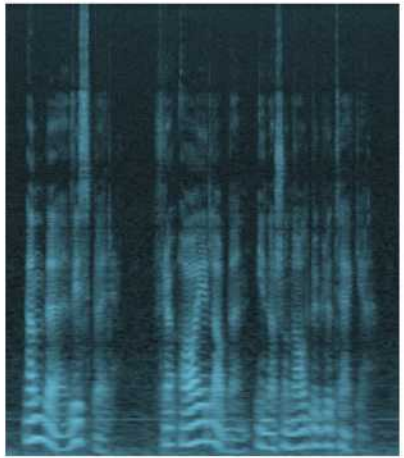
NB-Whisper

Per Egil Kummervold
per.kummervold@nb.no

AI-lab
National Library of Norway

How do you solve speech to text?

- Lets try converting phonems to graphems



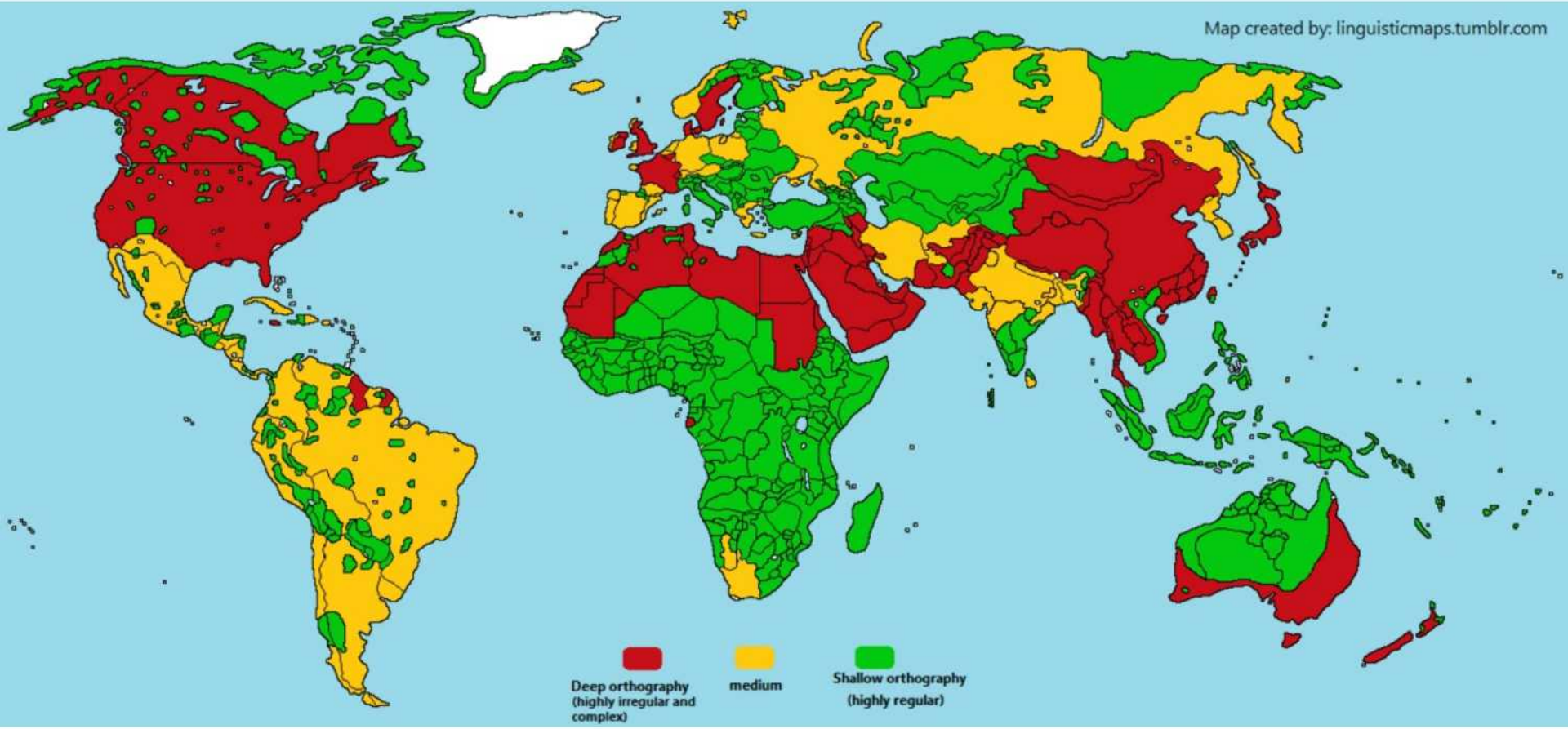
How do you solve speech to text?

