# TranscribeX: LLM-Enhanced ASR Transcription


**Grace LeFevre**

Summer 2024 PhD Intern, RAILS @ Vail (https://freeclimb.com/research)


RTC VoiceTech Talk

*October 8, 2024*

# Motivation

- When transcribing telephony audio, ASR engines often produce noisy outputs with high word error rates (WER)
- This impacts the efficacy of downstream analyses that process this transcribed text (intent determination, sentiment analysis, data aggregation, etc.)

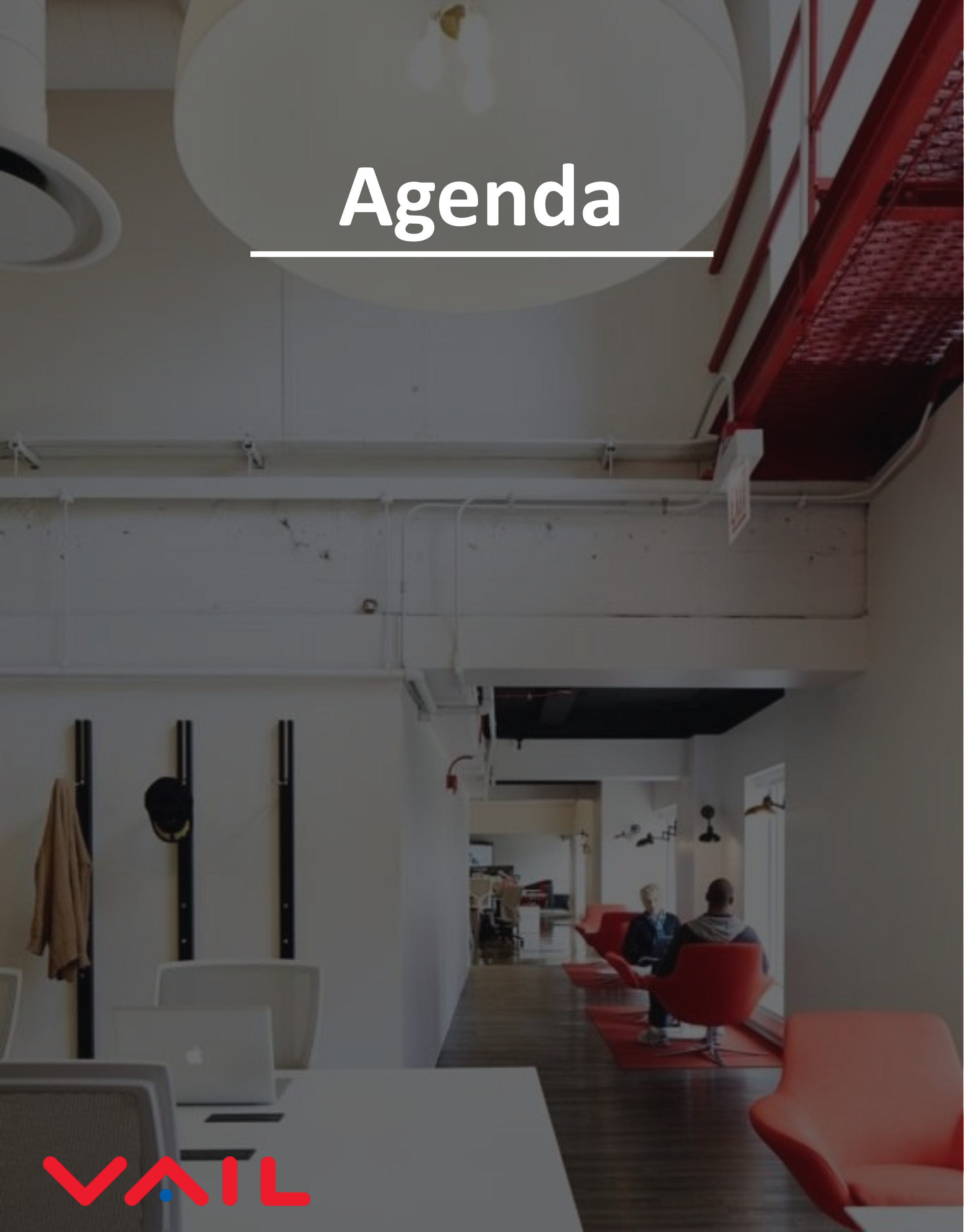
# Project Goals

**Main goal:** explore possibilities for improving telephony-based ASR transcripts using Large Language Models (LLMs)

# Project Goals

**Main goal:** explore possibilities for improving telephony-based ASR transcripts using Large Language Models (LLMs)

- Try a range of possible methods using any combinations of ASRs and/or LLMs
- Not operating under practical constraints
- Will inform us on how to create a system that can perform these edits under practical constraints



# Agenda

1. **Model Overview**
2. **Experiment #1**
    a. Common ASR errors
    b. LLM Choice method
    c. Targeted improvements
3. **Experiment #2**
    a. Dataset distribution
    b. Targeted improvements
4. **Summary**

