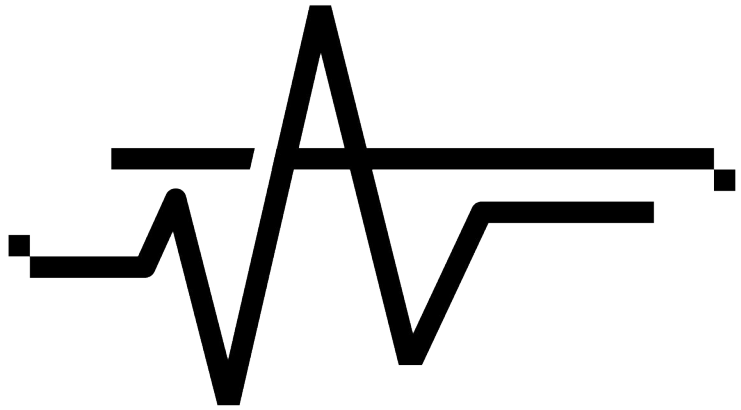


# ALVENIR

SPEECH TECHNOLOGIES

# CoRal project: Speech-to-text models

## Presentation Outline

1. About me and Alvenir
2. Introduction to the CoRal project
3. Performance of the Røst-wav2vec2-v2 models
4. Comparison of Wav2vec2 and Whisper errors
5. Dialect experiment on Wav2vec2 models

ALVENIR
SPEECH TECHNOLOGIES

# About me

*Marie Juhl Jørgensen*

- Junior Machine Learning Engineer at Alvenir
- B.Sc. in Artificial Intelligence and Data Science from DTU
- Pursuing a M.Sc. in Human-Centered Artificial Intelligence at DTU
  - Currently writing my thesis:

    *Physics Constrained Diffusion Models for Climate related Forecasting*

# About Alvenir

We are a spin-out from DTU and continue to collaborate with leading institutes

We deliver various AI-solutions within speech to text technologies

We are specialized in assisting advisors in the financial sector

ALVENIR
SPEECH TECHNOLOGIES

# About Alvenir

*Speech Technologies for the financial sector*

BØRSEN

## Velliv satser på AI-agenter: Vil frigøre tid og styrke kundemødet

*Tina Dupont, direktør for Kunder og Indsigter i Velliv*
*Foto: Velliv*

MERETE SKOV PEDERSEN
22. AUG. 2025 KL. 16.47
ANNONCØRBETALT INDHOLD AF ALVENIR APS

**Den finansielle sektor har længe været frontløber for digitalisering. Nu står branchen over for næste bølge: AI-agenter. Velliv viser vejen i samarbejde med AI-virksomheden Alvenir.**

**Online ASR**
(Automatic Speech Recognition)

**Live Summarization**

**Specialized Agentic AI**

# The CoRal project

*Making speech technologies accessible for all dialects in Denmark*

- 1000 + hours of speech
- Open-source models for speech recognition and generating synthetic speech

# The CoRal project

The project has released CoRal-v2 dataset

- Read-aloud train dataset: 250,000 samples
- Conversation train dataset: 43,000 samples

We train and publish open-source models for speech-to-text (named røst) and text-to-speech (

The speech-to-text models are evaluated using

- Word Error Rates (WER)

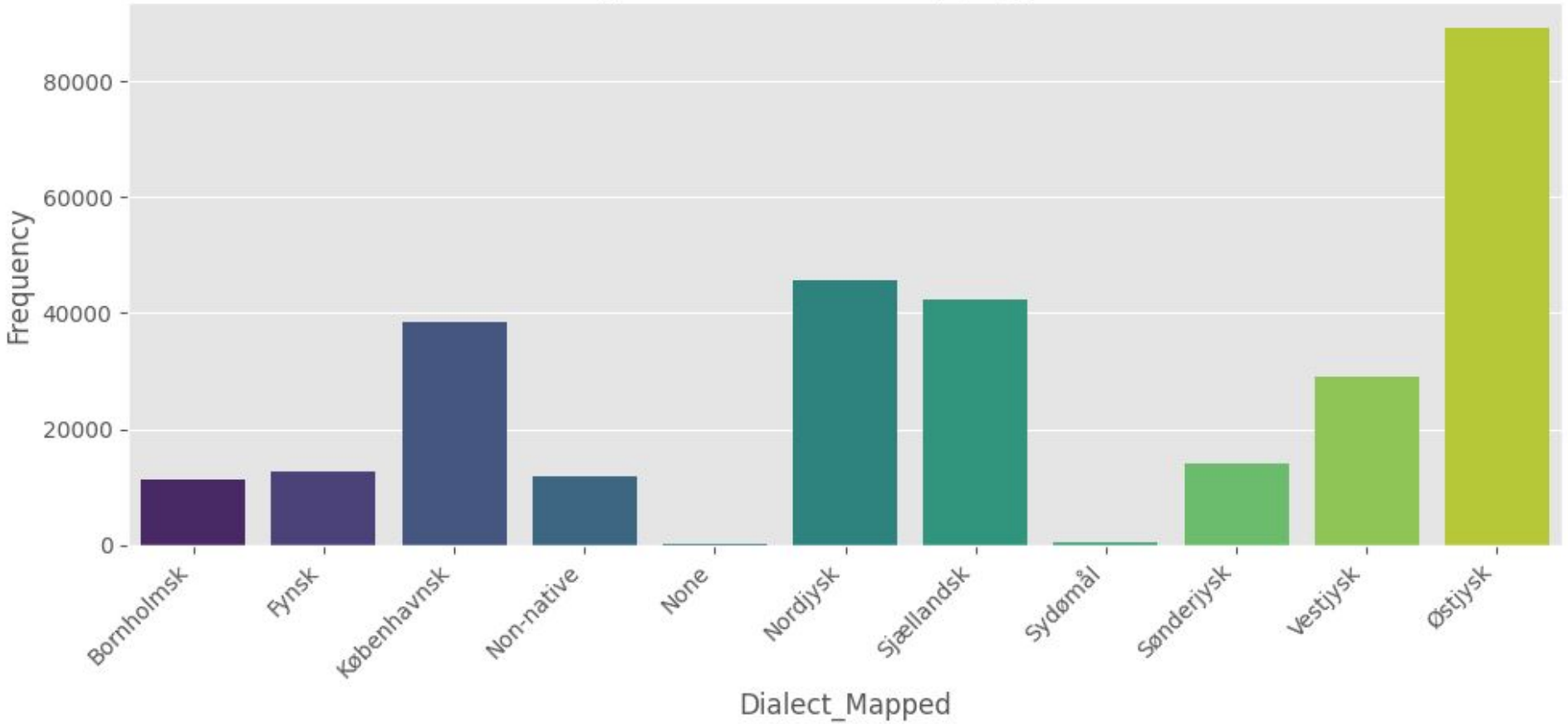
**CoRal-v3 Dataset and Models are released in the beginning of 2026**