- Lets try converting phonems to graphems
- We have been trying this for 70 years

How do you solve speech to text?

- Lets try converting phonems to graphems
- We have been trying this for 70 years
- It does not work

… there are two reasons it does not work



Some Challenges - Nouns, Infinitives and Pronouns

English	Norwegian Bokmål	Norwegian Nynorsk
The church	Kirka / kirken	Kyrkja
The week	Uka / uken	Veka
To do	Å gjøre	Å gjere / Å gjera
To swim	Å svømme	Å symje / Å symja

	English	Norwegian Bokmål	Norwegian Nynorsk
Written	I	Jeg	Eg
Spoken	/aɪ/, /ʌɪ/	/eː/, /eːg/, /e͡ɪg/, /je͡ɪ/, /jeː/, /jæː/, /e͡ɪ/, /æː/, /æːg/, /iː/	

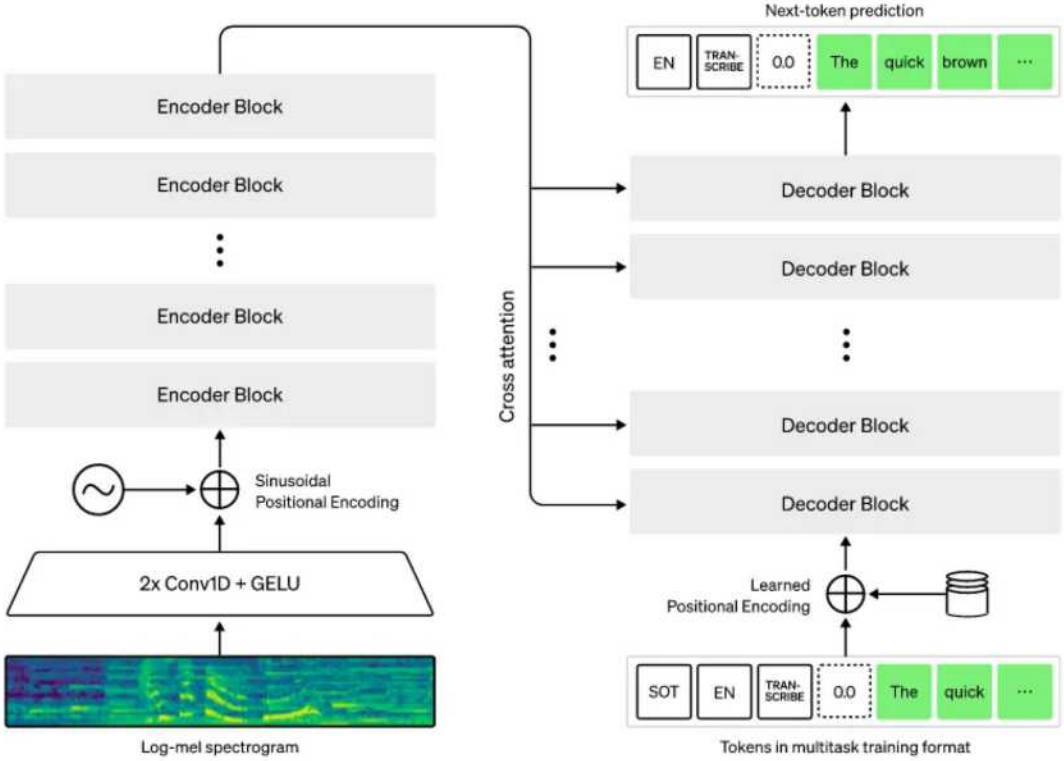
Second Reason - Are you really sure you want this…?