# Model Overview

ASR models:

- Whisper
  - OpenAI's generative transformer-based model
- Speechmatics
  - Traditional ASR that uses an acoustic model and a language model
- Google telephony
  - Google Cloud's speech-to-text model specifically trained for transcribing telephony audio

LLM: Meta's Llama-3-70B

VAIL

# Experiment #1

Dataset #1

- 911 short audio files
- Domain: customer experience surveys

Goal: develop a method that uses an LLM to improve transcription quality for this dataset

# Measuring ASR Performance

- ASR performance is measured using Word Error Rate (WER), the ratio of errors to total words in a transcript
  - A lower WER indicates better performance
  - A transcript with no errors has a WER of 0%

# ASR Performance

| ASR | WER |
|---|---|
| Whisper | 10.8% |
| Speechmatics | 15.8% |
| Google telephony | 12.1% |

# ASR Performance

| ASR | WER |
|---|---|
| Whisper | 10.8% |
| Speechmatics | 15.8% |
| Google telephony | 12.1% |

- Decent performance overall but ASRs still make significant errors
- Room for potential LLM improvement

# ASR Errors: Examples

Most of the time, at least one ASR is correct (or *more* correct than the others)

# ASR Errors: Examples

Most of the time, at least one ASR is correct (or *more* correct than the others)

- Correct transcription: "she was very helpful and considerate"
- **Whisper: "she was very helpful and considerate"**
- Speechmatics: "she was very helpful, if you consider"
- Google telephony: "she was very helpful inconsiderate"

VAIL

# ASR Errors: Examples

Most of the time, at least one ASR is correct (or *more* correct than the others)

- Correct transcription: "she answered all my questions in a reasonable manner, very *politely* and on time"
- Whisper: "she answered all my questions in a reasonable manner, very *quietly* and on time"
- **Speechmatics: "she answered all my questions in a reasonable manner, very *politely* and on time"**
- Google telephony: "she answered all my questions in a reasonable manner, very *poorly* and on time"

# ASR Errors: Examples

Most of the time, at least one ASR is correct (or *more* correct than the others)

- Correct transcription: “very helpful”
- Whisper: “It’s very careful”
- Speechmatics: “They were here for”
- **Google telephony: “very helpful”**

# “Best choice” performance

- Calculated by taking the best-performing (lowest WER) ASR transcript for each document in the dataset

# “Best choice” performance

- Calculated by taking the best-performing (lowest WER) ASR transcript for each document in the dataset
- Taking overall WER of “best” transcripts gives the *empirical minimum WER* for this dataset

# "Best choice" performance

- Calculated by taking the best-performing (lowest WER) ASR transcript for each document in the dataset
- Taking overall WER of "best" transcripts gives the empirical minimum WER for this dataset

| | |
|---|---|
| Whisper | 10.8% |
| Speechmatics | 15.8% |
| Google telephony | 12.1% |
| *Empirical minimum* | *7.4%* |

# LLM Choice Method

Taking the “best choice” transcription for each documents requires pre-existing knowledge of ground truth

LLM choice method: Given multiple ASR transcriptions, can an LLM choose the best one?

# LLM Choice Method: prompt

Prompt includes descriptions of:

```
You are a helpful transcription error correction assistant. I have a telephony dataset consisting of customers answering survey questions about their experience speaking to a customer service representative. I transcribed an audio file from this dataset using three Automatic Speech Recognition models. The first ASR model is Whisper, a generative transformer-based model. The second ASR model is Speechmatics, a traditional ASR that uses an acoustic model and a language model. The third ASR model is a Google Cloud model trained to transcribe telephony audio. Overall, Whisper is the best performing model and Speechmatics is the worst performing model, but all three models make mistakes sometimes. … (cont.)
```

# LLM Choice Method: prompt

Prompt includes descriptions of:

- **dataset domain**

You are a helpful transcription error correction assistant. I have **a telephony dataset consisting of customers answering survey questions about their experience speaking to a customer service representative**. I transcribed an audio file from this dataset using three Automatic Speech Recognition models. The first ASR model is Whisper, a generative transformer-based model. The second ASR model is Speechmatics, a traditional ASR that uses an acoustic model and a language model. The third ASR model is a Google Cloud model trained to transcribe telephony audio. Overall, Whisper is the best performing model and Speechmatics is the worst performing model, but all three models make mistakes sometimes. … (cont.)

# LLM Choice Method: prompt

Prompt includes descriptions of:

- dataset domain
- **ASR models**

You are a helpful transcription error correction assistant. I have a telephony dataset consisting of customers answering survey questions about their experience speaking to a customer service representative. I transcribed an audio file from this dataset using three Automatic Speech Recognition models. **The first ASR model is Whisper, a generative transformer-based model. The second ASR model is Speechmatics, a traditional ASR that uses an acoustic model and a language model. The third ASR model is a Google Cloud model trained to transcribe telephony audio.** Overall, Whisper is the best performing model and Speechmatics is the worst performing model, but all three models make mistakes sometimes. … (cont.)