Coral-v2 Conversation

Coral-v2 Read-aloud

# The CoRal data grouped by dialect

Coral-v2 Read-aloud and Conversation Data mapped onto dialect categories



# Dataset additions

## Alvenir-OSS

- Open subtitles
- Read-aloud
- Simulate conversational words

## Alvenir-Wiki

- Wikipedia, various topics
- Read-aloud
- Different speakers, different setting than CoRal

WIKIPEDIA
The Free Encyclopedia

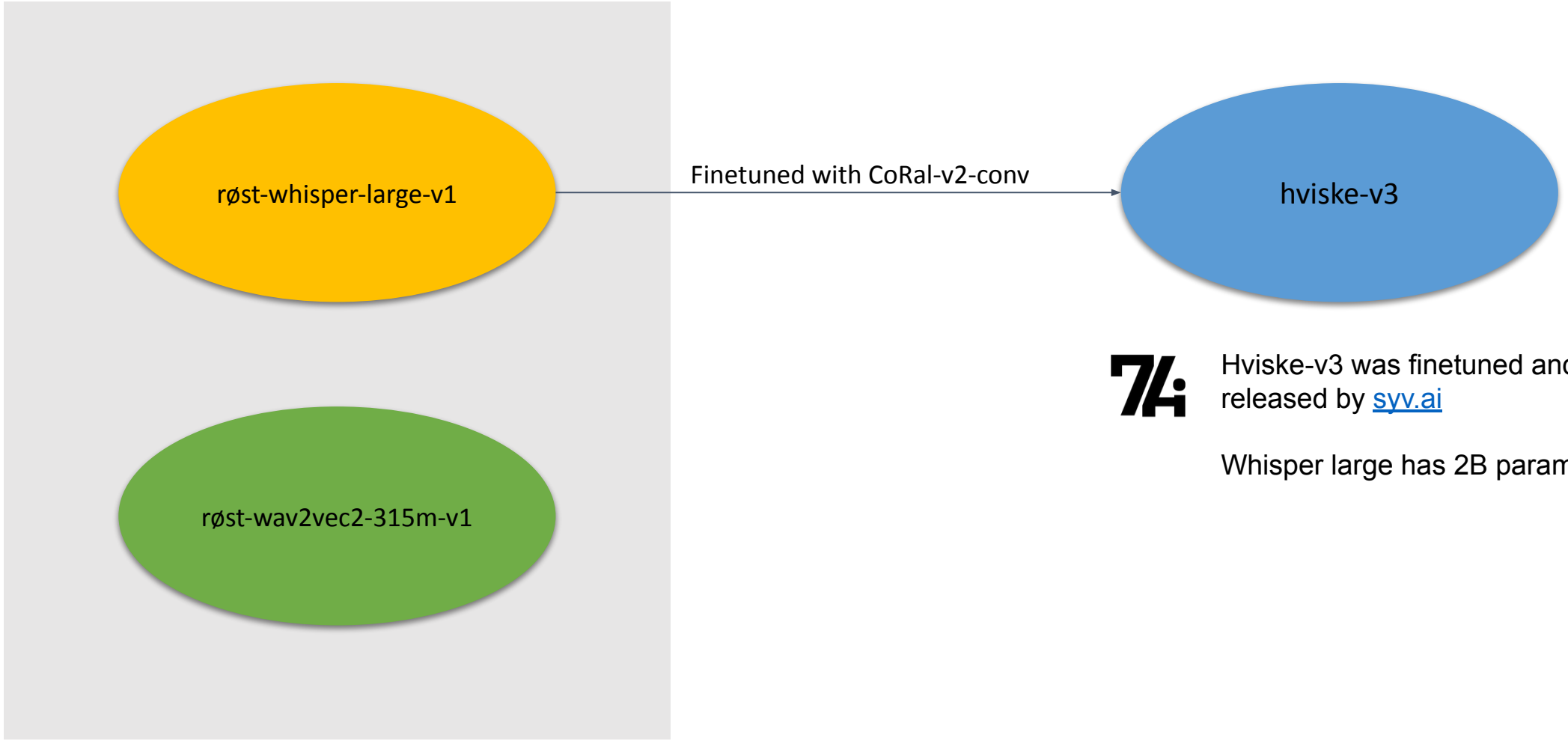


Link to both sets: https://huggingface.co/datasets/Alvenir/alvenir_asr_da_eval

# CoRal-v2 Wav2vec2 Model overview

# Models for comparison

**Release v1**




Hviske-v3 was finetuned and released by [syv.ai](syv.ai)

Whisper large has 2B parameters

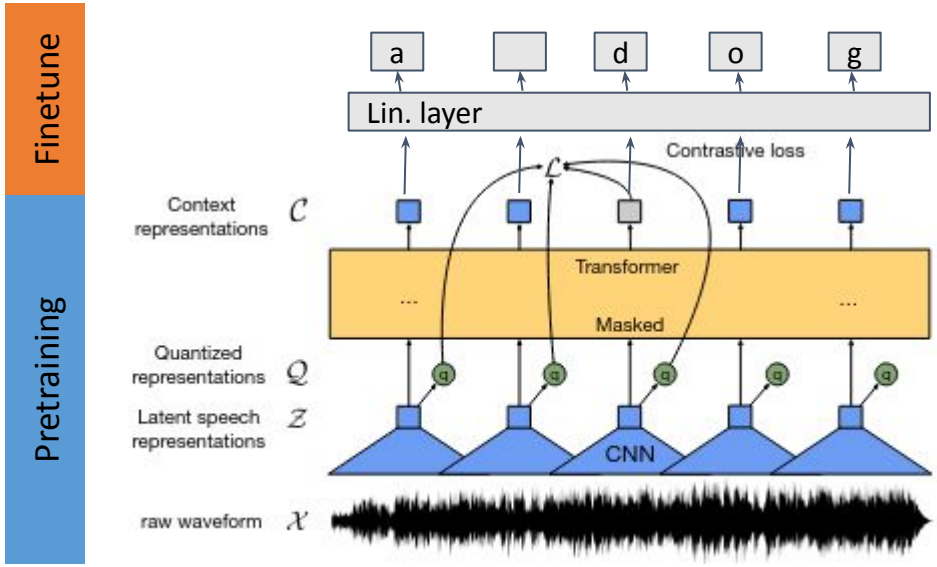
# Wav2vec2 and Whisper models

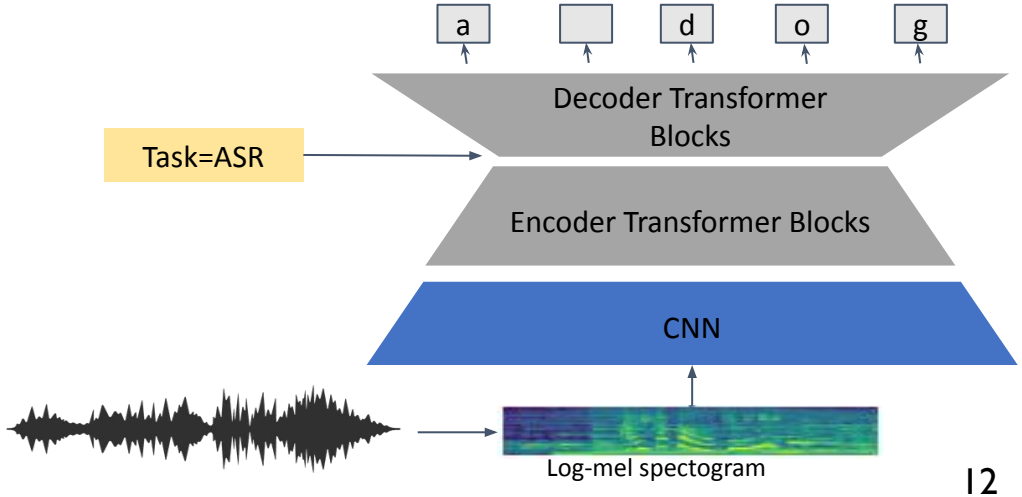
## Wav2Vec2

- Developed by Facebook AI (Meta)
- Self-supervised pretraining to learn features by masking, finetune linear layer to project to vocabulary
- For røst models: 3-gram Language Model is used to post-process

