Kunstig Intelligens på dansk



DTHS – Teknologiseminar 2024

Torben Blach – Alexandra Instituttet

ALEXANDRA INSTITUTTET

ABOUT

*The Alexandra Institute is a non-profit commercial **consultant** company that works with **applied IT research and development**. Our mission is to create value, growth and welfare in our society with innovative, IT-based products and services.*

CLIENT EXAMPLES

TORBEN BLACH
Senior Projekt- og Programleder
AI Lab, Århus

Torben.blach@alexandra.dk
+45 9155 8799

Sammen kommer vi #forandigitalt

ALEXANDRA INSTITUTTET

GRAND SOLUTION PROGRAMME

Digitalization, technology, and innovation

BUDGET

14.217.380 DKK / 22.172.400 DKK

CoRal

Danish **Co**nversational and **R**ead-**al**oud Speech Dataset

ALVENIR

Corti

AGENCY FOR DIGITAL GOVERNMENT
MINISTRY OF FINANCE

ALEXANDRA INSTITUTTET

UNIVERSITY OF COPENHAGEN

Innovationsfonden

Så forstå det dog.....



Scottish Elevator - Voice Recognition - ELEVEN !

PROJECT VISION

To build a Danish speech dataset that elevates Danish speech technology and help creation of new companies, jobs and growth in the speech tech industry and in society in general

WHY A DANISH DATASET?

Speech tech in general is a growing industry, with an estimated annual compound growth of 16-19% in the next five years.

But the majority of speech tech development is targeted towards high resource language such as English.

Danish, a small and low-resource language, risks falling behind and missing revenue from this technology.

Modern speech tech is based on ML algorithms, which require large amounts of data. **Lack of Danish data** is the main problem for moving the industry forward.



INHIBITORS

Speech data are expensive to produce

Danish is a low-resource language

CONSEQUENSES

The entry barrier is too high for most companies to cross alone

Only few demographics are supported

WE NEED TO INVEST IN DANISH SPEECH TECH OTHERWISE WE RISK THAT COMPANIES WILL CHOOSE NOT TO DEVELOP PRODUCTS AND SERVICES FOR THE DANISH MARKET

EFFECTS



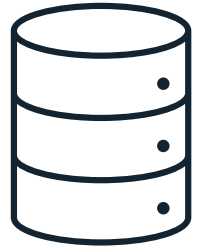
Danish companies ability to compete will be lowered as speech tech takes off in other languages



DK citizens/consumers will not be able to use speech tech products and services despite Denmark being one of the most digitalized countries

AIM

BUILD A HIGH QUALITY DANISH DATASET THAT COMPANIES CAN USE TO BUILD SPEECH TECH PRODUCTS AND SERVICES



1000-1500 h.