“Men den debatt vi har hatt i samfunnet i de siste fem årene, faktisk, og mer enn det, ja, fem årene, for det var fra niogåtti, fire år, har jo blant annet dreiet seg om at noe av det folk, det har vært mulig å oppvise engasjement på, har vært hva man mener om EF, spesielt hvis man er mot, altså engasjement mot EF har vært noe av det engasjement som har vært mobilisert faktisk, og der har ikke kvinnene vært noe mindre engasjert enn menn i det.”

Gro Harlem Brundtland
Kapital 13/1993 - Niels Christian Geelmuyden

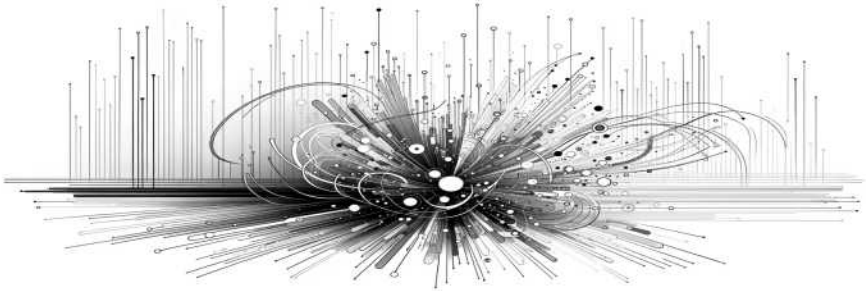
Why does it not work?

- Orthographic distance
  - Varies between languages
  - Varies between dialects
- We do not want it
  - Only in very few settings we want verbatim transcriptions
  - Spontaneous speech and written text is very different
  - In most cases we want semantic transcriptions (to a varying degree)