## Whisper

- Developed by OpenAI
- Trained for multiple tasks, ASR, translation, voice activity detection etc.
- End-to-end Transformer architecture: handles speech input directly to text output.



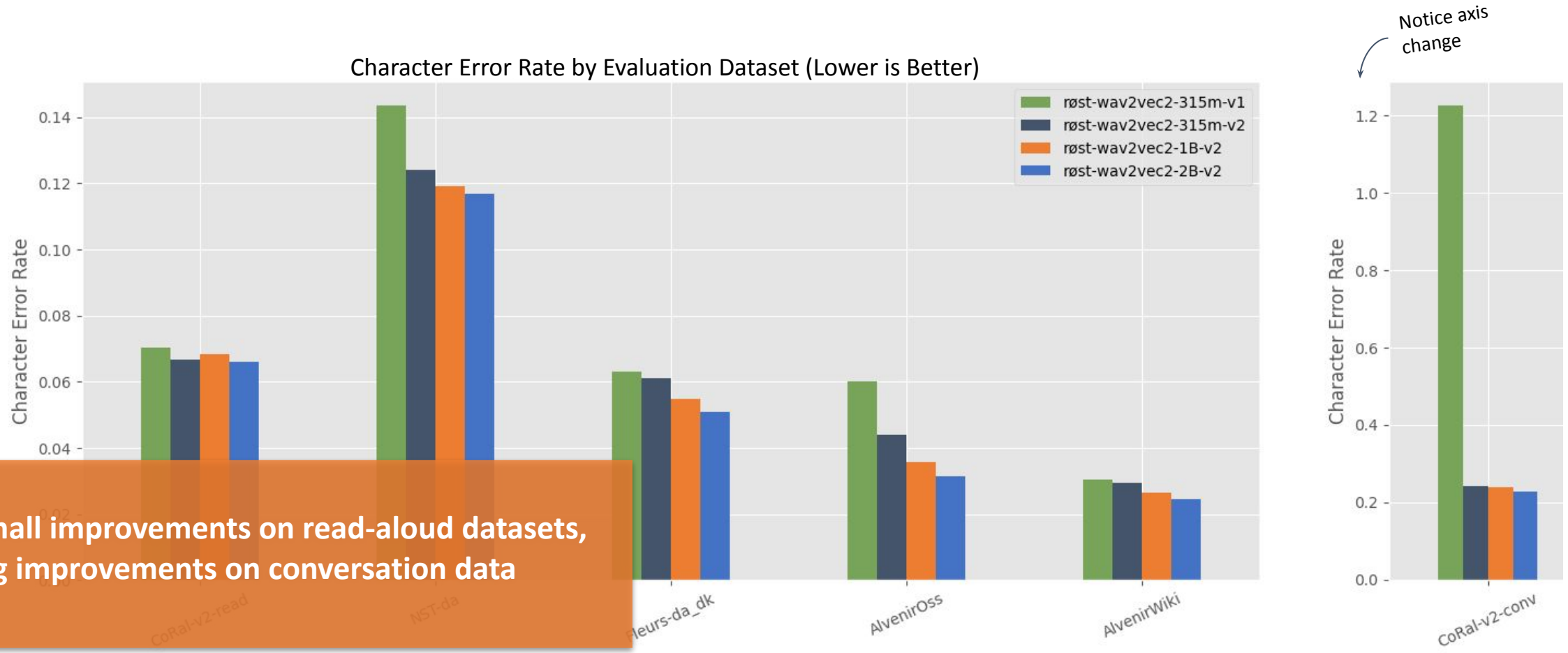






Wav2vec2 drawing from: Baevski et. al, *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*,2020