- **Conversational speech (<500 h)**
- **Read aloud (<1000 h)**
- **Annotated**

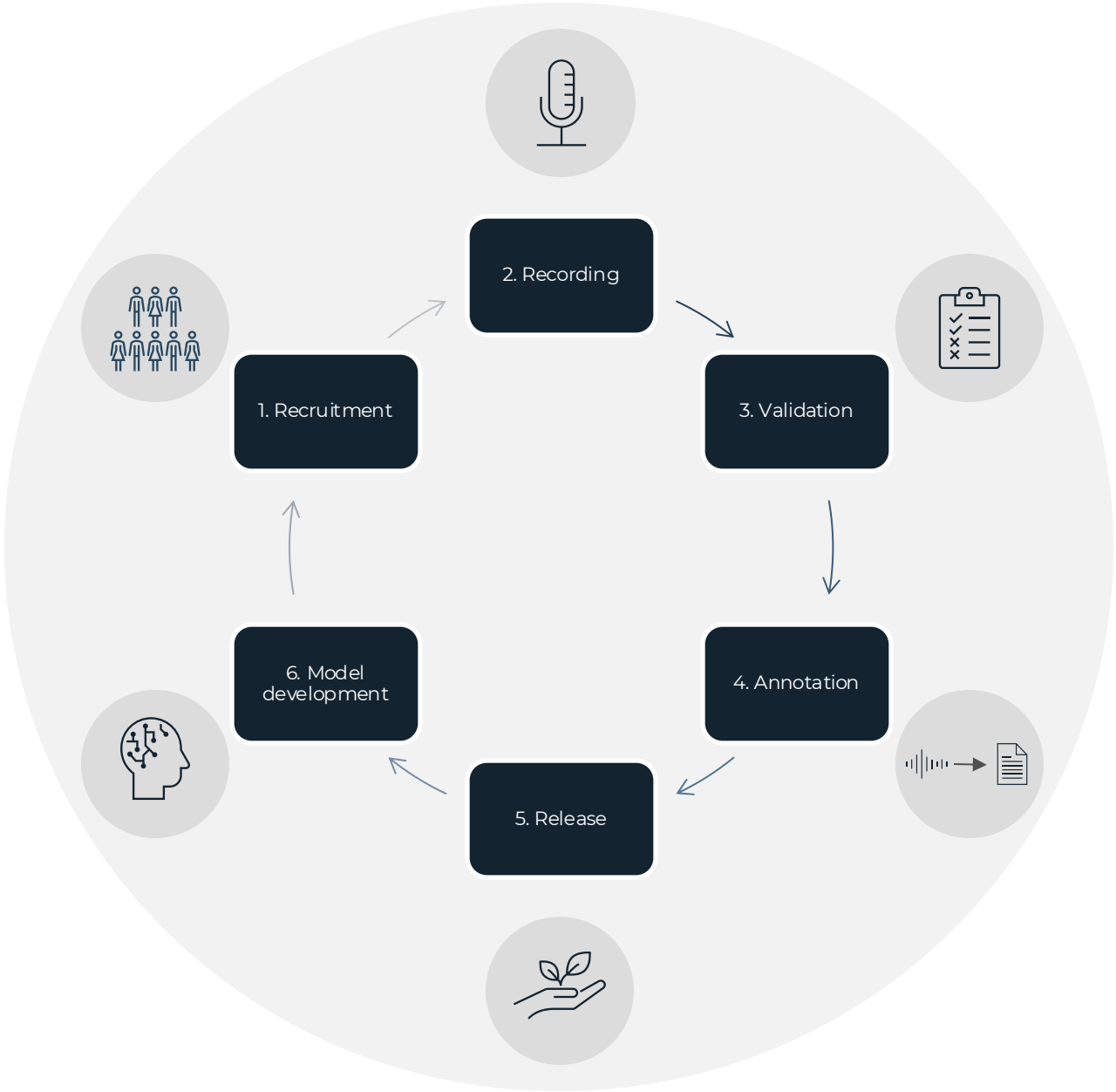
OPEN SOURCE



THE DATASET WILL BE PRODUCED IN FOUR ITERATIONS

PRELIMNARY PLAN

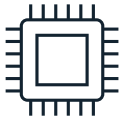




FAIR SPEECH DATA



SPEECH MODELS



ONLINE RESOURCES



MAIN TEAM



Jens Kaas Benner
HEAD OF AI AND DATA ANALYTICS
THE ALEXANDRA INSTITUTE



Anders Jess Pedersen
SENIOR AI SPECIALIST
THE ALEXANDRA INSTITUTE



Dan Sattrup Nielsen
SENIOR AI SPECIALIST
THE ALEXANDRA INSTITUTE

Torben Blach
SENIOR PROJECT MANAGER
THE ALEXANDRA INSTITUTE

Sif Bernstorff Lehmann
ANTHROPOLOGIST
THE ALEXANDRA INSTITUTE



Martin Carsten Nielsen
CO-FOUNDER
ALVENIR



Jane Gyberg Rasmussen
HEAD OF DIVISION, TECHNOLOGY & DATA
AGENCY FOR DIGITAL GOVERNMENT