VAIL

# LLM Choice Method: prompt

Prompt includes descriptions of:

- dataset domain
- ASR models
- **comparative ASR performance on dataset**

```
You are a helpful transcription error correction assistant. I have a telephony dataset consisting of customers answering survey questions about their experience speaking to a customer service representative. I transcribed an audio file from this dataset using three Automatic Speech Recognition models. The first ASR model is Whisper, a generative transformer-based model. The second ASR model is Speechmatics, a traditional ASR that uses an acoustic model and a language model. The third ASR model is a Google Cloud model trained to transcribe telephony audio. Overall, Whisper is the best performing model and Speechmatics is the worst performing model, but all three models make mistakes sometimes. … (cont.)
```

# LLM Choice Method: prompt

Prompt directs LLM to:

```
Given the transcriptions produced by these ASR models, your task is to choose which transcription you think is most likely to be the correct transcription. A correct transcription should be semantically coherent, fit the customer service survey context described above, and stick as closely as possible to the content of the original audio file. It is likely that all the transcriptions contain inaccuracies, but please choose the one you think is most correct.

…

<formatting instructions>

<ASR transcriptions>
```

# LLM Choice Method: prompt

Prompt directs LLM to:

- **Choose the ASR transcription most likely to be true to the original audio file**

Given the transcriptions produced by these ASR models, your task is to **choose which transcription you think is most likely to be the correct transcription.** A correct transcription should be semantically coherent, fit the customer service survey context described above, and **stick as closely as possible to the content of the original audio file**. It is likely that all the transcriptions contain inaccuracies, but please choose the one you think is most correct.

…

<formatting instructions>

<ASR transcriptions>

VAIL

# Results

All data:

| | |
|---|---|
| Whisper | 10.8% |
| Speechmatics | 15.8% |
| Google telephony | 12.1% |
| **LLM choice** | **9.1%** |
| *Empirical minimum* | *7.4%* |

- LLM choice method achieves an overall WER improvement on this dataset!

VAIL

# Results

All data:

| | |
|---|---|
| Whisper | 10.8% |
| Speechmatics | 15.8% |
| Google telephony | 12.1% |
| **LLM choice** | **9.1%** |
| *Empirical minimum* | *7.4%* |

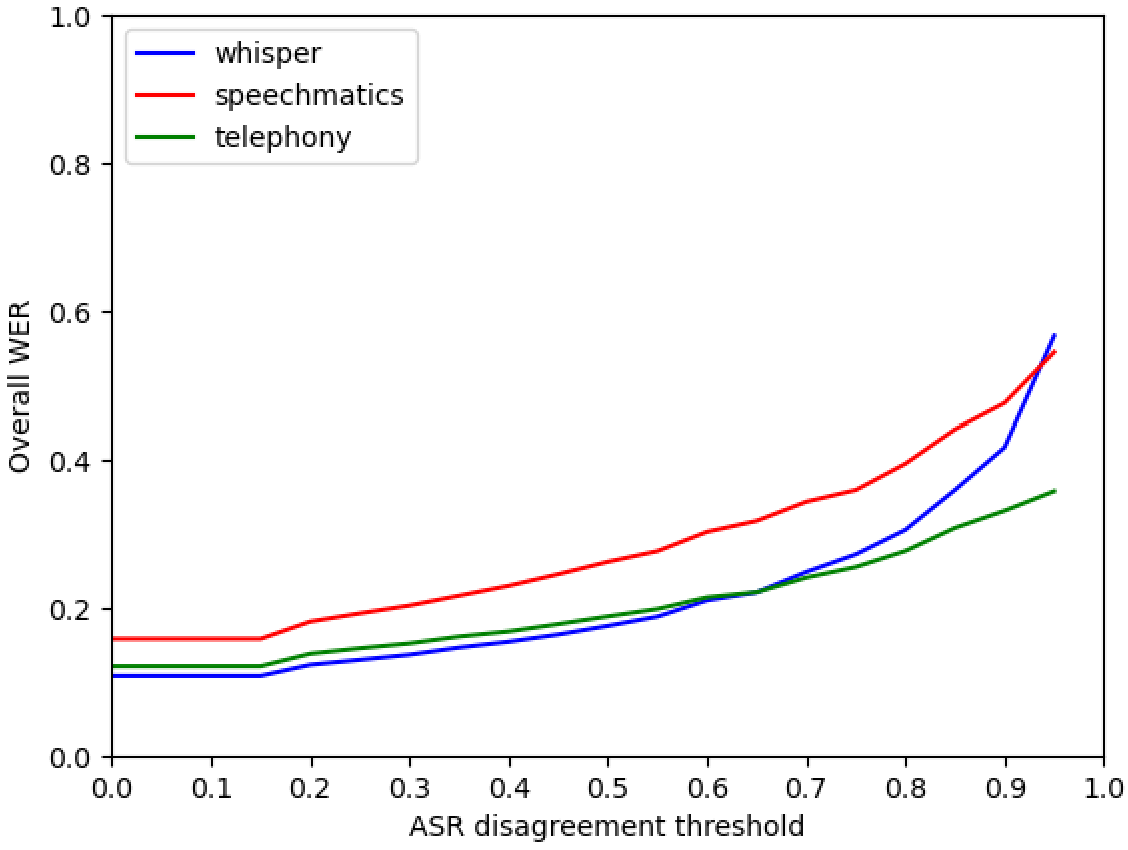
- LLM choice method achieves an overall WER improvement on this dataset!
- Can we maximize the method's usefulness by targeting specific documents for LLM improvement?