# Coral Wav2vec2 models - 2nd release

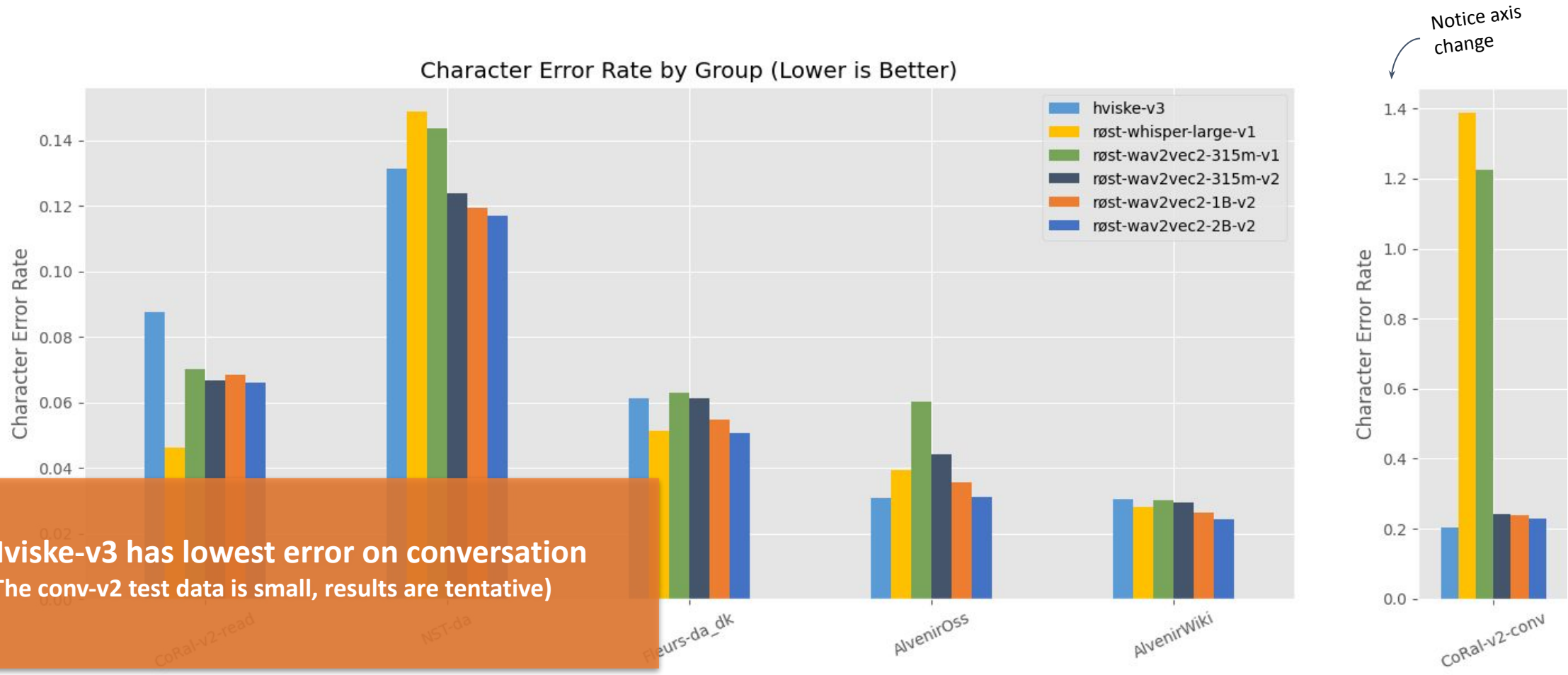
Comparing to 1st CoRal model release




**Small improvements on read-aloud datasets, big improvements on conversation data**

# Coral Wav2vec2 models - 2nd release

Comparing to Whisper models

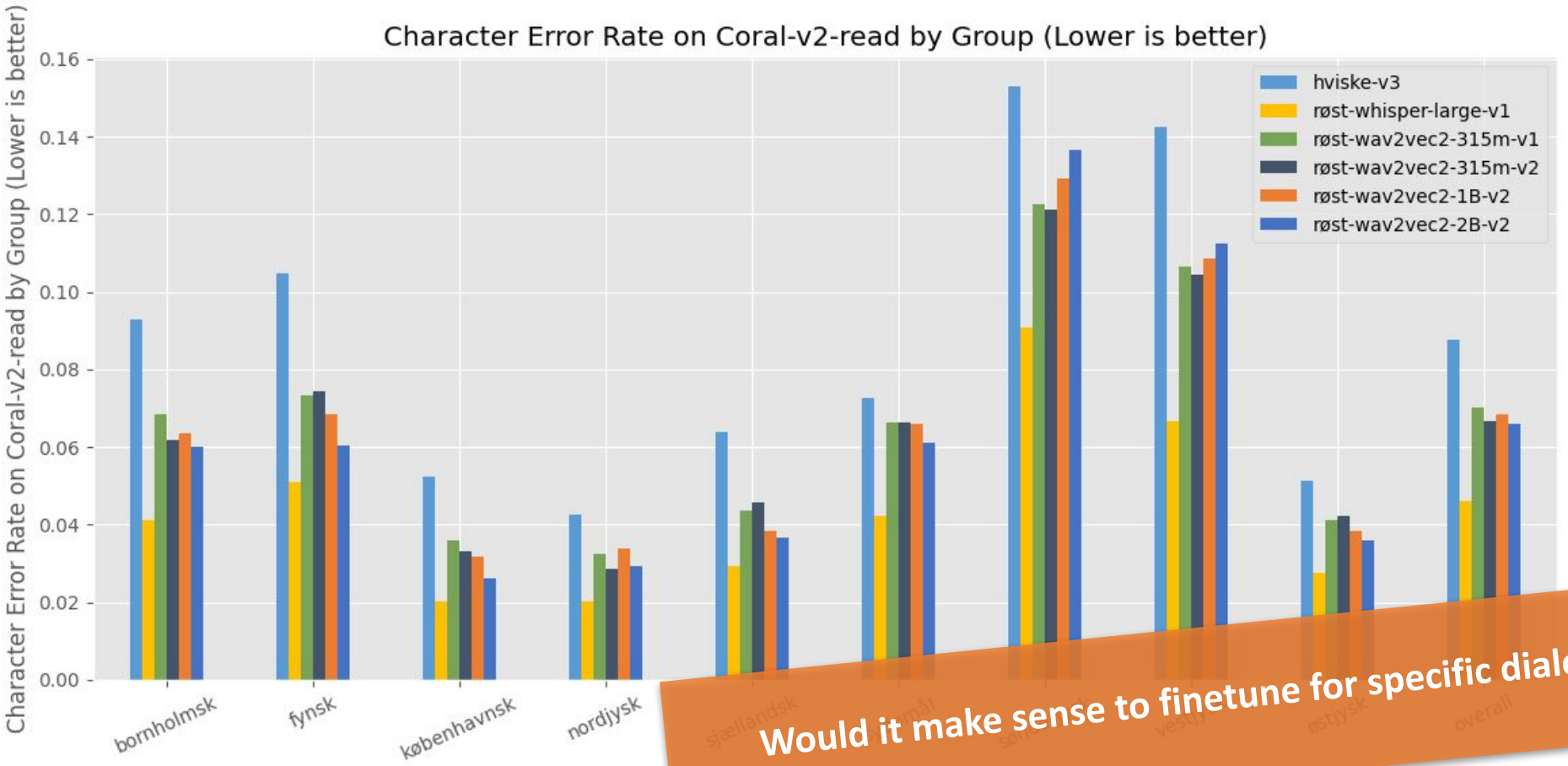



**Hviske-v3 has lowest error on conversation**

**(The conv-v2 test data is small, results are tentative)**

# Coral Wav2vec2 models - 2nd release

Significant difference in performance on dialects




**Would it make sense to finetune for specific dialects?**

# Wav2vec2 learn phonetic sounds while Whisper learns language

**Wav2vec2 Model Transcription:**
*det er produceret af thomas helme og indspillede i easy sound recording studio i københavn*

- **Character Error Rate (CER):** 4.4%
- **Word Error Rate (WER):** 13.3%

**Whisper Model Transcription:**
*det er produceret af thomas helmig og indspillet i easy sound recording studio i københavn*

- **Character Error Rate (CER):** 0%
- **Word Error Rate (WER):** 0%

**Can lead to inflated WER scores**

# When is Whisper under-performing?

**True sentence:**

det gør det ikke

ja

ja så ja

**Wav2vec2 Model Transcription:**

det gør det ikke

ja

ja ja
- CER: 37.5%, WER: 33.3%

**Whisper Model Transcription:**

det gør det ikke det gør det ikke
- CER: 106%, WER: 100%

det var der ikke så mange
- CER: 1200%, WER: 600%

der er jo rigtig mange frivillige dernede
- CER: 463%, WER: 233%

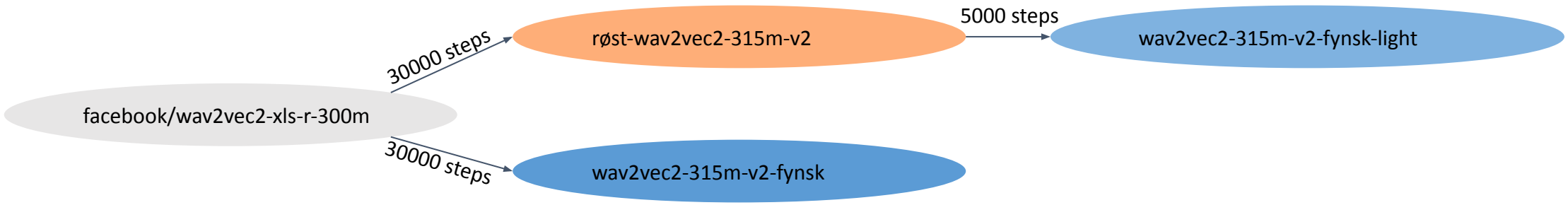
**Our experience: Whisper is less robust, since it can hallucinate and ignore audio**