Josephine Worm Andersson
SPECIAL ADVISOR
AGENCY FOR DIGITAL GOVERNMENT



Anders Søgaard
PROFESSOR
UNIVERSITY OF COPENHAGEN



Anna Katrine Van Zee
SENIOR CONSULTANT
UNIVERSITY OF COPENHAGEN



Lars Maaløe
CTO/ASSOCIATE PROFESSOR
CORTI/DTU

SOUNDING BOARD

PARTNERS

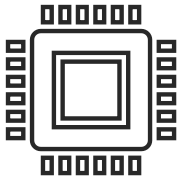
CoRal

(Danish **Co**nversational and **R**ead-**al**oud Speech Dataset)



1000-1500 h. speech data

- Conv. speech (<500 h)
- Read aloud (<1000 h)
- Annotated



ML speech models

- Speech recognition
- Text to speech
- Acoustic models
- Domain specific models



Open source

- DaNLP
- www.sprogteknologi.dk
Open source hubs



+500 companies

- Industry engagement
- Conferences/meetups
- Education
- Online resources



Duration: 3 years - primo 2023 - ultimo 2025

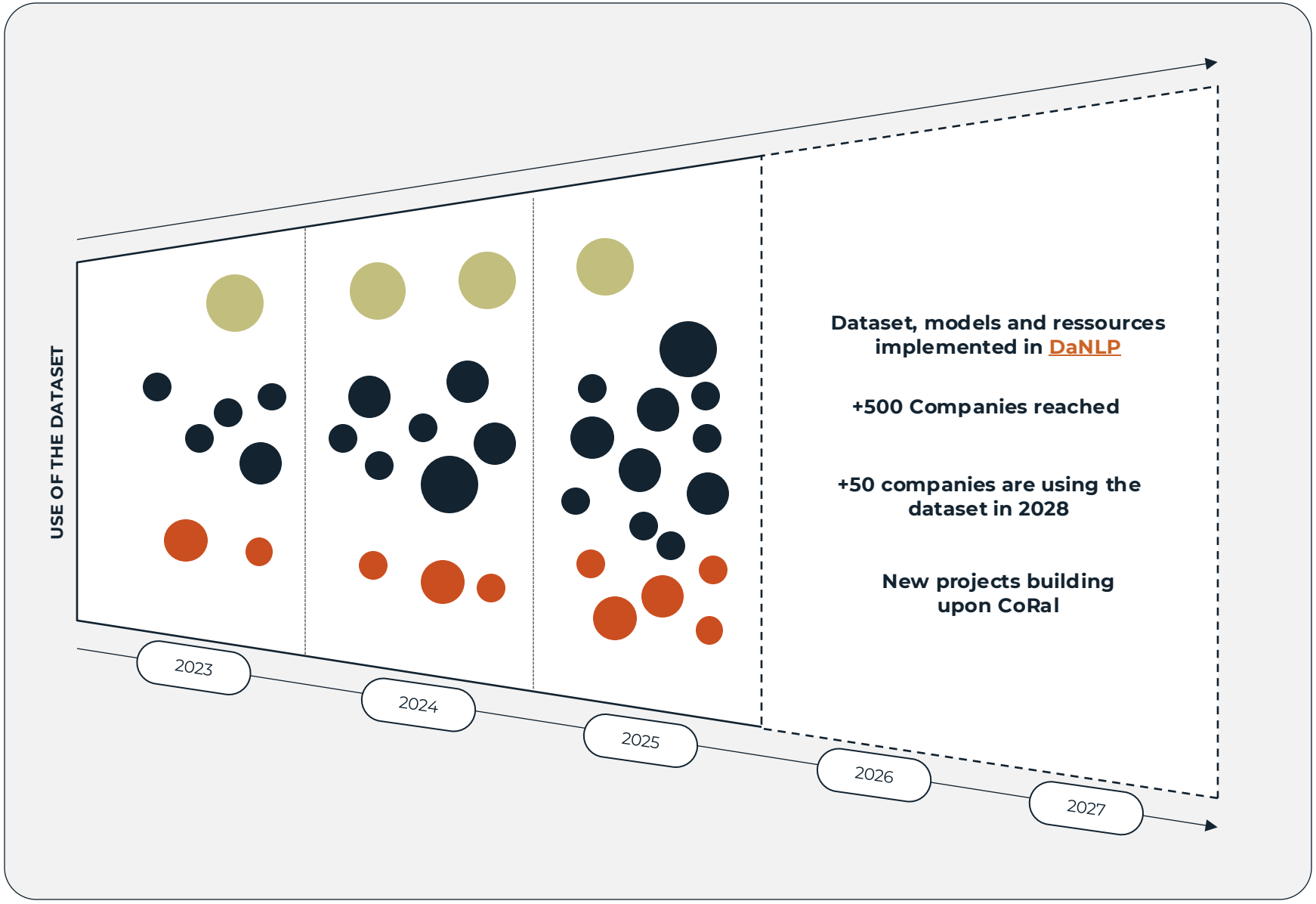


Budget: DKK 14.219.811 (IFD) / DKK 22.963.455 (Full budget)



Economic potential: DKK 150.000.000

IMPLEMENTATION AND PROJECT AFTERLIFE



INDUSTRY COLLABORATION
- Industry board meetings (Four total)

OUTREACH AND COMMUNICATION
- Host two conferences
- Open seminars on data production annually (Five total)
- Participation in two scientific conferences annually (Five total)
- Participation in two industry conferences annually (Five total)
- 1-3 media articles annually (Six total)

COMMMUNITY BUILDING
- +10 ressources in DaNLP (tutorials, notebooks, etc.)

Scope of activity
small → large

Recordings

Read Aloud

- Content based on random sentences from Wikipedia
- 20 min. wo/ noise
- 20+20 min. w/ noise (traffic and people)