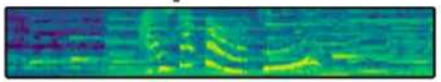
illustrasjon: OpenAI artikkel

Multidimensional embeddings



encoder

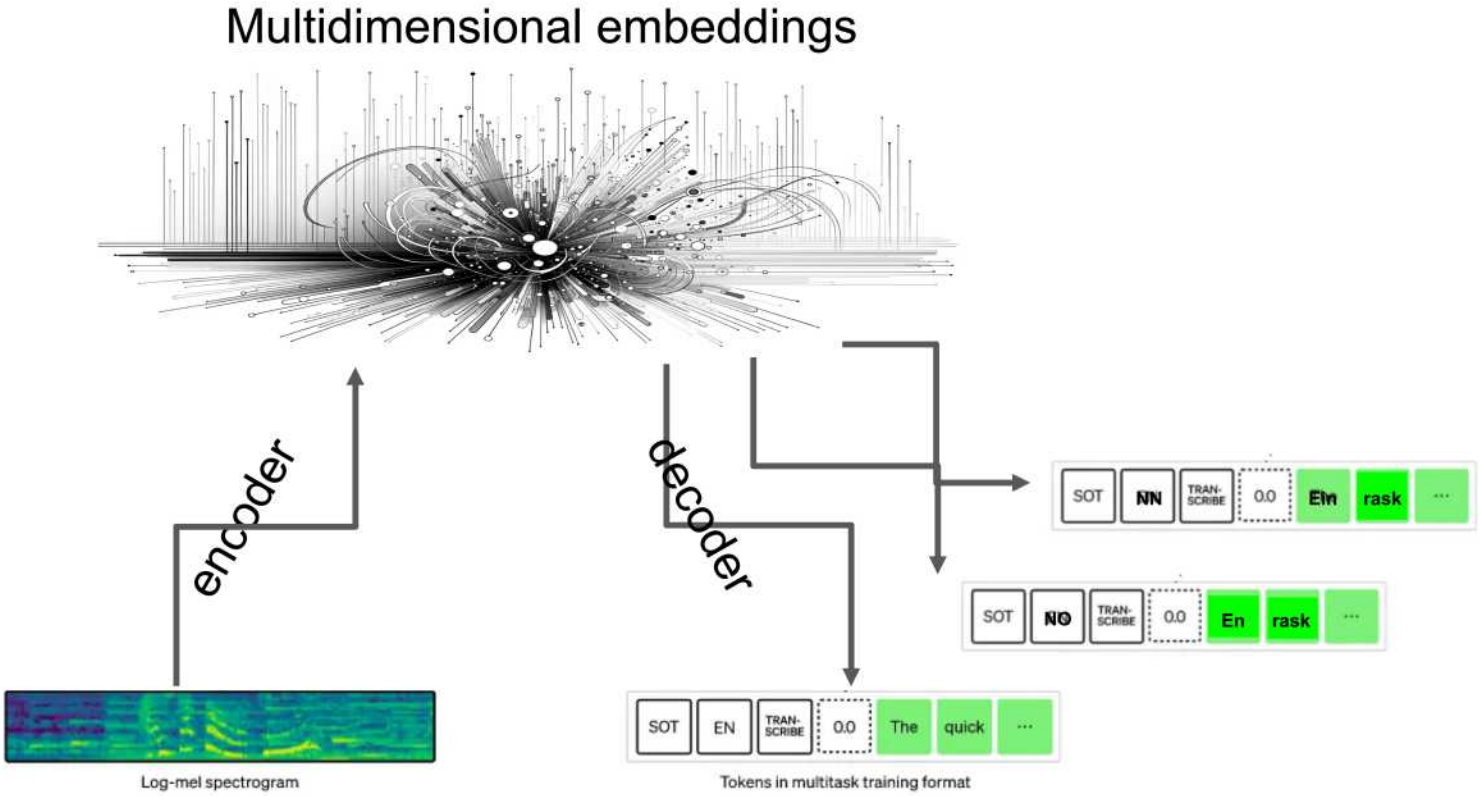
decoder



Log-mel spectrogram

Tokens in multitask training format

illustrasjon: OpenAI artikkel + ChatGPT



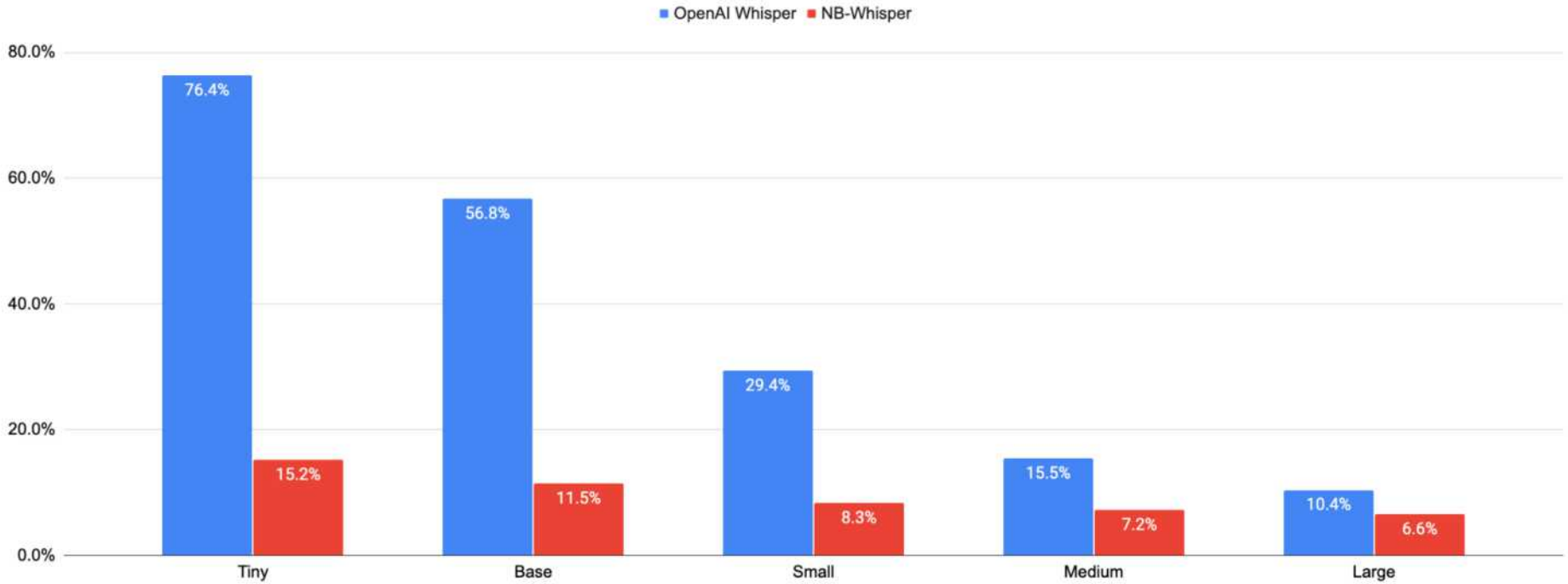
OpenAI Whisper Architecture + ChatGPT

Training corpus

Dataset	Stage 1 (hours)	Stage 2 (hours)
NRK - Subtitles	16,518	2,478
NRK - No caption	715	312
Audio Books	2,461	2,275
The NST Dataset	260	490
The Stortinget Speech Corpus	2,230	523
Total	**22,184**	**6,078**

- 30 second clips
- More than 8 million clips
- Quality of the model == How clean your data is

Fleurs - Bokmål



Whispering in Norwegian: Navigating Orthographic and Dialectic Challenges - Preview - https://arxiv.org/abs/2402.01917