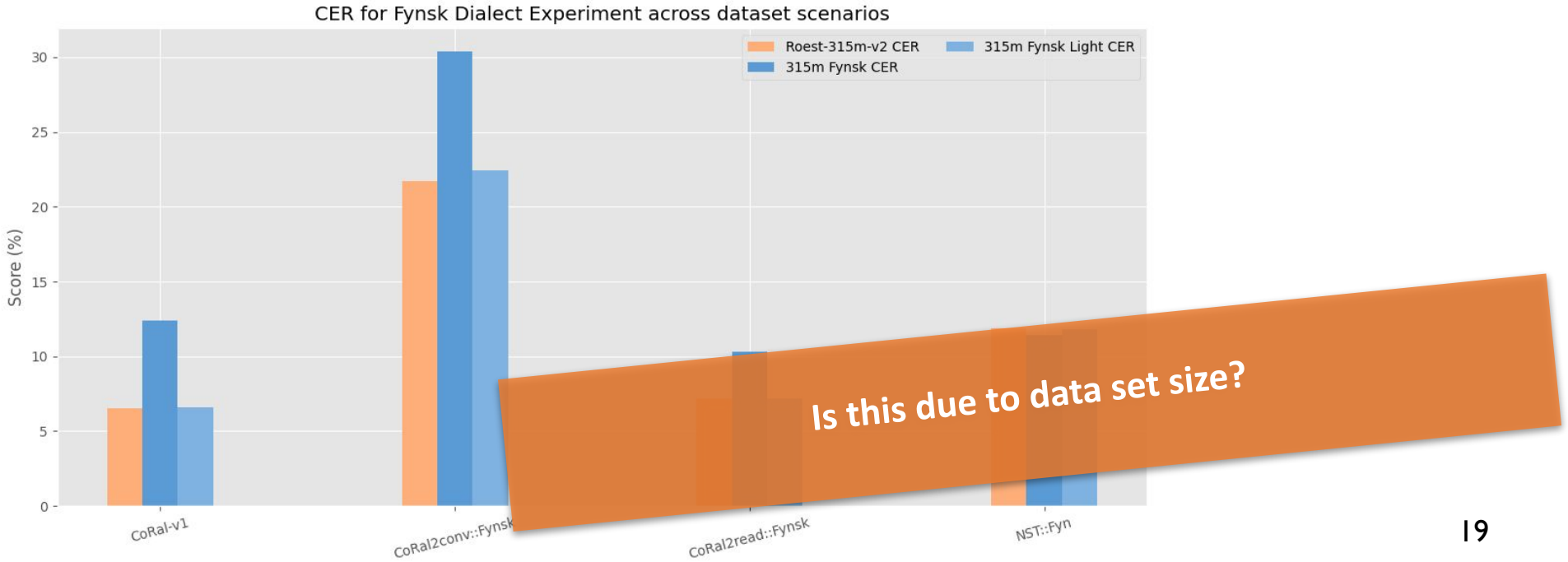
# Experiment

*Does training on a single dialect improve performance in that domain?*

ALVENIR
SPEECH TECHNOLOGIES

# Fynsk dialect models do not improve performance on dialect

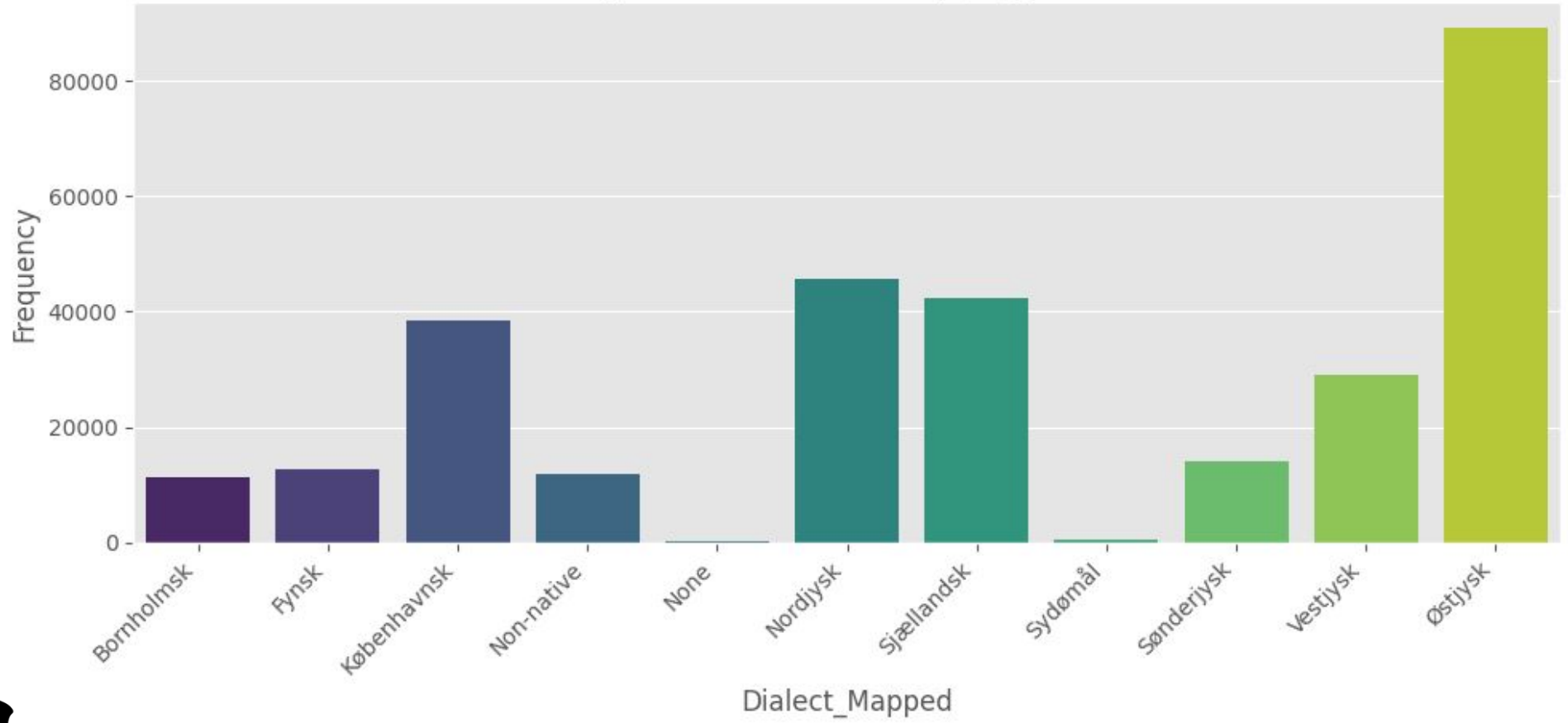



**Is this due to data set size?**

# The data is not evenly distributed across dialects

Coral-v2 Read-aloud and Conversation Data mapped onto dialect categories



# Extended Dialect Experiment

- How is performance affected if we hold the dataset size constant and train on diverse data vs dialect specific data?
  - **The "all" models are trained on a subset of data** randomly sampled so it contains the relative dialect distribution from CoRal-v2