Conversation

- In groups of 2
- 20 min free conversation
- 20 min Ideo cards w/ noise





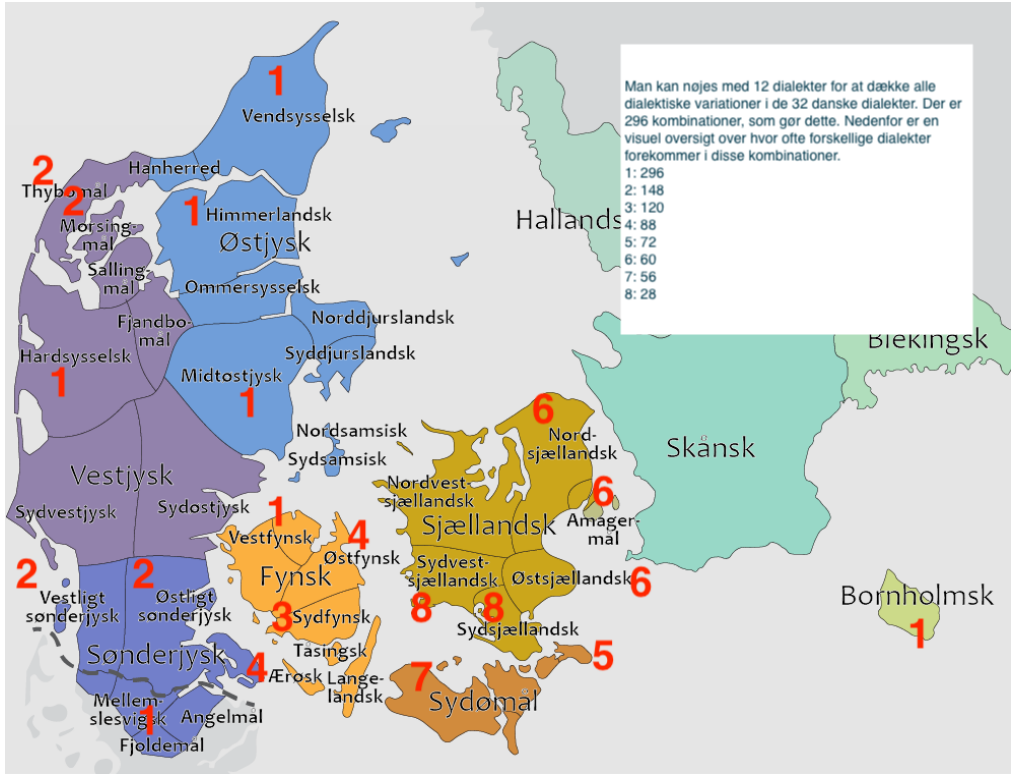
ANALYSE AF DIALEKTER SOM BAGGRUND FOR REKRUTTERINGS-GRID

Fokus for første iteration af dataproduktion er *dialekter*, fordi:

- Dialekter er den sværeste opgave at løse, grundet mange dialekter (32, med hver forskellige variationer).
- Det forventes at det relativt store antal timer, der skal indsamles, vil sikre dækning af krav for køn og alder.
- Et stort antal timer optaget I hovedstadsområdet vil sikre dækning af krav for accent.

Analyse af dialekterne viser at man kan dække samtlige dialekttræk med:

- 7 nødvendige dialekter (eksempelvis Bornholmsk)
- 5 dialekter valg (eksempelvis, vælg Thybomål eller Morsingmål)





Godaw' do: Nu skal danske dialekter samles i en digital bank

Projektet skal ... at stemmestyret teknologi kan forstå de ...kter.

...for tredje gang på tre dage:
27-årig tilstår stribevis af bilindbrud

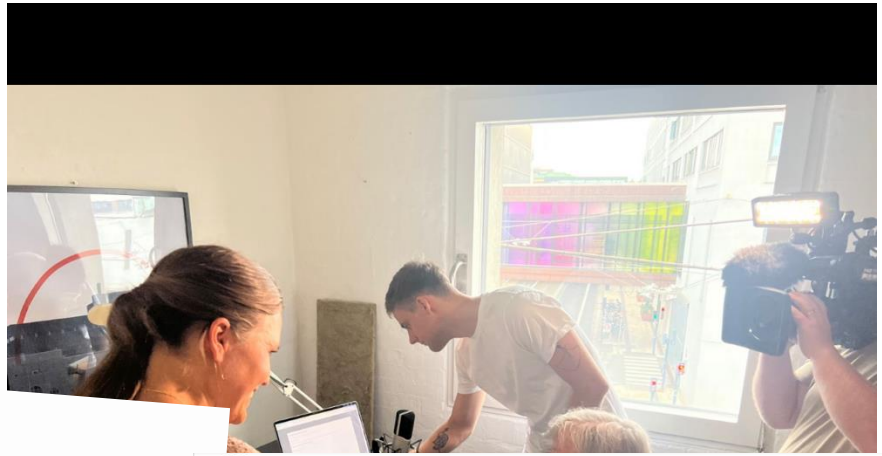
I løbet af få dage har der været indbrud i omkring tyve biler

Biblotek søger Silkebår-stemmer

Lokale stemmer ska gør mobiltelfoner og computere bedre te at forstå os

LÆS I MORGEN

Ramt af PTSD: Jagt på pirater



Aalborg Bibliotekerne
22 t. ·

Hvassåå Aaller'n! Prøliåhørheriggå...

Hvis du, eller nogen du kender, taler med dialekt eller accent, så vil Alexandra Instituttet gerne bede om din hjælp.

De skal nemlig bruge optagelser af forskellige dialekter og accenter, som skal hjælpe computere med at blive bedre til at forstå menneskelig tale på dansk.

I sidste ende skal arbejdet hjælpe med at gøre livet lettere for medborgere med særlige behov.

