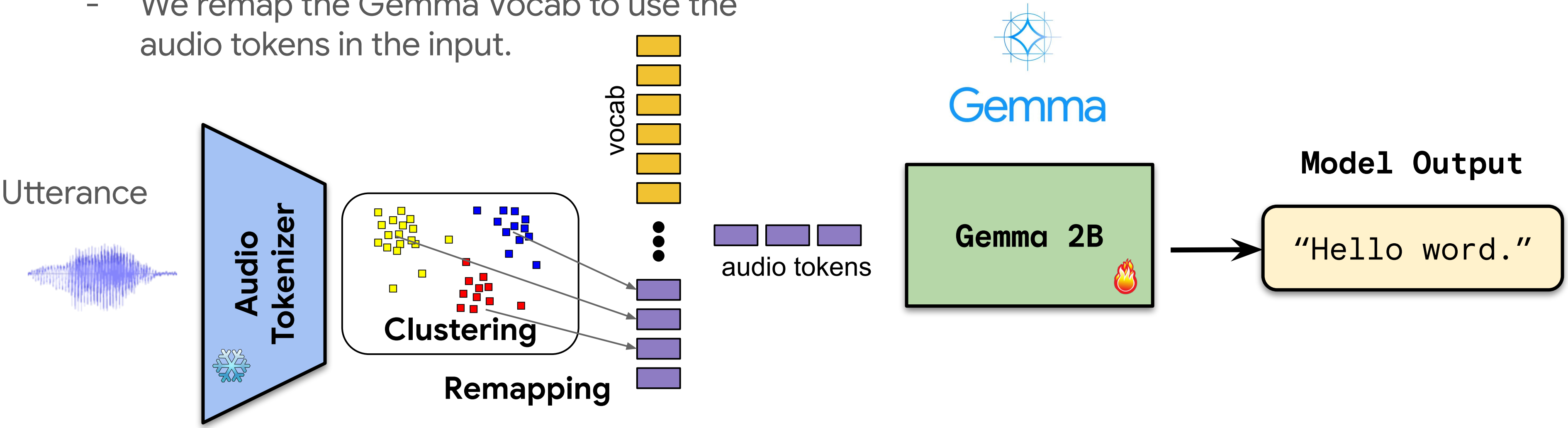


Specifically replace the low-frequency tokens

How do you turn an LLM into an ASR model?

Tokenization of the audio

- We cluster embeddings to 1024 tokens from the Librispeech Corpus.
- We remap the Gemma Vocab to use the audio tokens in the input.



Specifically replace the low-frequency tokens

Let's train it.

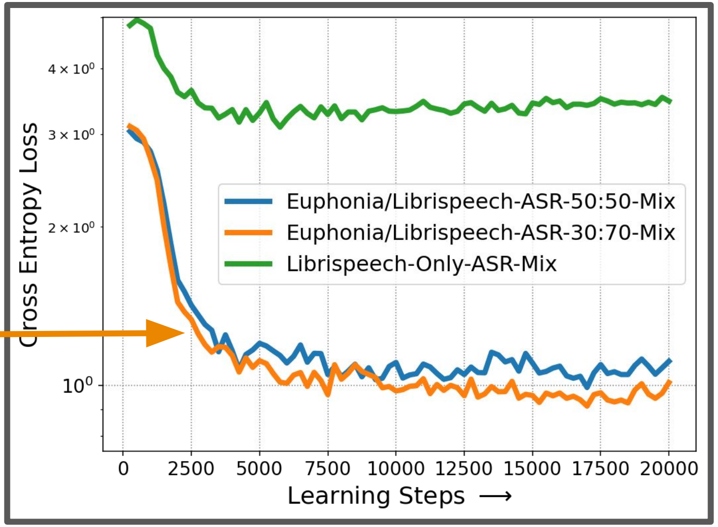
- First train on Librispeech
  - Librispeech: 1000 hrs of audio from books
- Then adapt to disordered speech
  - Euphonia also ~1000 hrs of prompted audio

