

# NB-Whisper

Per Egil Kummervold  
[per.kummervold@nb.no](mailto:per.kummervold@nb.no)



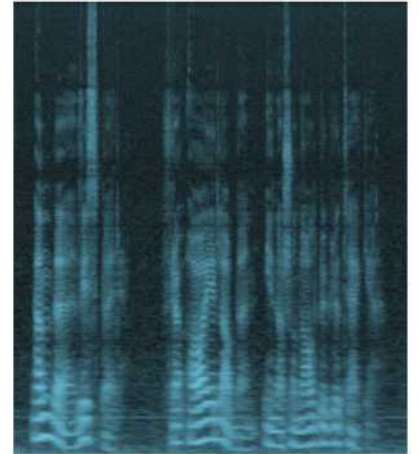
**AI-lab**

National Library of Norway



How do you solve speech to text?

- Lets try converting phonemes to graphemes



How do you solve speech to text?

- Lets try converting phonemes to graphemes
- We have been trying this for 70 years

How do you solve speech to text?

- Lets try converting phonemes to graphemes
- 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
<b>Written</b>	I	Jeg	Eg
<b>Spoken</b>	/aɪ/, /ɪ/	/e:/, /e:g/, /eɪ̯g/, /jɛi̯/, /je:/, /jæ:/, /eɪ̯/, /æ:/, /æ:g/, /i:/	

## Second Reason - Are you really sure you want this...?

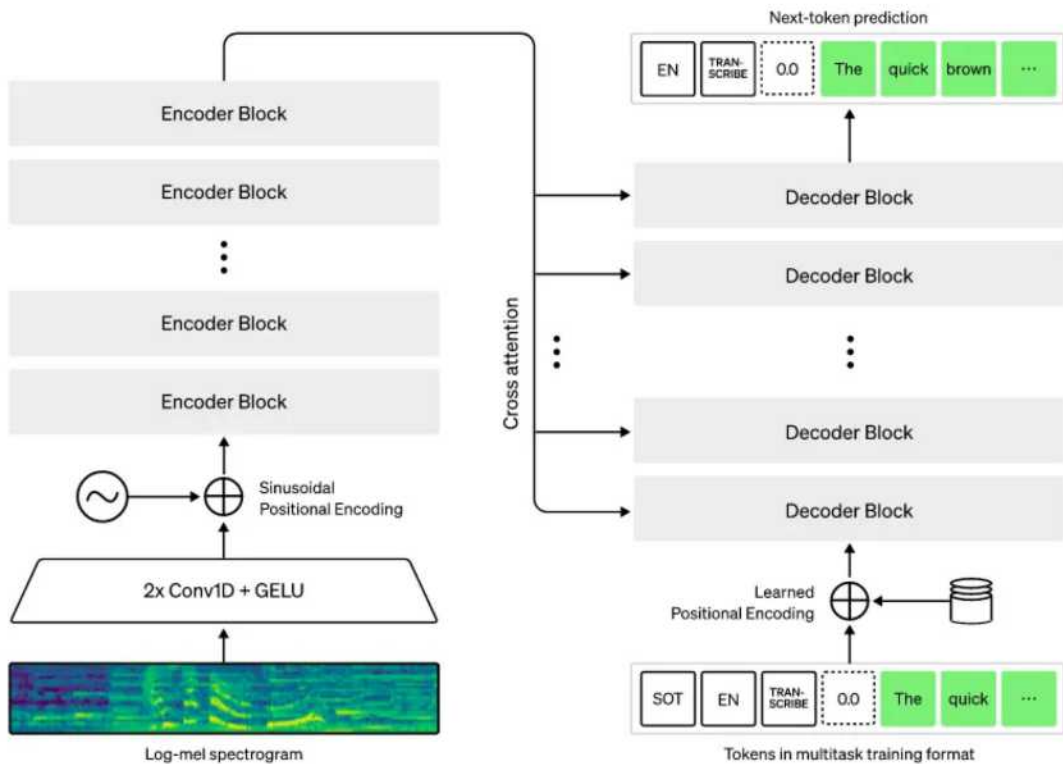
*“Men den debatt vi har hatt i samfunnet i de siste fem årene, faktisk, og mer enn det, ja, fem årene, for det var fra niogåtti, fire år, har jo blant annet dreiet seg om at noe av det folk, det har vært mulig å oppvise engasjement på, har vært hva man mener om EF, spesielt hvis man er mot, altså engasjement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