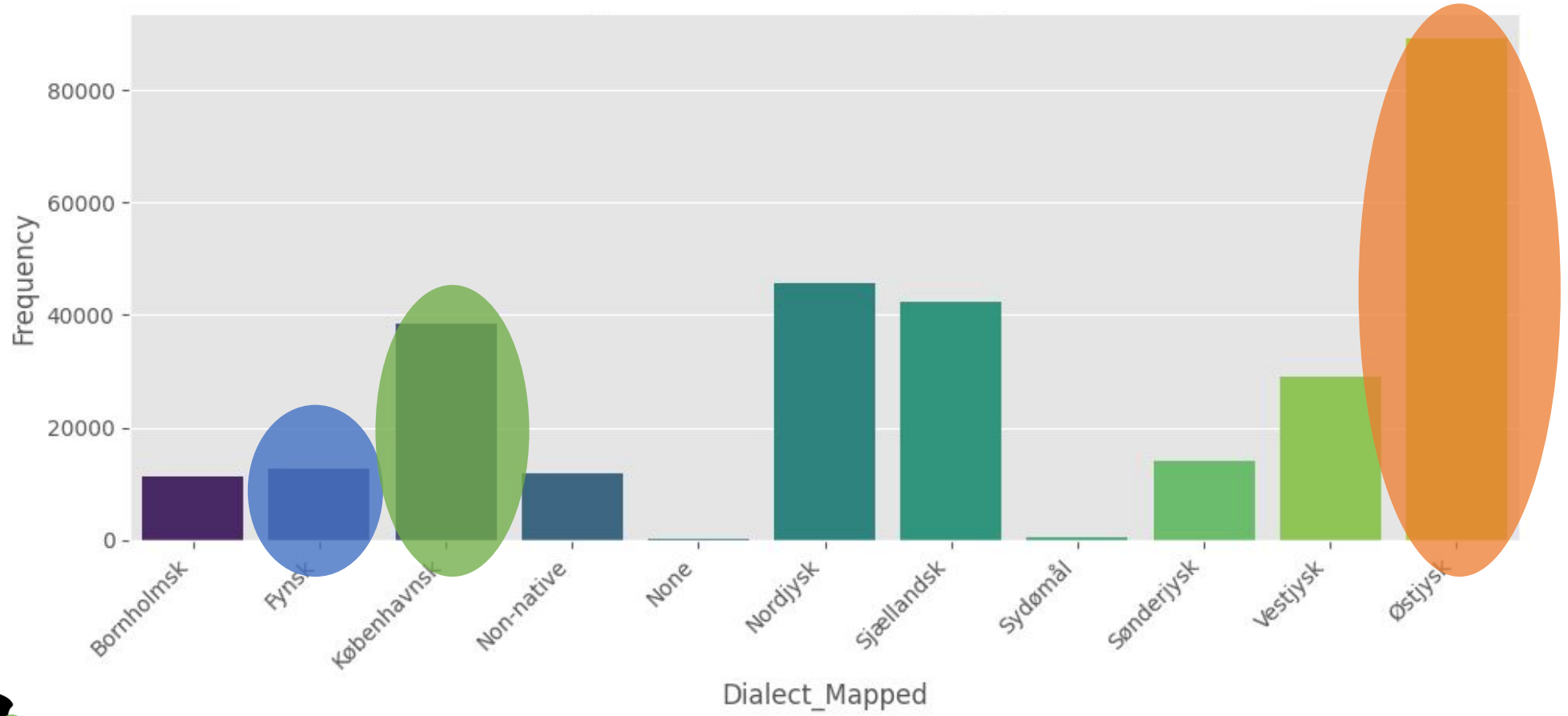
Entire dataset, colors represent dialects

Fynsk datasize 11236

"All" datasize 11236

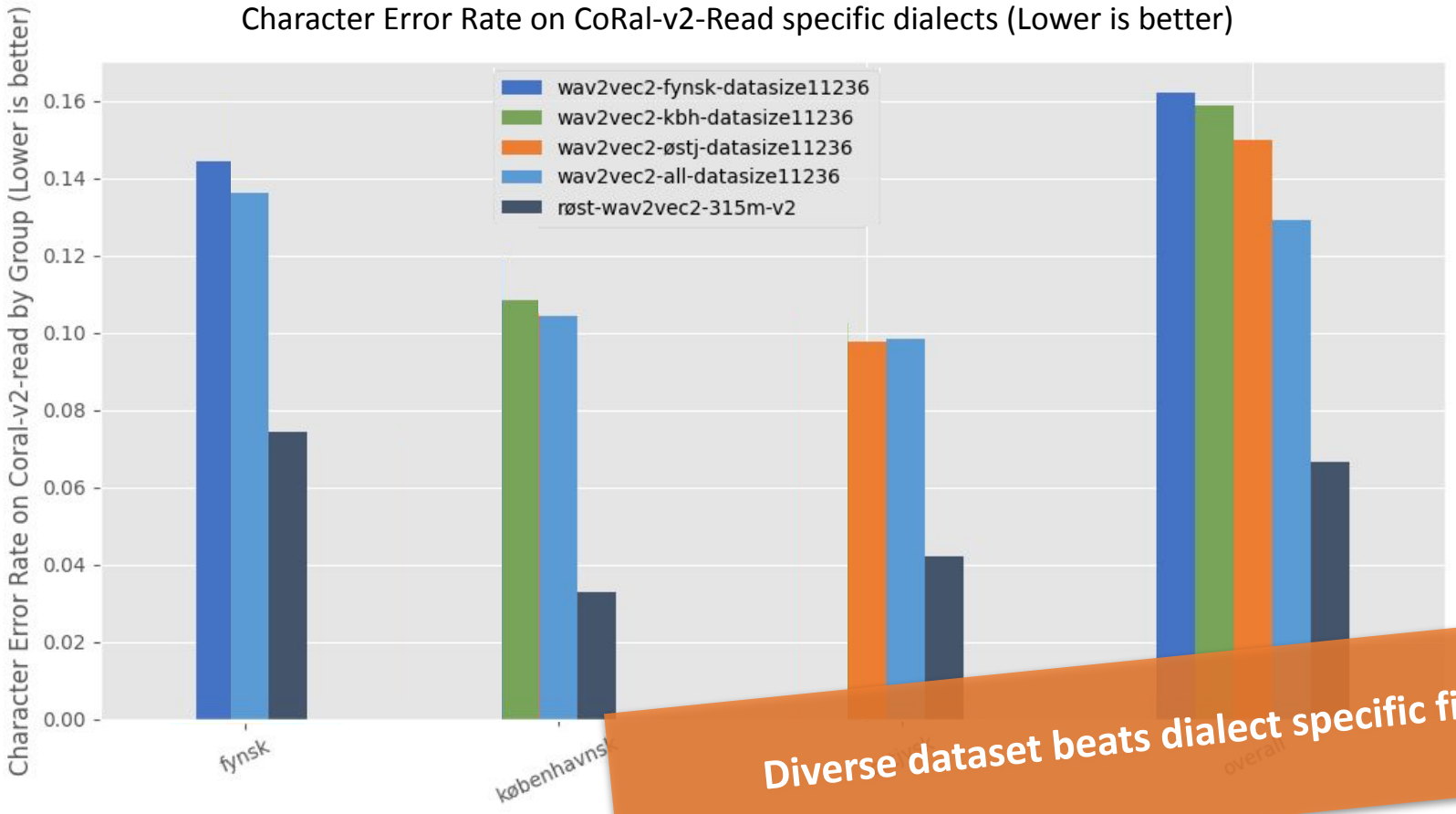
# The data is not evenly distributed across dialects