# Targeted Improvement Strategy

- For this dataset, **ASR disagreement** effectively measures transcription quality
- *The more the ASR transcriptions disagree, the less accurate they all are overall*

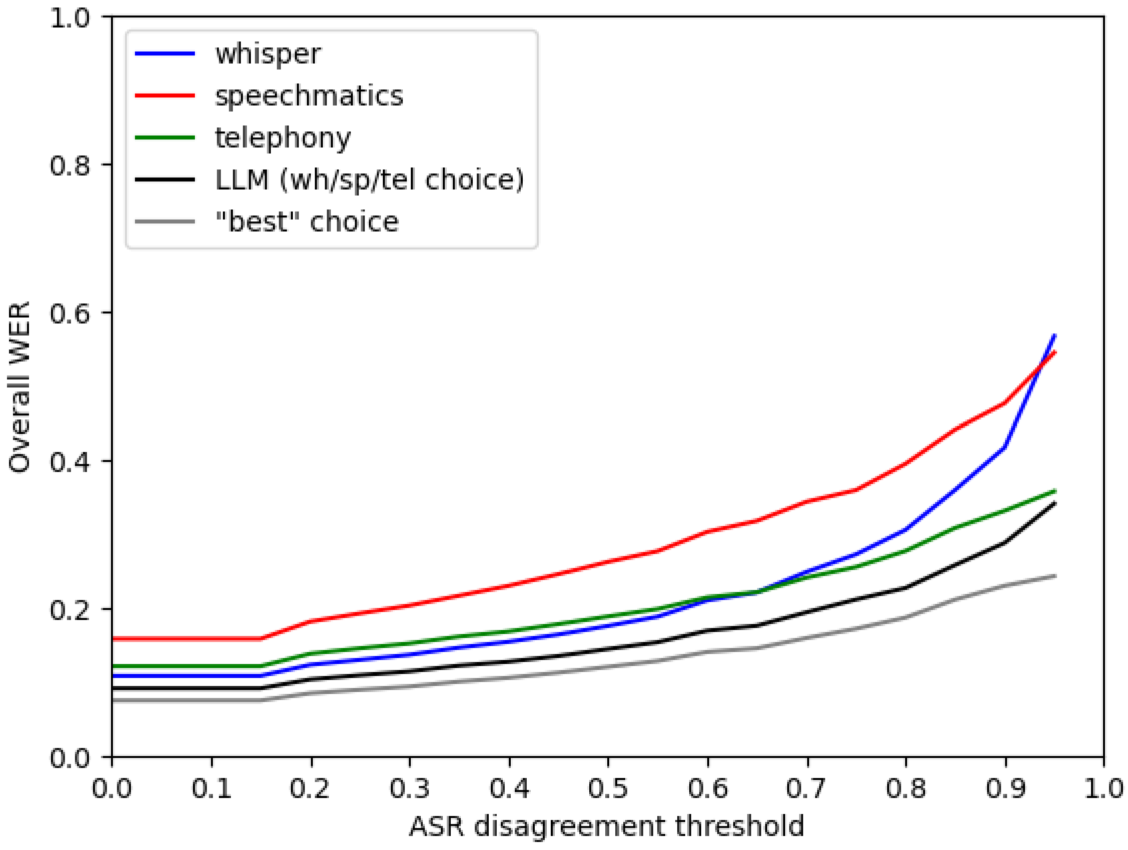
# Targeted Improvement Strategy

- For this dataset, **ASR disagreement** effectively measures transcription quality
- *The more the ASR transcriptions disagree, the less accurate they all are overall*