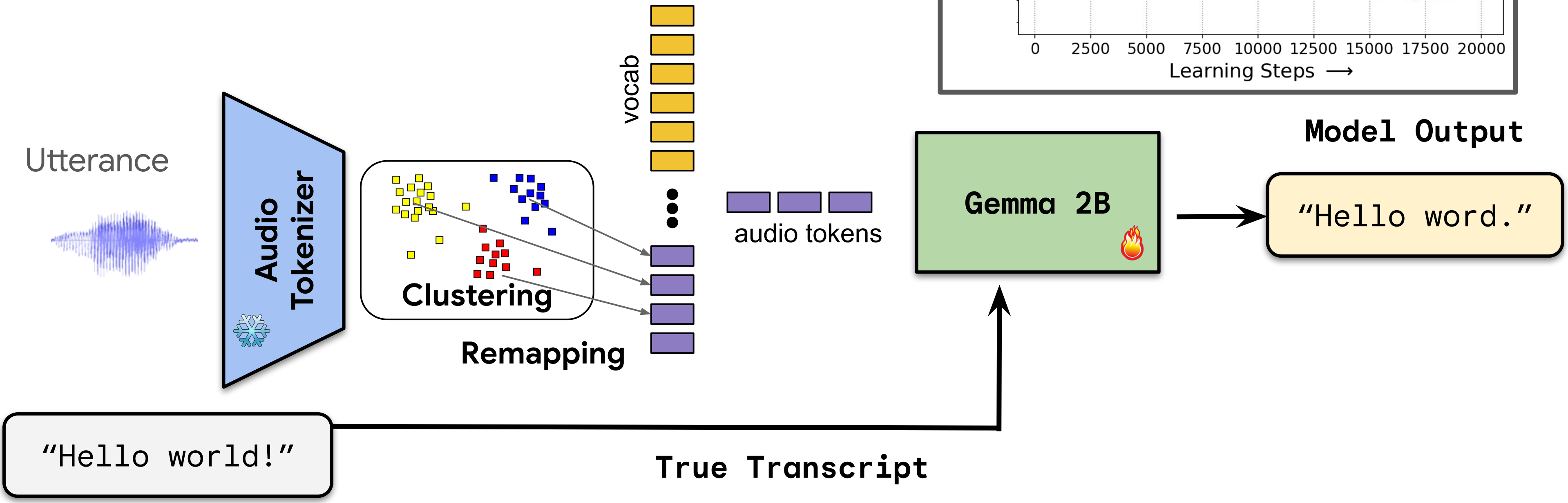
  - **Training**: 900k utterances, 1246 speakers
  - **Test**: 5699 utterances, 200 speakers

Supervised Fine Tuning

Mixture of Librispeech and Euphonia Audio

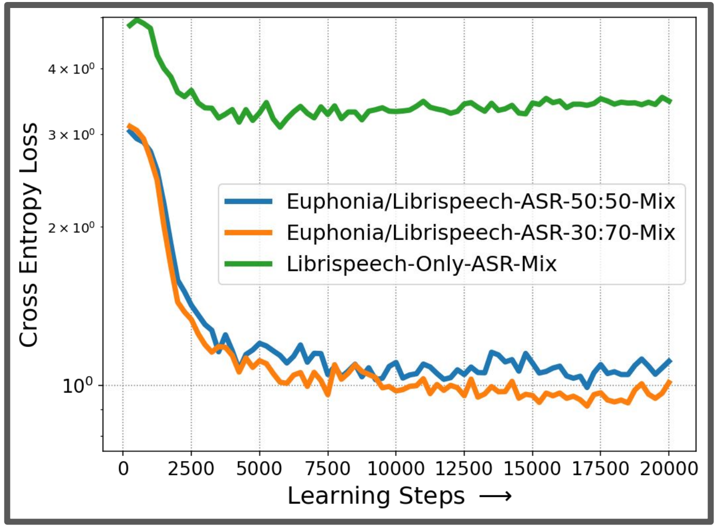
- Augmenting the SFT mixture with ASR data gives improved performance.





How well does it work?

TABLE I: Training the LLM on ASR data with a 30:70 mix of Euphonia:Librispeech leads to significant (*) improvements on Euphonia and little loss on Librispeech. ↑ and ↓ indicate higher or lower is better respectively. **bold** shows best score.