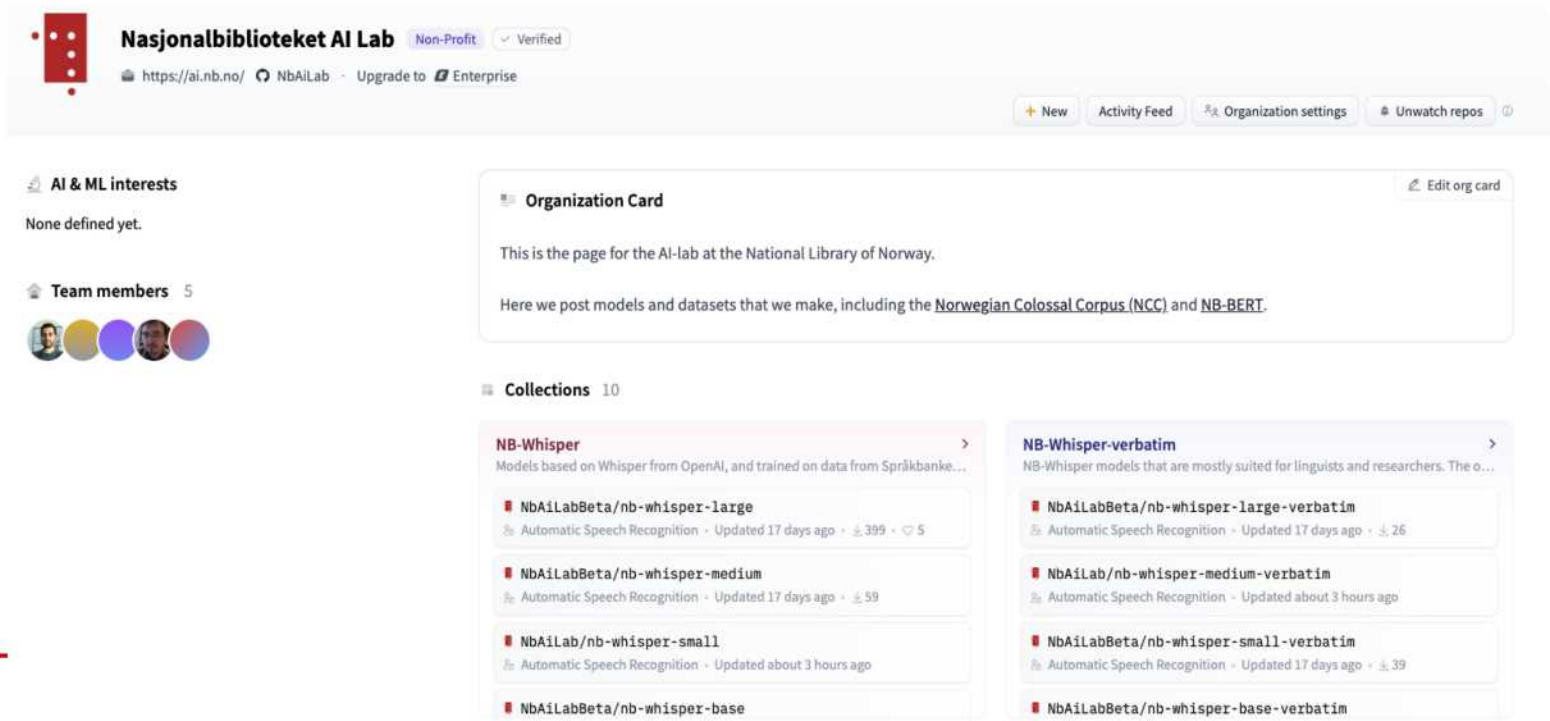
External evaluation - WER

Bokmål	WER %
NB-Whisper Verbatim	11,89
NB-Whisper	18,71
Google USM	20,84
NB-Wav2Vec2	22,32
Microsoft Azure	24,63
Open AI Whisper	30,17
Google Cloud long	30,40
NB-Whisper Semantic	31,53

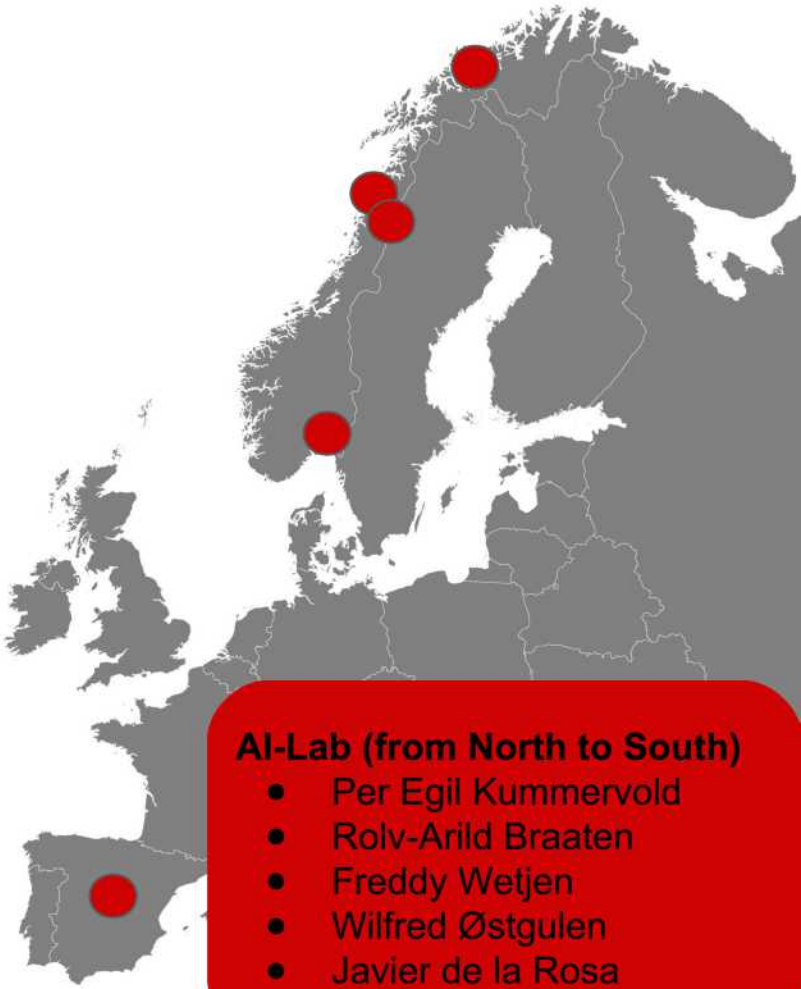
External evaluation - SemDist

Bokmål	SemDist
NB-Whisper	0,09
NB-Whisper Semantic	0,11
NB-Whisper Verbatim	0,12
Microsoft Azure	0,17
OpenAI Whisper	0,20
NB-Wav2Vec2	0,21
Google USM	0,26
Google Cloud long	0,38

https://ai.nb.no
https://huggingface.co/NbAiLab



https://ai.nb.no



AI-Lab (from North to South)

- Per Egil Kummervold
- Rolv-Arild Braaten
- Freddy Wetjen
- Wilfred Østgulen
- Javier de la Rosa

National Library of Norway

Google Research

Thanks!

Questions?

Per Egil Kummervold
per.kummervold@nb.no

AI-lab
National Library of Norway