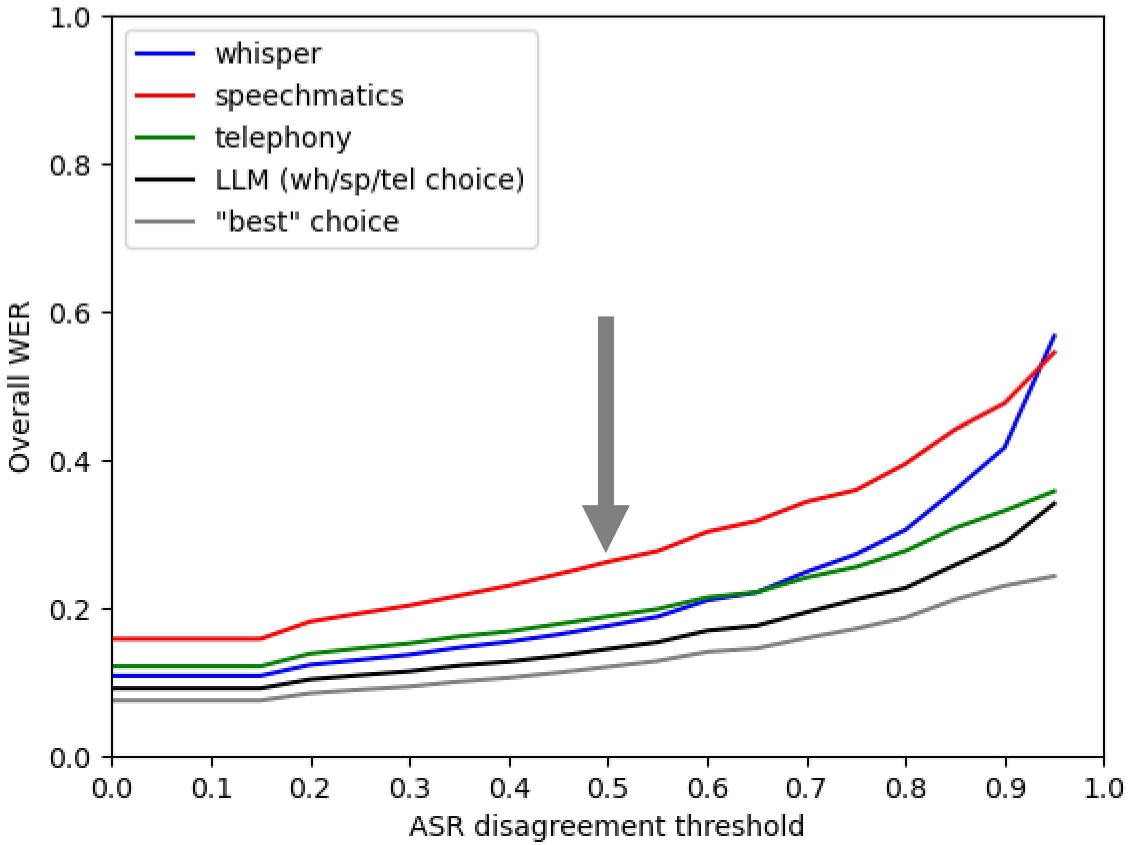
# Targeted Improvement Strategy

- Documents with higher ASR disagreement benefit more from the LLM choice method



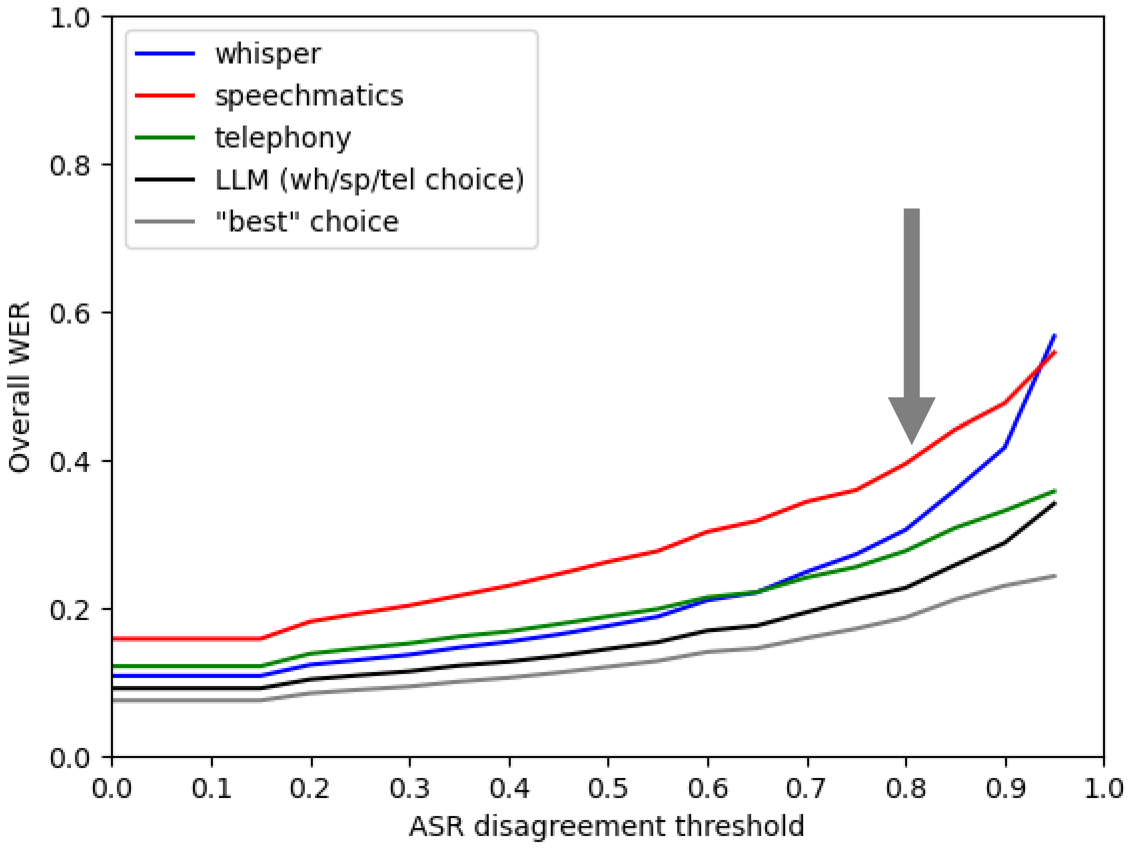
# Results: Targeted Improvement

| | |
|---|---|
| Whisper | 17.5% |
| Speechmatics | 26.2% |
| Google telephony | 18.8% |
| **LLM choice** | **14.4%** |
| *Empirical minimum* | *12.0%* |



# Results: Targeted Improvement

| | |
|---|---|
| Whisper | 30.5% |
| Speechmatics | 39.5% |
| Google telephony | 27.7% |
| **LLM choice** | **22.7%** |