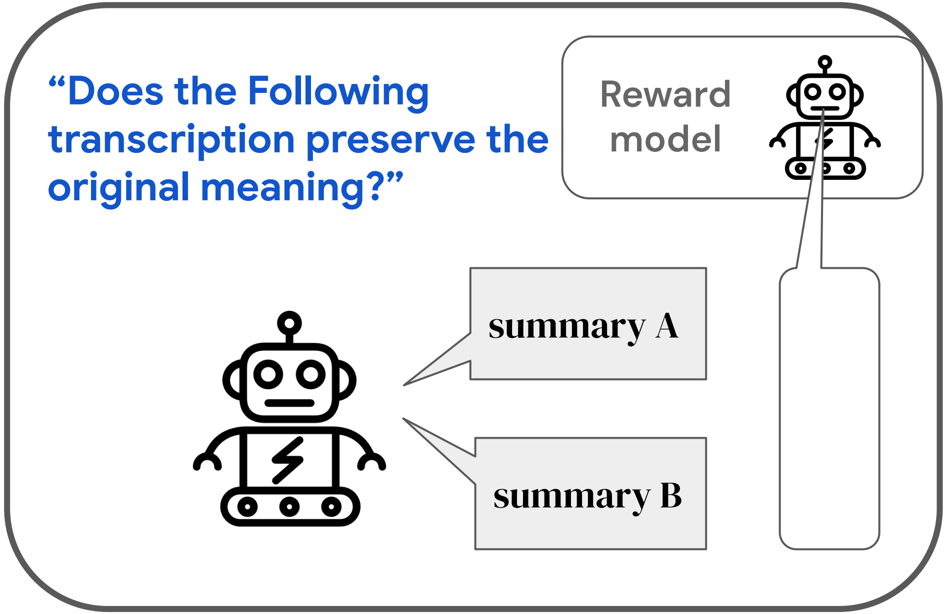


Dataset mixture	Euphonia Test		Euphonia Dev		Librispeech Dev	
	WER ↓	MP ↑	WER ↓	MP ↑	WER ↓	MP ↑
Librispeech Only	70.9	39.0	66.5	31.8	**17.1**	**86.6**
30:70 mixture	**50.4***	**48.2***	**47.3***	**48.1***	17.2	85.6

Can RL can help generalize further than SFT on Disordered Speech Data?

We need a reward

- Can meaning preservation be a reward?



Example: Meaning preservation as reward

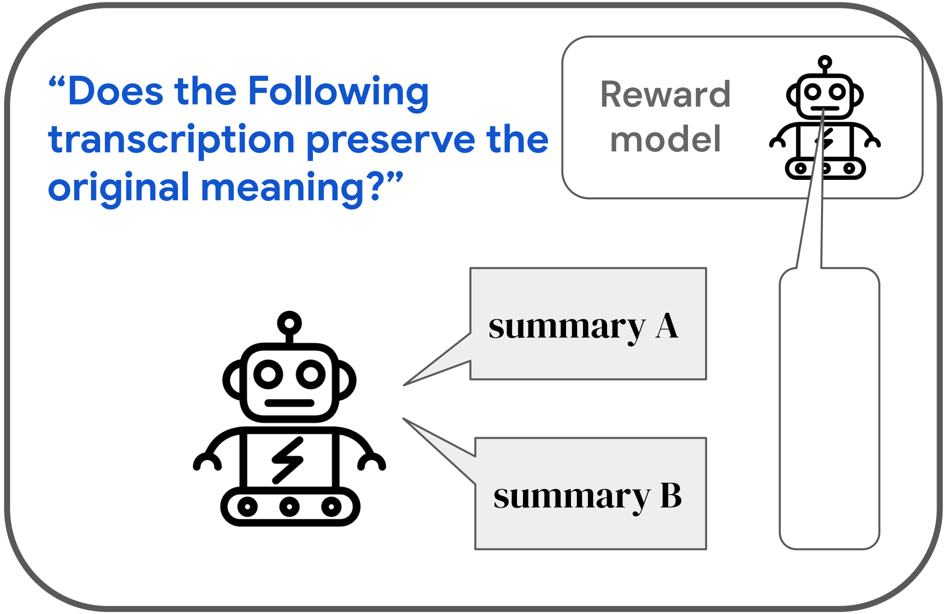
Insight: High word errors can still preserve meaning !

Ground Truth: `“Not so good today”`

Output A: `“not so good to the.”`

Output B: `“not so good to day.”`

Both have same same WER, but B Preserves Meaning.



Meaning preservation as a reward

Conferences > ICASSP 2024 - 2024 IEEE Inter...