Coral-v2 Read-aloud and Conversation Data mapped onto dialect categories



# Does training across dialects improve performance?

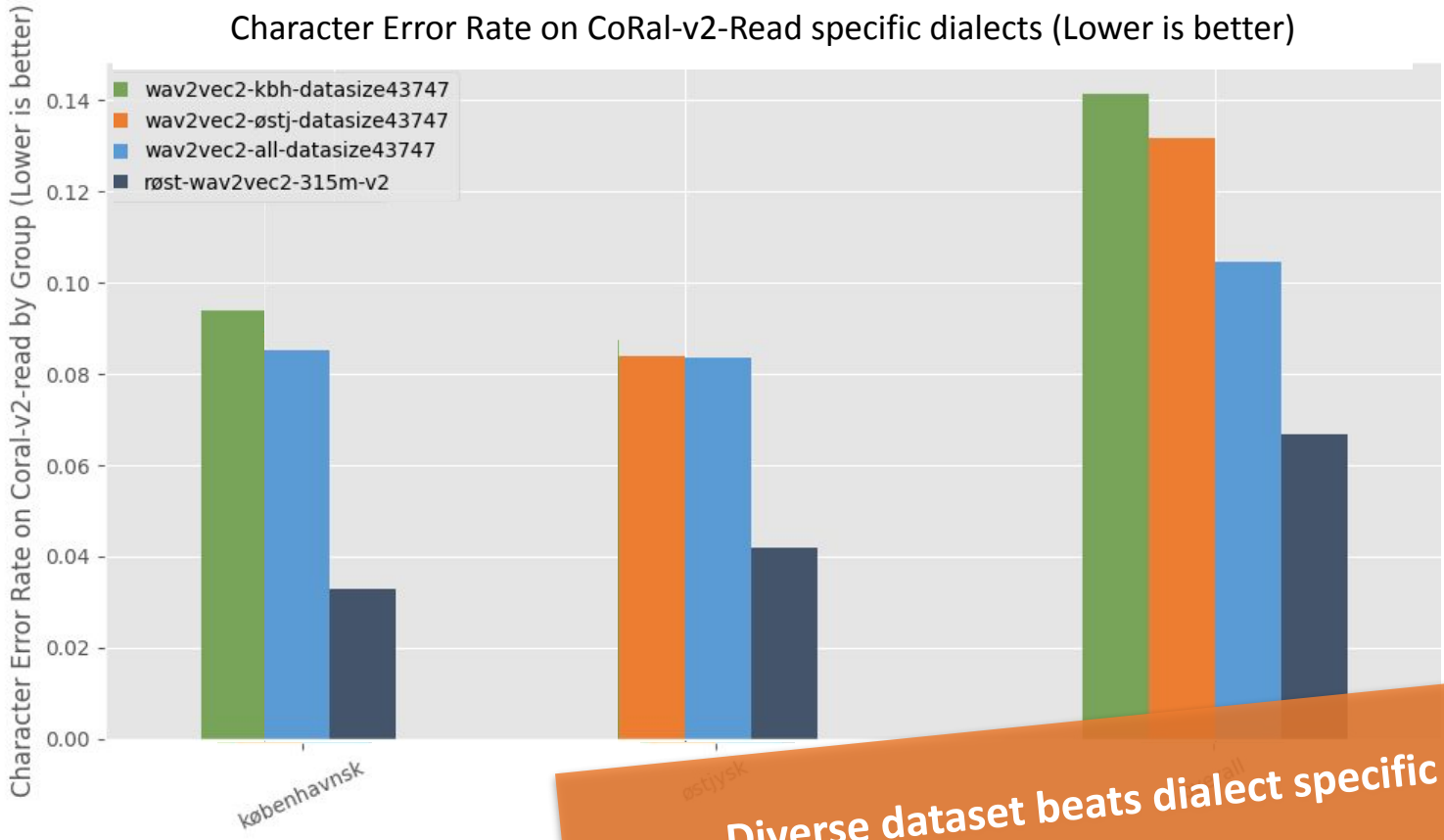
Datasets limited to length of available **“Fynsk”** data = 11 236 samples



**Diverse dataset beats dialect specific finetuning**

# Does training across dialects improve performance?

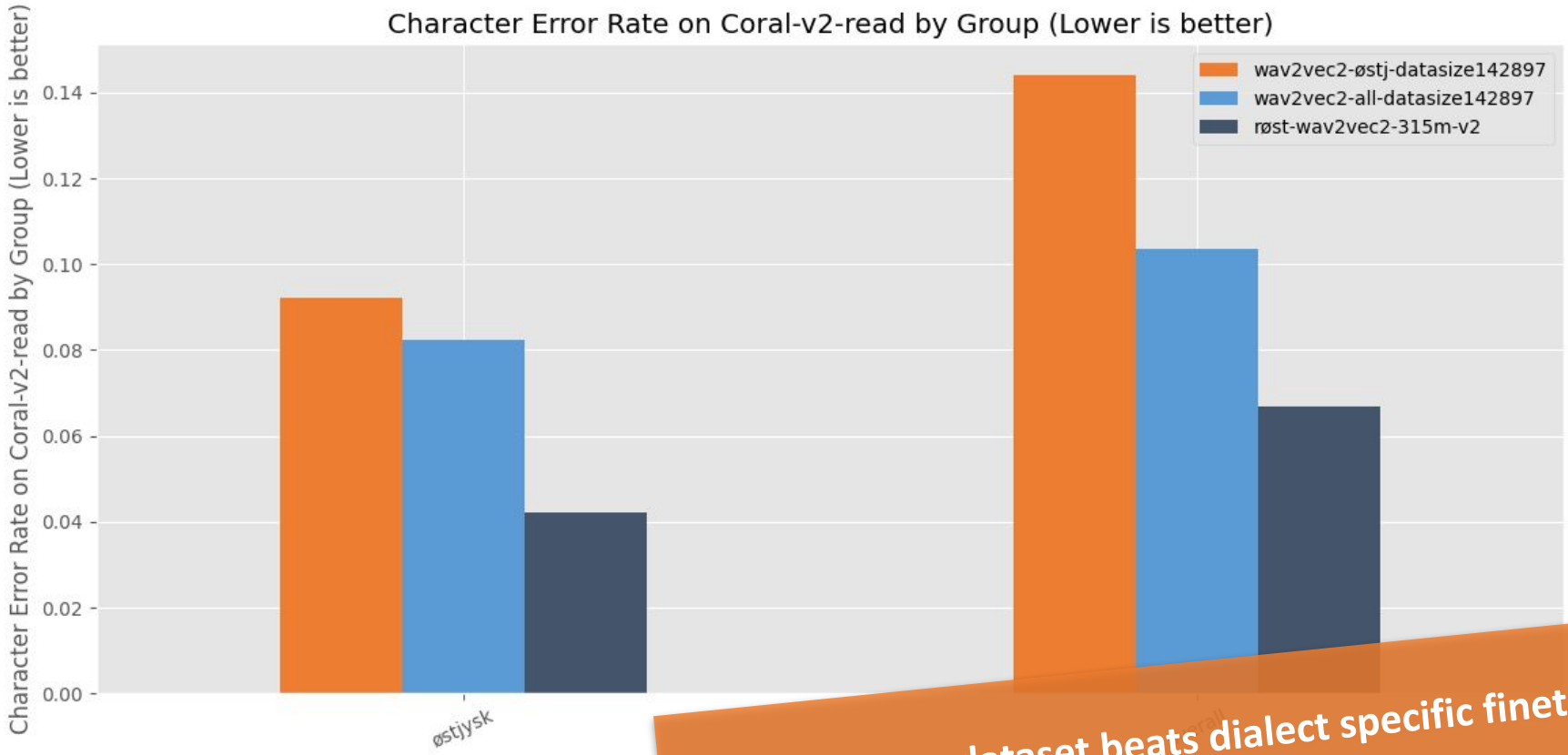
Datasets limited to length of available **“Københavnsk”** data = 43 747 samples