| *Empirical minimum* | *18.7%* |



# Experiment #1 Summary

- Proof-of-concept
  - LLM choice method achieved performance improvement on this dataset
- Targeted improvement strategy
  - Focus on documents with high ASR disagreement

# Bonus Results: 2 ASR Choice Method

Q: Does the method still work

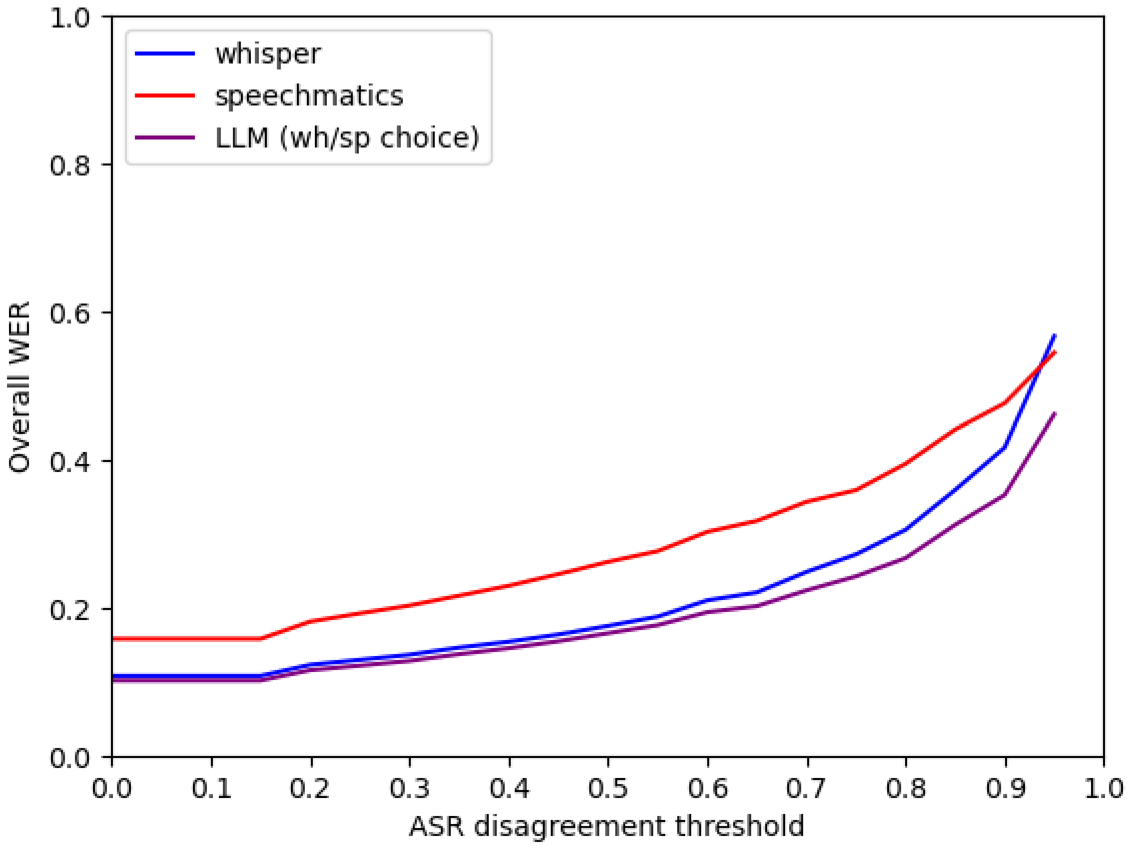
if we only provide the LLM two

ASR options to choose from?

# Bonus Results: 2 ASR Choice Method

Q: Does the method still work if we only provide the LLM two ASR options to choose from?