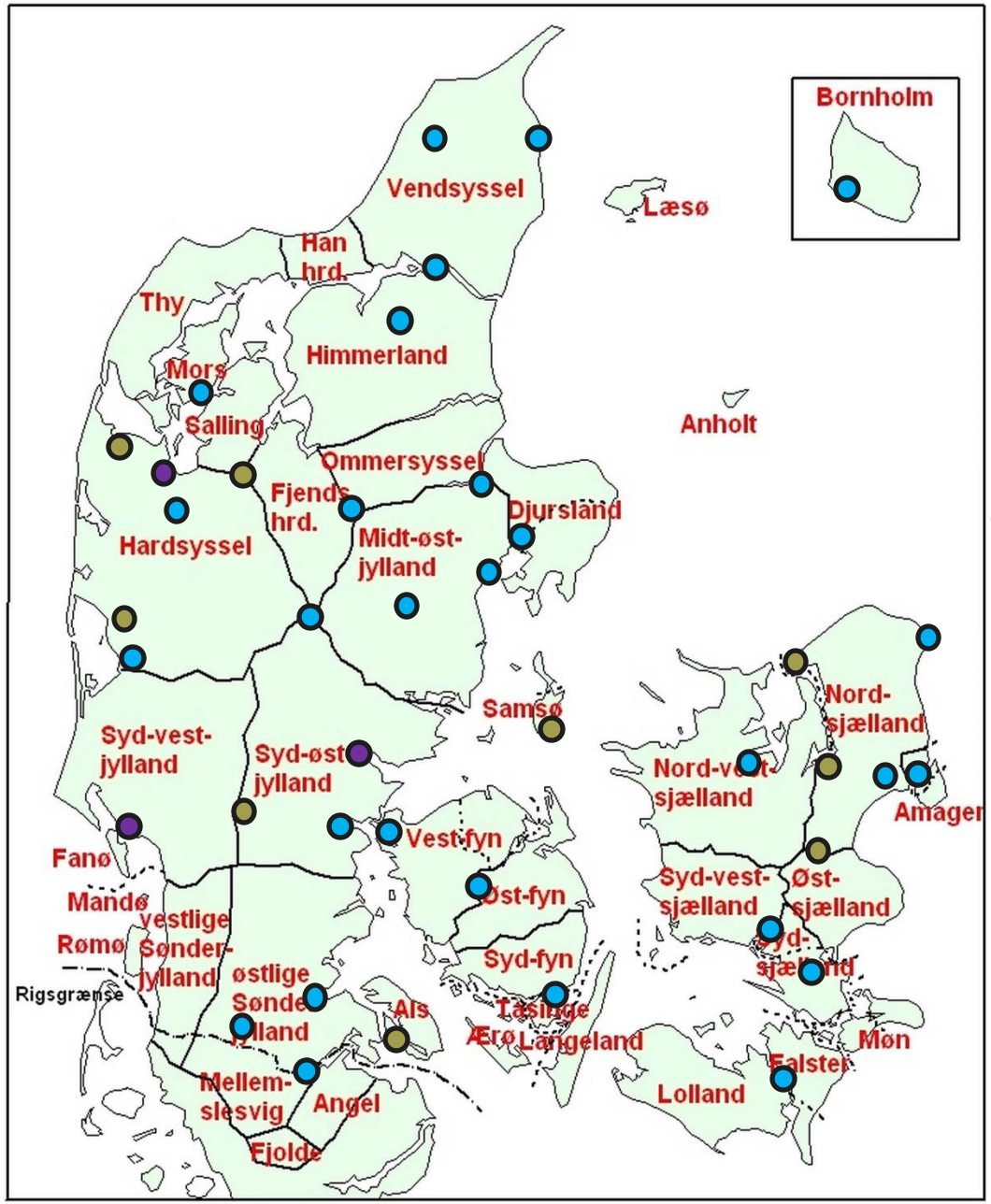
Du donerer omtrent to timer af din tid til samtaler og tekstoplæsning, som optages på Hovedbiblioteket, og til gengæld modtager du et gavekort på 400 kr., som du kan bruge i en stor del af byen butikker.

Du kan læse mere og tilmelde dig via vores hjemmeside her:
https://www.aalborgbibliotekerne.dk/.../lad-din-stemme...