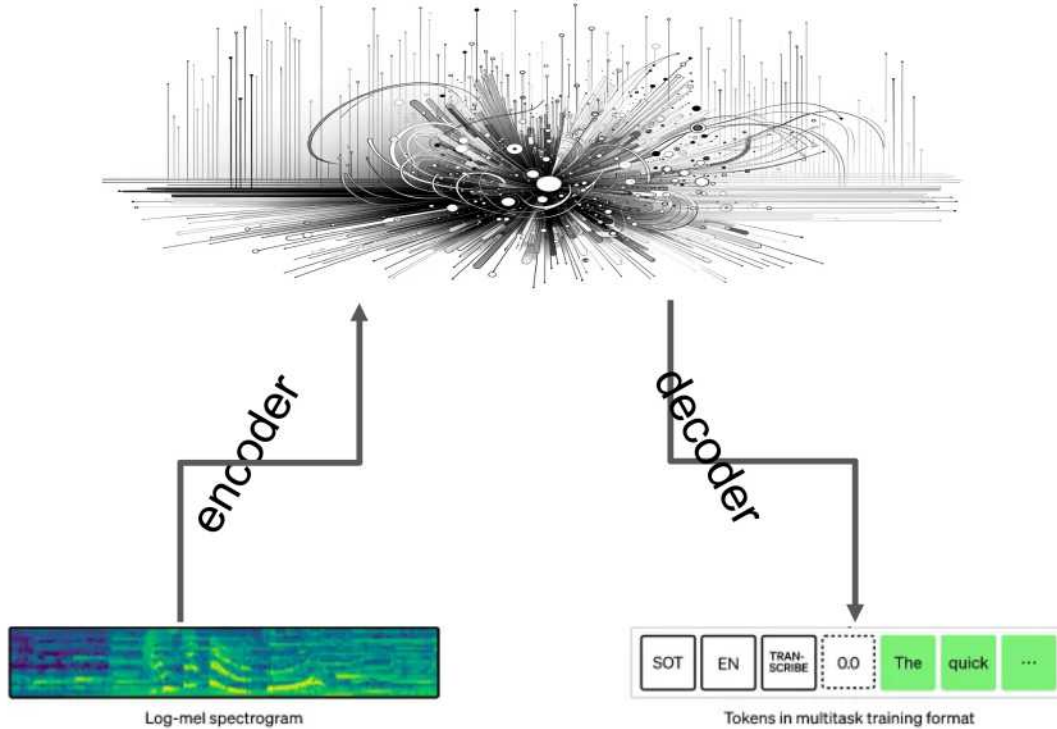
Why does it not work?

- 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)