Large Language Models As A Proxy For Human Evaluation In Assessing The Comprehensibility Of Disordered Speech Transcription

Publisher: IEEE Cite This PDF

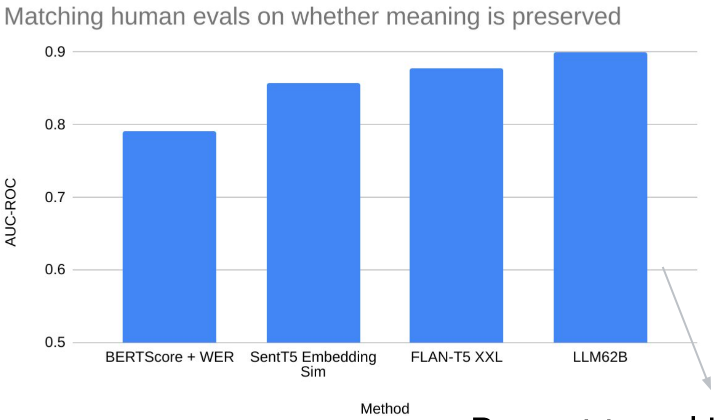
Katrin Tomanek ; Jimmy Tobin ; Subhashini Venugopalan ; Richard Cave ; Katie Seaver ; Jordan R. Green **All Authors**

in ICASSP 2024

Meaning preservation as a reward

Train models to predict human labels of whether meaning was preserved

ASR Transcript



Is meaning preserved?

Prompt-tuned LLM does best
(+ case-study on model deployment of SI-ASR vs personalized)

This work: we retrain Gemma 2B as a reward model achieving AUC ~0.88

Using Meaning Preservation as a Reward signal

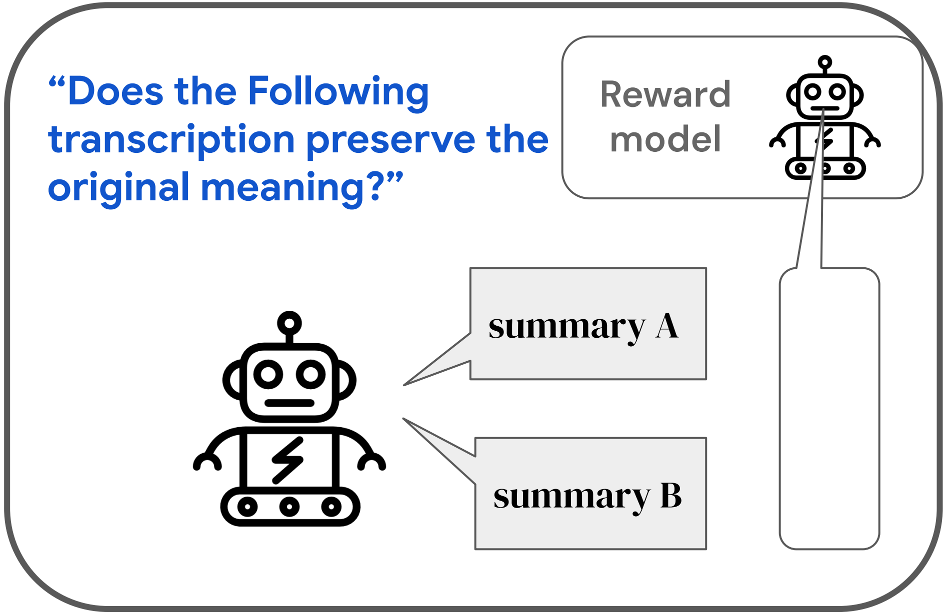
Insight: **High word errors can still preserve meaning !**

Ground Truth: `“Not so good today”`

Output A: `“not so good to the.”`

Output B: `“not so good to day.”`

Both have same same WER, but B Preserves Meaning.