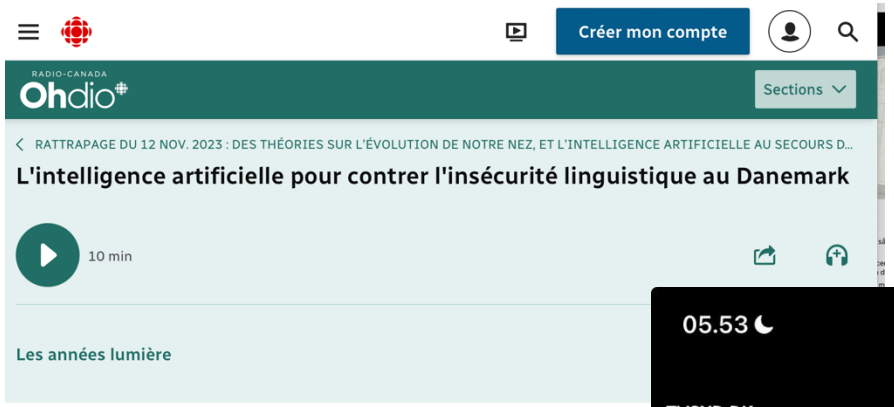
Del gerne opslaget, hvis du kender nogen, der kunne være gode at få med i projektet.

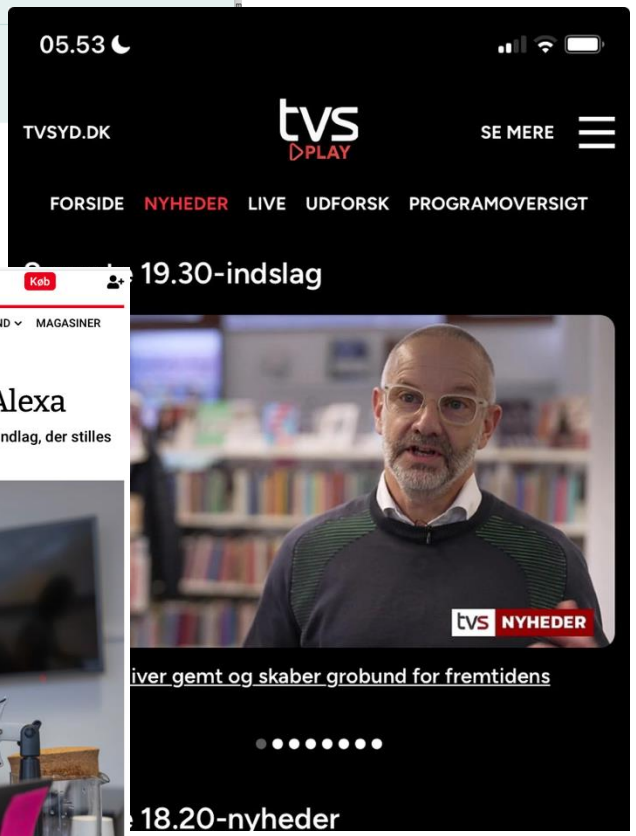


DIALEKT-OMRÅDER I DANMARK



Nordisk Forskningsinstitut
Afd. for Dialektforskning



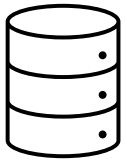




Hvem kommer?



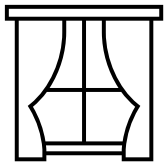
Dataindsamling



~ 567 timer



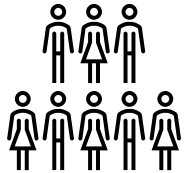
Optagelser
265.000 oplæsninger
556 samtaler



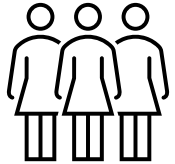
Hjemmeoptagelser
~ 38 personer



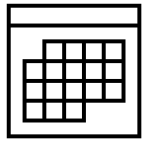
Lokationer
34 forskellige



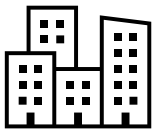
Antal indtalere
593 personer
i alderen 11-87 år