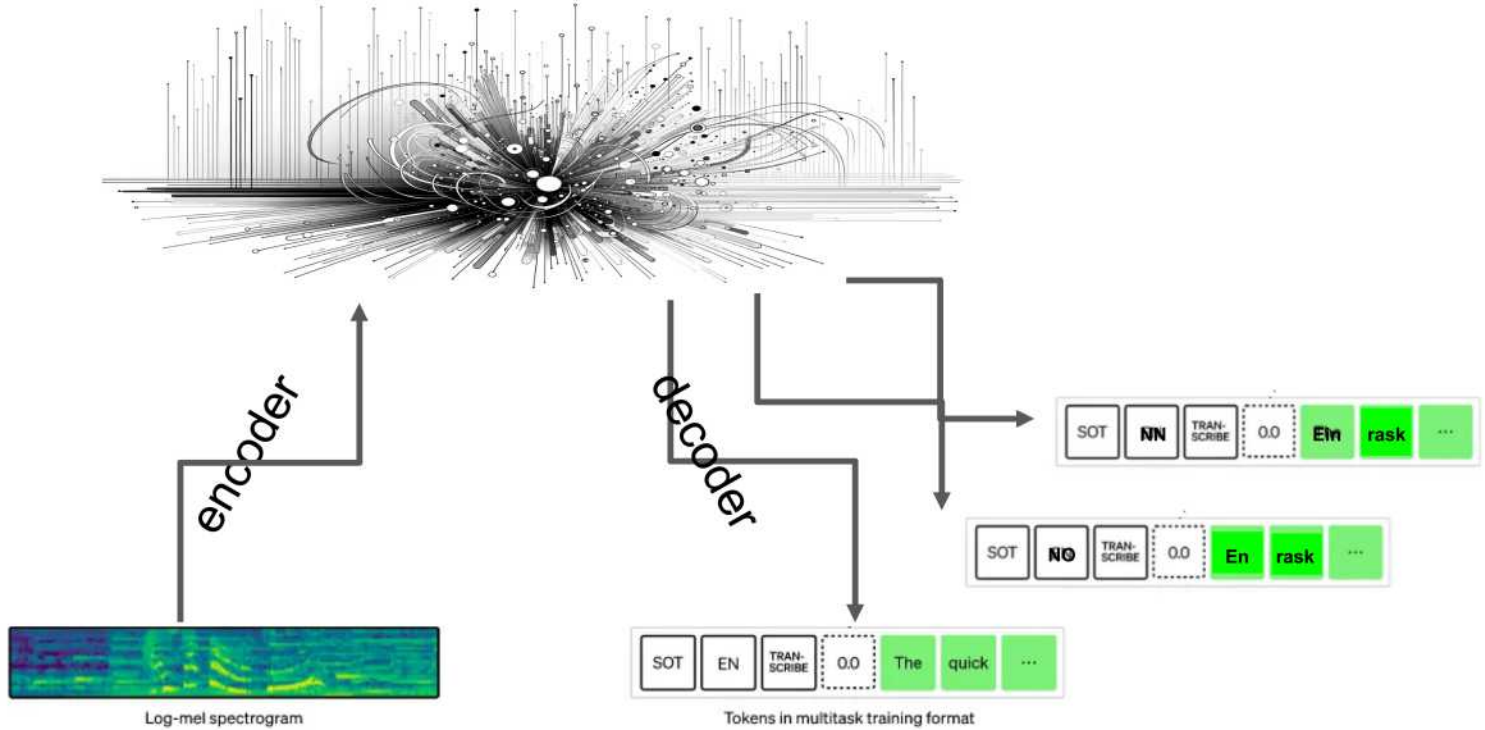




# Multidimensional embeddings



# Multidimensional embeddings

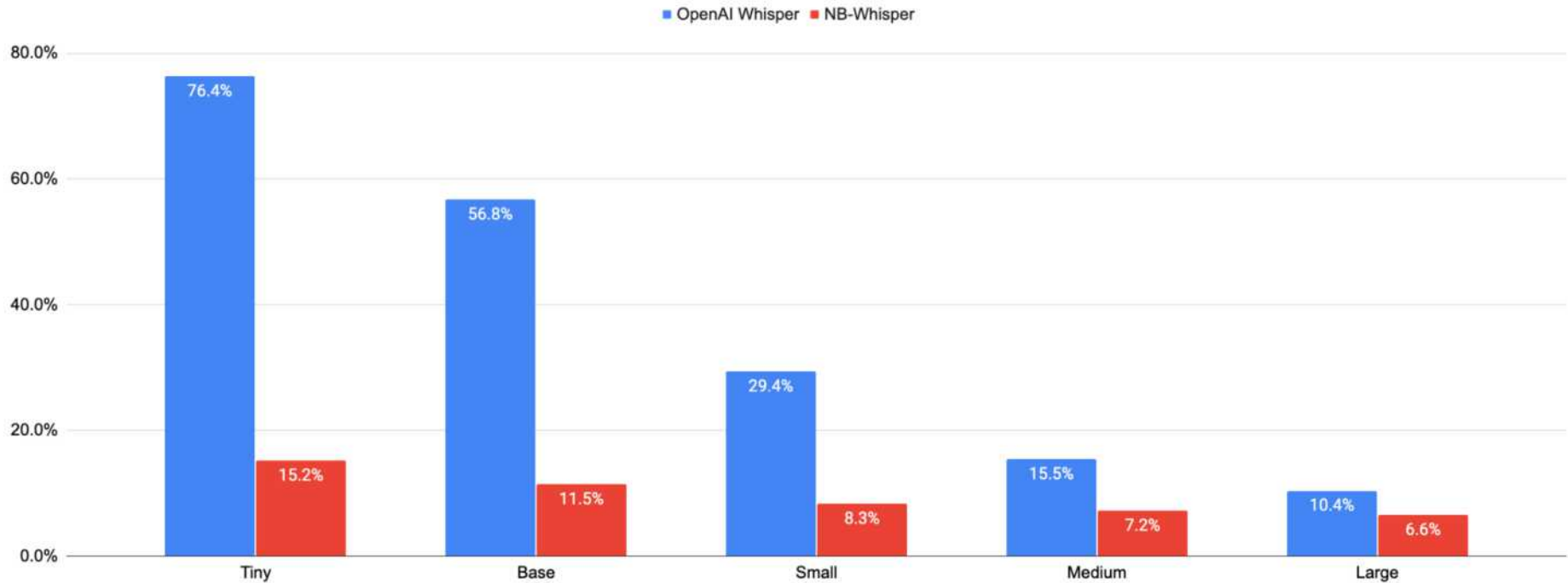


# Training corpus

<b>Dataset</b>	<b>Stage 1 (hours)</b>	<b>Stage 2 (hours)</b>
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
<b>Total</b>	<b>22,184</b>	<b>6,078</b>

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

## Fleurs - Bokmål



## External evaluation - WER

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

## External evaluation - SemDist

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

<https://ai.nb.no>

<https://huggingface.co/NbAiLab>

The screenshot shows the Hugging Face profile page for 'Nasjonalbiblioteket AI Lab'. The profile is a 'Non-Profit' organization, verified, with a website link to 'https://ai.nb.no/'. It has 5 team members and 10 collections. The 'AI & ML interests' section is currently empty. The 'Organization Card' contains a description: 'This is the page for the AI-lab at the National Library of Norway. Here we post models and datasets that we make, including the Norwegian Colossal Corpus (NCC) and NB-BERT.' The 'Collections' section is divided into two columns. The left column lists 'NB-Whisper' models: 'NbAiLabBeta/nb-whisper-large', 'NbAiLabBeta/nb-whisper-medium', 'NbAiLab/nb-whisper-small', and 'NbAiLabBeta/nb-whisper-base'. The right column lists 'NB-Whisper-verbatim' models: 'NbAiLabBeta/nb-whisper-large-verbatimim', 'NbAiLab/nb-whisper-medium-verbatimim', 'NbAiLabBeta/nb-whisper-small-verbatimim', and 'NbAiLabBeta/nb-whisper-base-verbatimim'. Each model entry includes a brief description, the number of likes, and the time since it was updated.

**Nasjonalbiblioteket AI Lab** Non-Profit Verified

<https://ai.nb.no/> NbAiLab Upgrade to Enterprise

+ New Activity Feed Organization settings Unwatch repos

AI & ML interests  
None defined yet.

Team members 5

Organization Card Edit org card

This is the page for the AI-lab at the National Library of Norway.

Here we post models and datasets that we make, including the Norwegian Colossal Corpus (NCC) and NB-BERT.