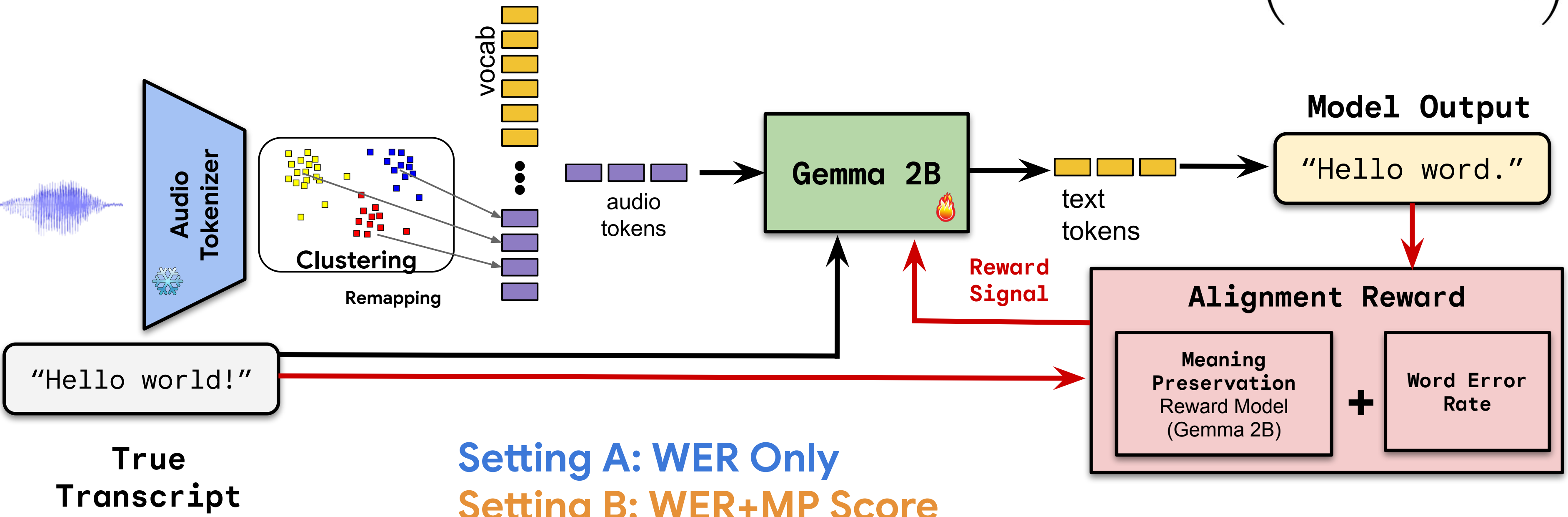


$$R(\boldsymbol{x},\boldsymbol{y};\boldsymbol{y}^*) := \gamma \cdot \overset{\text{Reward Model}}{\mathtt{MP}(\boldsymbol{y},\boldsymbol{y}^*)} + \ln\left(1 - \overset{\text{Ground Truth}}{\mathtt{WER}(\boldsymbol{y},\boldsymbol{y}^*)}\right)$$

We use meaning preservation and WER to *align* the model

Proximal Policy Optimization

$$R(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{y}^*) := \gamma \cdot \mathrm{MP}(\boldsymbol{y}, \boldsymbol{y}^*) + \ln\Big(1 - \mathtt{WER}(\boldsymbol{y}, \boldsymbol{y}^*)\Big)$$



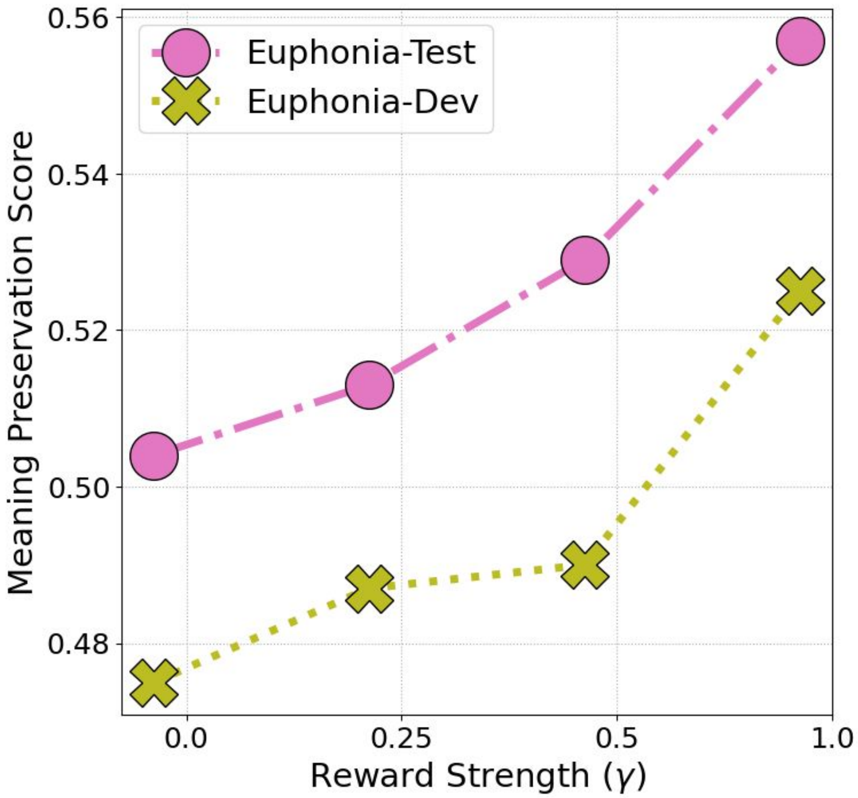
Setting A: WER Only

Setting B: WER+MP Score

Results

RLHF w/ MP Reward

- Significant improvement in MP.