**Diverse dataset beats dialect specific finetuning**

# Does training across dialects improve performance?

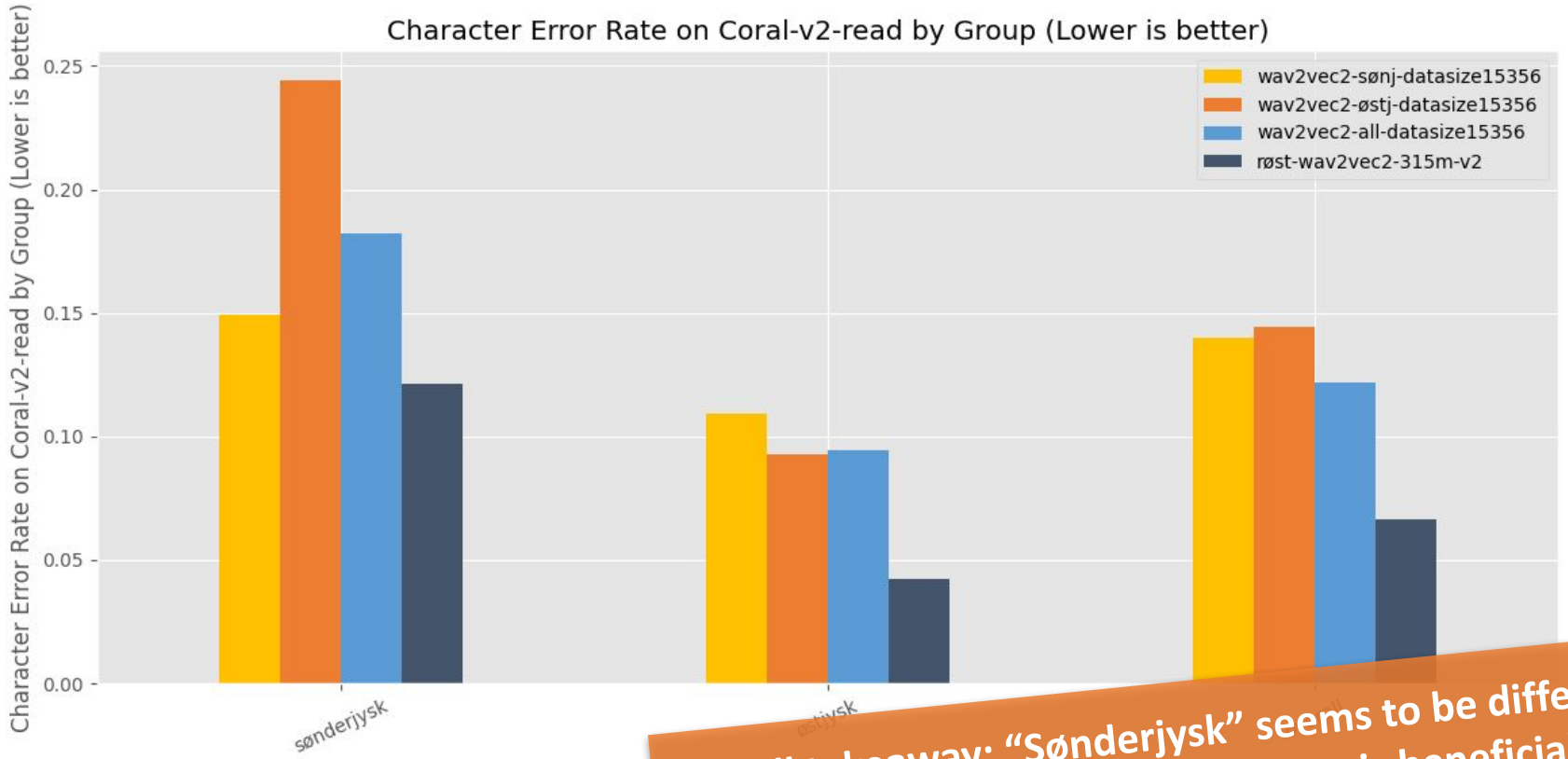
Datasets limited to length of available **“Østjysk”** data = 142 897 samples



**Diverse dataset beats dialect specific finetuning**

# Does training across dialects improve performance?

Datasets limited to length of available **“Sønderjysk”** data = 15 356 samples



**Small takeaway: “Sønderjysk” seems to be different enough that specific finetuning is beneficial**

# To conclude

*Training on all dialects generally performs better than finetuning for specific dialect across data sizes -* ***There is cross-dialect learning***

Coral data is available at: https://huggingface.co/CoRal-project

ALVENIR
SPEECH TECHNOLOGIES

# Thank you to our collaborators

*The CoRal project is funded by the Danish Innovation Fund and consists of the following partners:*

- Alexandra Institute
- University of Copenhagen
- Agency for Digital Government
- Alvenir
- Corti

We would like specifically to thank Dan Saattrup Nielsen, Alexandra Institute for (among other things) the repository work and Simon Leminen Madsen, Alexandra Institute for modelling work.

CoRal

# Contact


**Alvenir**

Applebys Plads 7
1411, København K

**Website**

www.alvenir.ai

**Email**

marie@alvenir.ai

**LinkedIn**

Marie Juhl Jørgensen


**Questions?**

ALVENIR
SPEECH TECHNOLOGIES