Collections 10

**NB-Whisper**

Models based on Whisper from OpenAI, and trained on data from Språkbanke...

- NbAiLabBeta/nb-whisper-large Automatic Speech Recognition · Updated 17 days ago · 399 · 5
- NbAiLabBeta/nb-whisper-medium Automatic Speech Recognition · Updated 17 days ago · 59
- NbAiLab/nb-whisper-small Automatic Speech Recognition · Updated about 3 hours ago
- NbAiLabBeta/nb-whisper-base

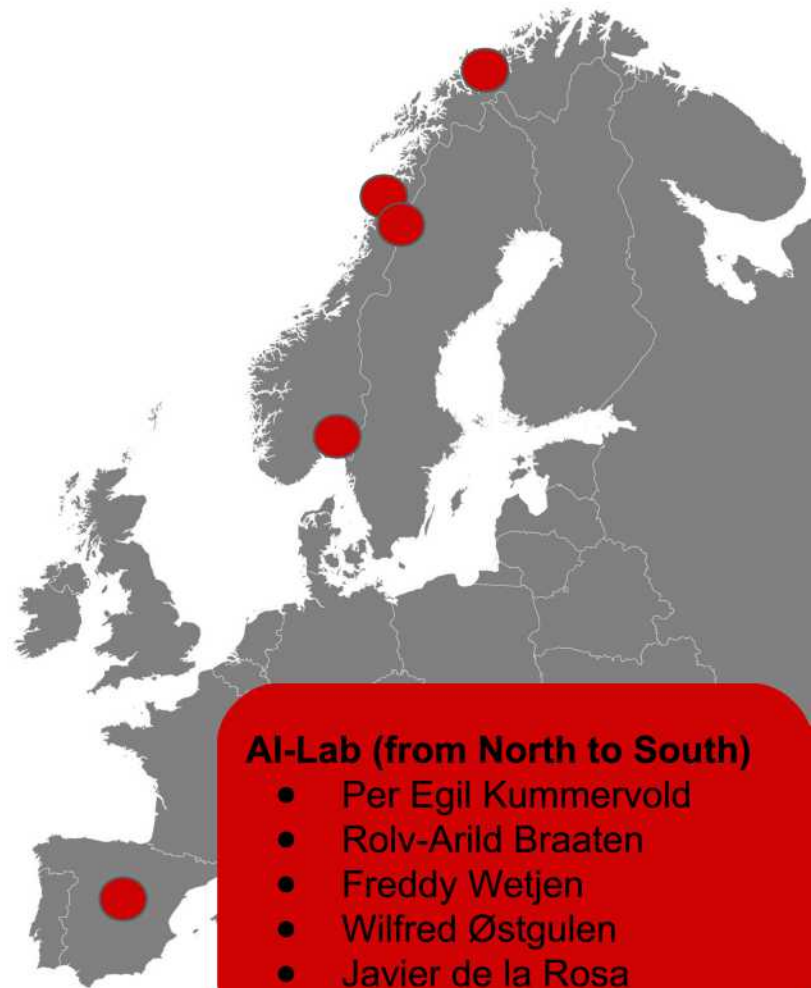
**NB-Whisper-verbatimim**

NB-Whisper models that are mostly suited for linguists and researchers. The o...

- NbAiLabBeta/nb-whisper-large-verbatimim Automatic Speech Recognition · Updated 17 days ago · 26
- NbAiLab/nb-whisper-medium-verbatimim Automatic Speech Recognition · Updated about 3 hours ago
- NbAiLabBeta/nb-whisper-small-verbatimim Automatic Speech Recognition · Updated 17 days ago · 39
- NbAiLabBeta/nb-whisper-base-verbatimim



<https://ai.nb.no>



**AI-Lab (from North to South)**

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

Thanks!

Questions?

Per Egil Kummervold ·  AI-lab  
[per.kummervold@nb.no](mailto:per.kummervold@nb.no) · National Library of Norway