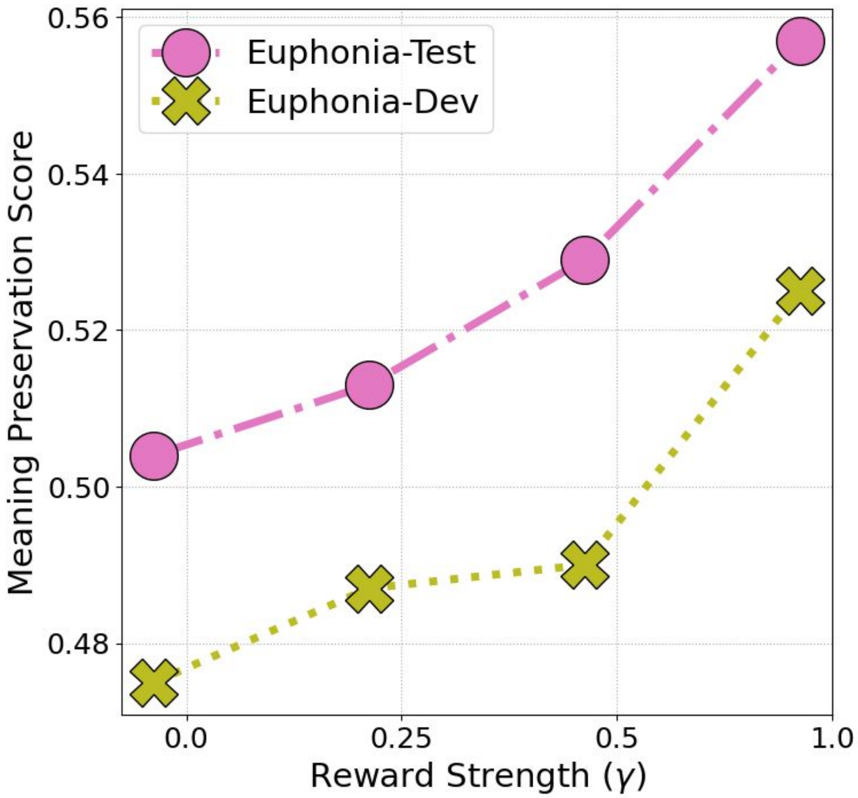


Results

RLHF w/ MP Reward

- Significant improvement in MP.
- No significant diff in WER.