A: Yes, but less well



# Experiment #2

<u>Dataset #2</u>

- 918 short audio files
- Same domain but different distribution than dataset #1

<u>Goal:</u> test LLM choice method's performance on a dataset with a different distribution

# Dataset #1 vs #2 Distributions

Different wordcount distributions

- Dataset #1: median word count = 16 words
- Dataset #2: median word count = 7 words

# Dataset #1 vs #2 Distributions

Different wordcount distributions

- Dataset #1: median word count = 16 words
- Dataset #2: median word count = 7 words

Different ASR performance trends

- On dataset #1, Whisper was the highest-performing ASR and Speechmatics was the lowest-performing ASR
- On dataset #2, this trend reverses & Speechmatics is highest-performing ASR

# Results

All data

| | |
|---|---|
| Whisper | 15.1% |
| Speechmatics | 12.1% |
| Google telephony | 15.2% |

# Results

<u>All data</u>

| | |
|---|---|
| Whisper | 15.1% |
| Speechmatics | 12.1% |
| Google telephony | 15.2% |
| **LLM choice** | **11.9%** |
| *Empirical minimum* | *7.8%* |

- LLM choice method achieves minimal overall improvement on this dataset

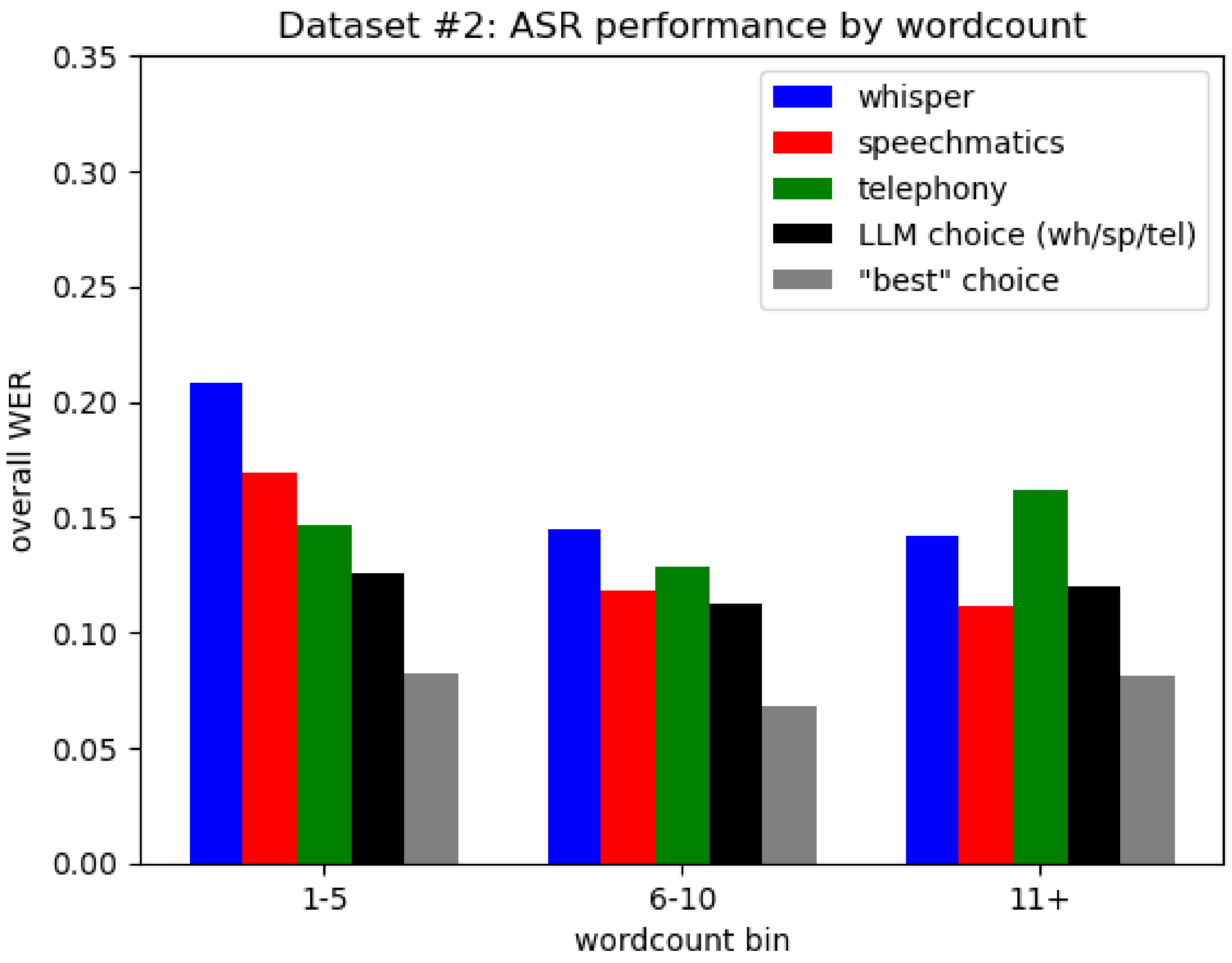
# Results

| All data | |
|---|---|
| Whisper | 15.1% |
| Speechmatics | 12.1% |
| Google telephony | 15.2% |
| **LLM choice** | **11.9%** |
| *Empirical minimum* | *7.8%* |