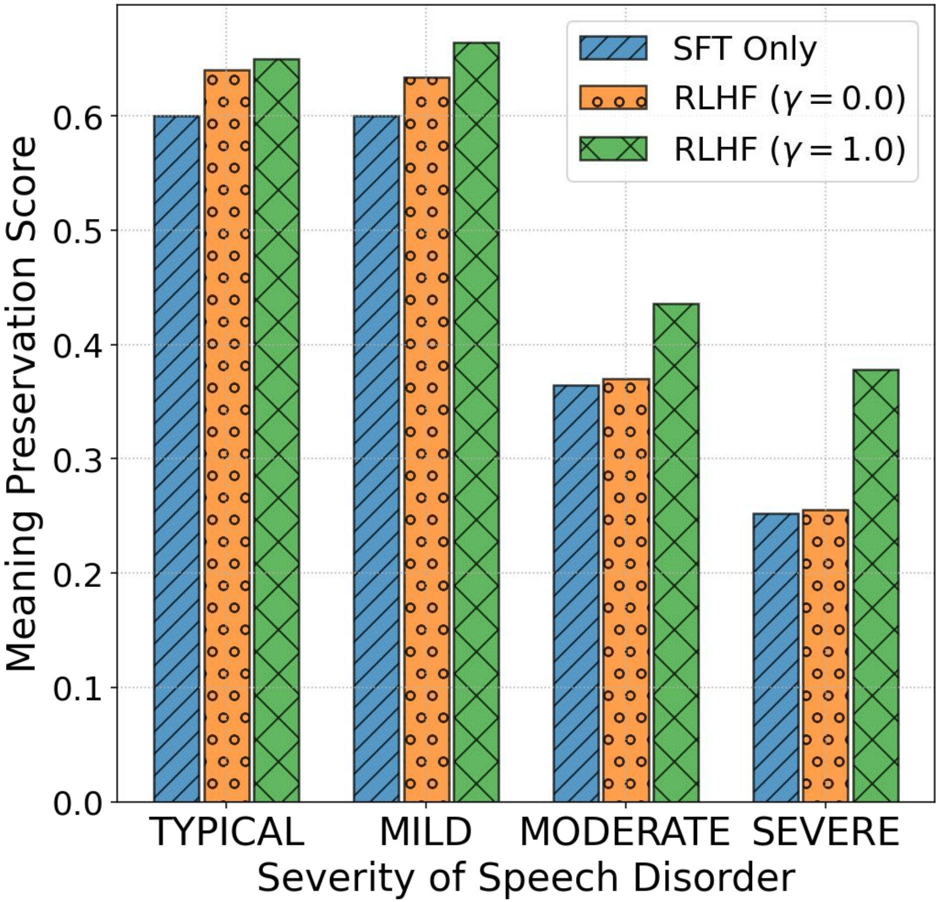
Tuning strategy	Euphonia Test WER ↓	Euphonia Test MP ↑	Euphonia Dev WER ↓	Euphonia Dev MP ↑	Librispeech Dev WER ↓	Librispeech Dev MP ↑
Base SFT model	50.4	48.2	47.3	48.1	17.2	85.6
Continued SFT	57.1	42.8	59.2	40.5	22.9	73.2
RLHF **WER** + **MP**						
WER ($\gamma = 0.00$)	**41.0**	50.4	**40.1**	47.5	**20.2**	75.7
+ **MP** ($\gamma = 0.25$)	41.7	51.3	41.7	48.7	22.4	74.7
+ **MP** ($\gamma = 0.50$)	41.2	52.9	41.1	49.0	23.9	72.2
+ **MP** ($\gamma = 1.00$)	42.6	**55.7***	42.9	**52.5***	22.0	**76.2***



Results

RLHF w/ MP Reward

- Significant improvement in MP.
- No significant diff in WER.
- **Gains more pronounced for more severe speech utterances.**

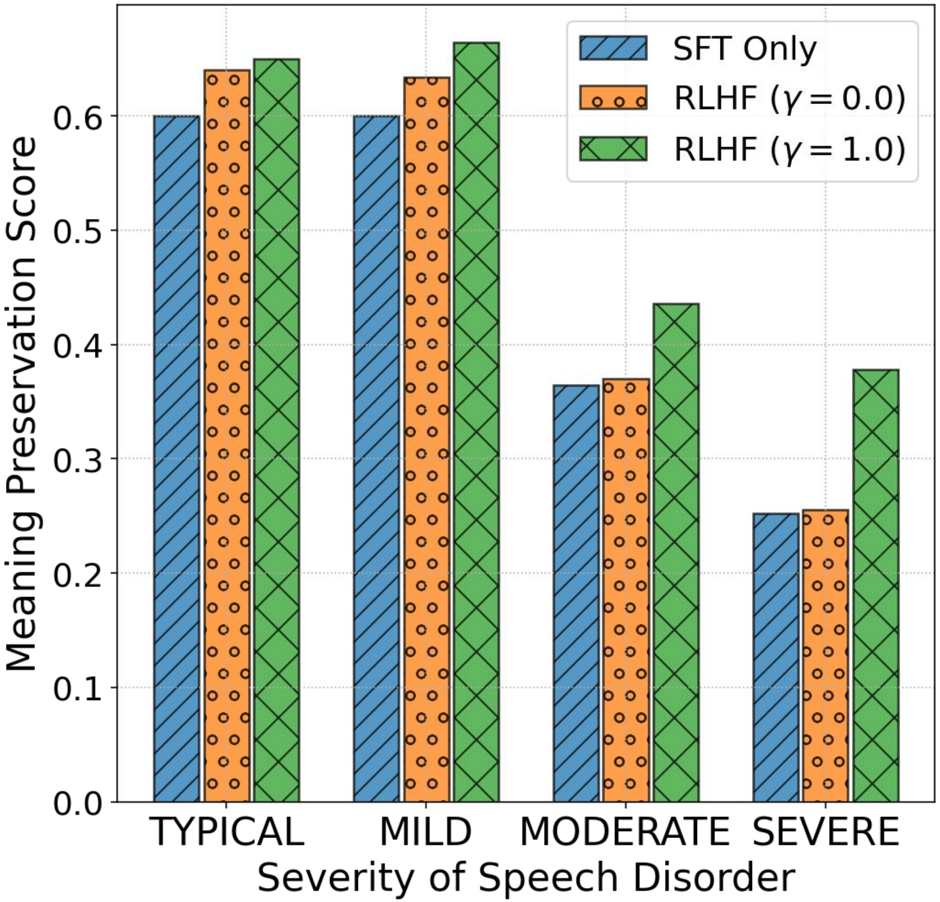


Results

Human Eval

- Significant correlation with auto-eval.
- Significant gain in MP.

Statistic (# samples = 220)	$\gamma = 0.0$	$\gamma = 1.0$
Average Primary Assessment (Human MP)	29.10%	40.45%
Accuracy (Human vs. Model MP)	85.90%	81.36%
Spearman (ρ) (Human vs. Model MP)	0.684*	0.639*



Examples

TABLE II: Examples selected based on human evaluation of transcripts on meaning preservation and error type of the RLHF models show that trading-off WER slightly for a significant gain in MP score ($\gamma = 1.00$) leads to better predictions overall.

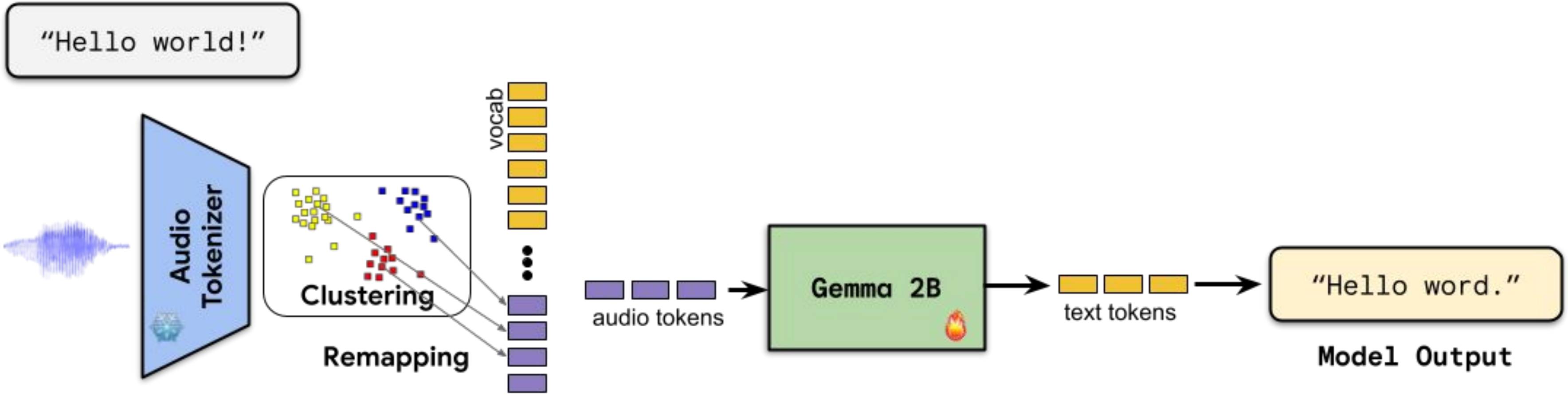
Ground Truth	Severity	RLHF ($\gamma = 0.0$)	WER	RLHF ($\gamma = 1.0$)	WER
”not so good today”	MILD	”not so good to the.”	(0.5)	”not so good to day.”	(0.5)
”every one of my family listens to music”	MODERATE	“every once in my frame and listen to music”	(0.62)	“everybody in my family listens to music”	(0.38)
“dancing is so much fun”	MODERATE	“that’s so much fun.”	(0.40)	“dancing so much fun.”	(0.20)
“are you comfortable?”	MODERATE	“are you going to school?”	(1.0)	“are you comfortable with it?”	(0.67)
“happy birthday dear friend.”	SEVERE	“absolutely your friend.”	(0.75)	“happy birthday to your friend.”	(0.50)
“as soon as possible”	SEVERE	“it soon adds pounds him volume”	(1.0)	“a soon as possible.”	(0.25)

WER alone as reward.

MP + WER together as reward does best.

Summary

- LLMs can be modified to recognize speech.
- SFT on a mix standard and disordered speech datasets helps.



Summary

- LLMs can be modified to recognize speech.
- SFT on a mix standard and disordered speech datasets helps.
- RL can help further generalize the model on disordered speech.
- Combination of Meaning Preservation and WER as reward signal works best.