- LLM choice method achieves minimal overall improvement on this dataset
- The method is more effective for shorter documents than for longer ones

VAIL

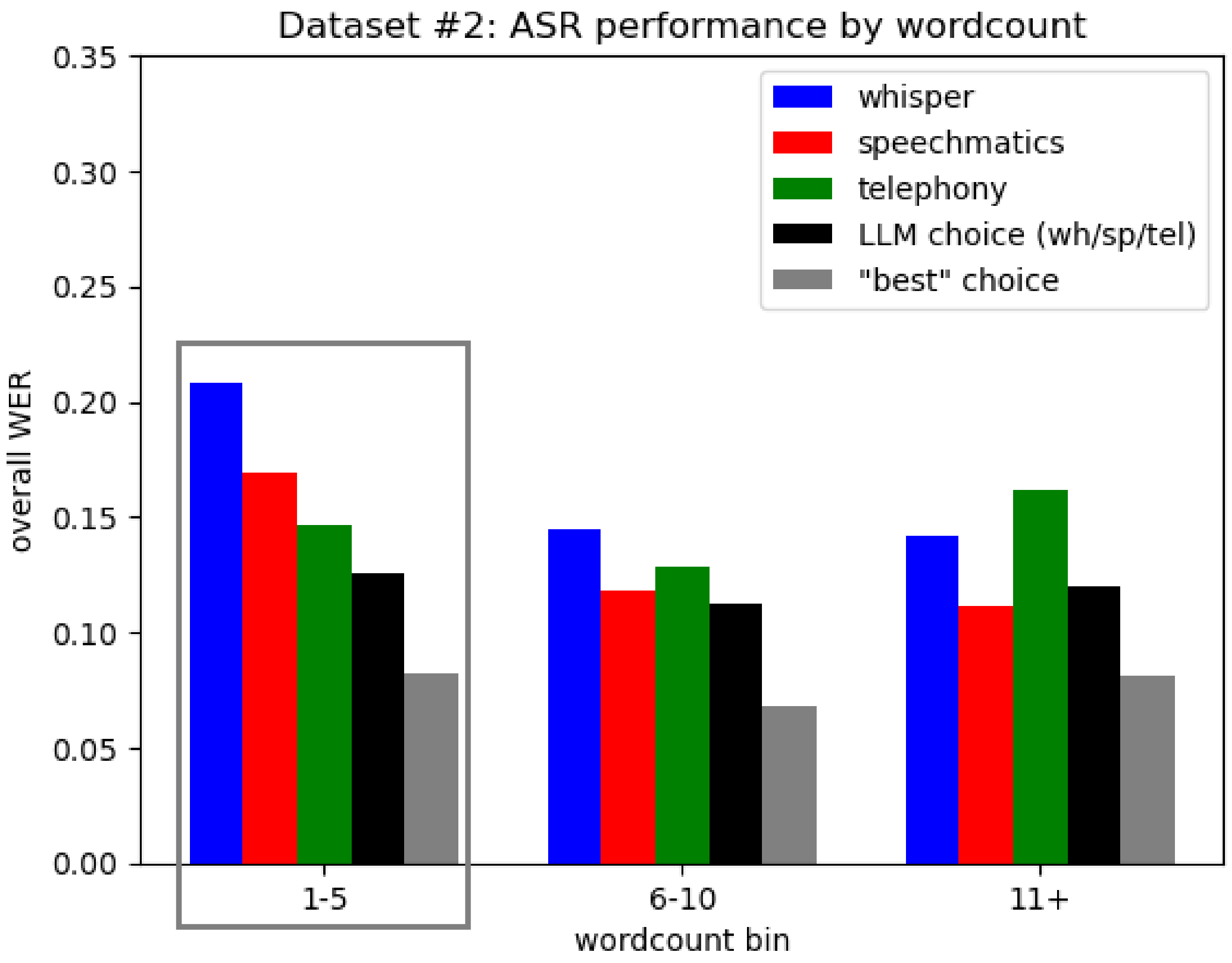
# Results: Targeted Improvement



# Results: Targeted Improvement

For 1-5 word documents:

| | |
|---|---|
| Whisper | 20.8% |
| Speechmatics | 16.9% |
| Google telephony | 14.6% |
| **LLM choice** | **12.6%** |
| *Empirical minimum* | *8.2%* |



# Experiment #2 Summary

- LLM choice method improved performance on short documents in this dataset
- Even within the same domain, a dataset with a different distribution may require a different targeted improvement strategy to benefit from LLM enhancement

# Takeaways & Future Work

- Small proof-of-concept that ASR transcriptions of telephony audio can be improved via LLM choice method
- Targeted transcription improvement using an LLM enhancement method requires strategy specific to both domain and distribution of dataset
- Ongoing work exploring other LLM enhancement methods


# Thank you!