Kønsfordeling
70 % Kvinder

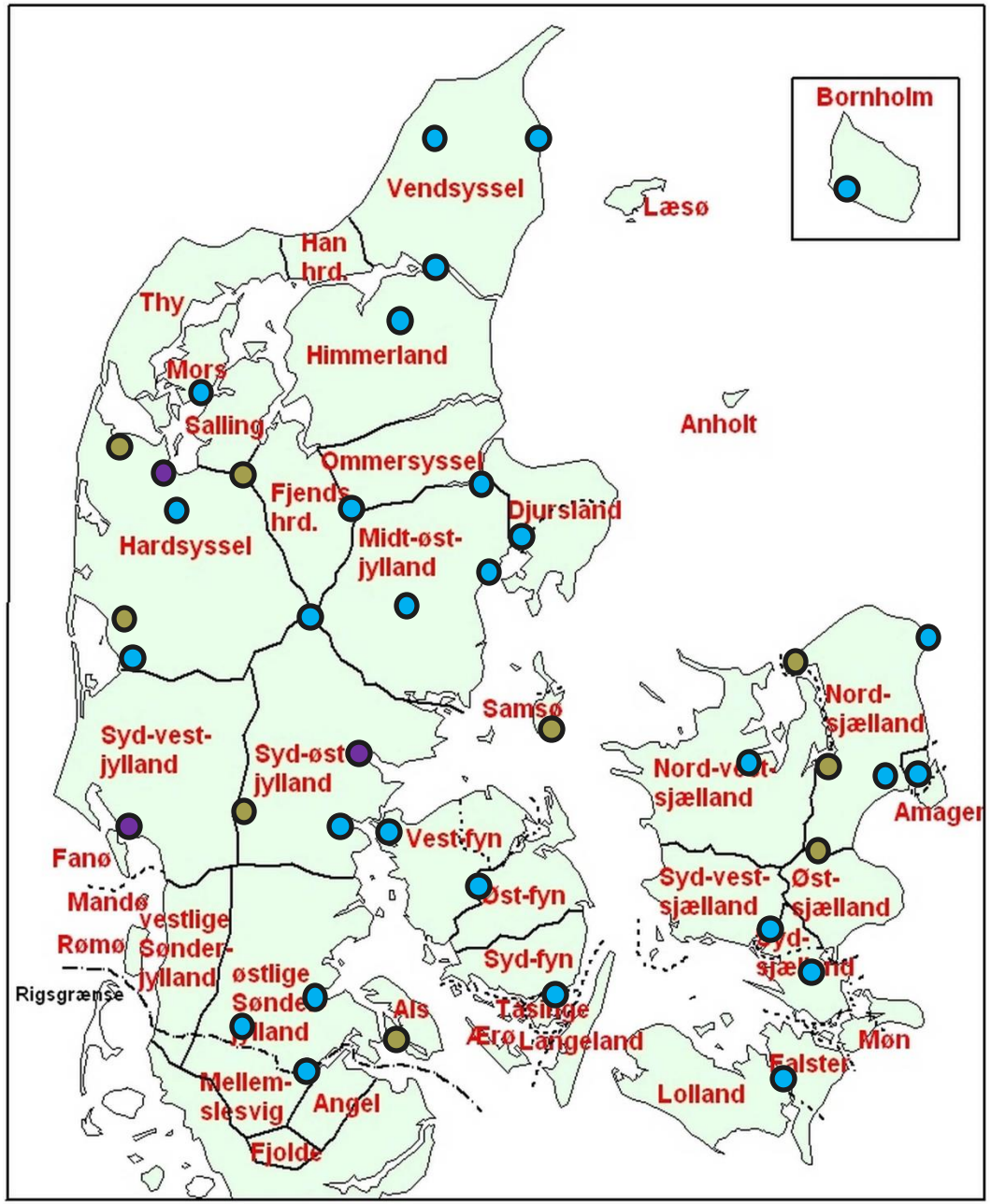


Optagedage
60 dage

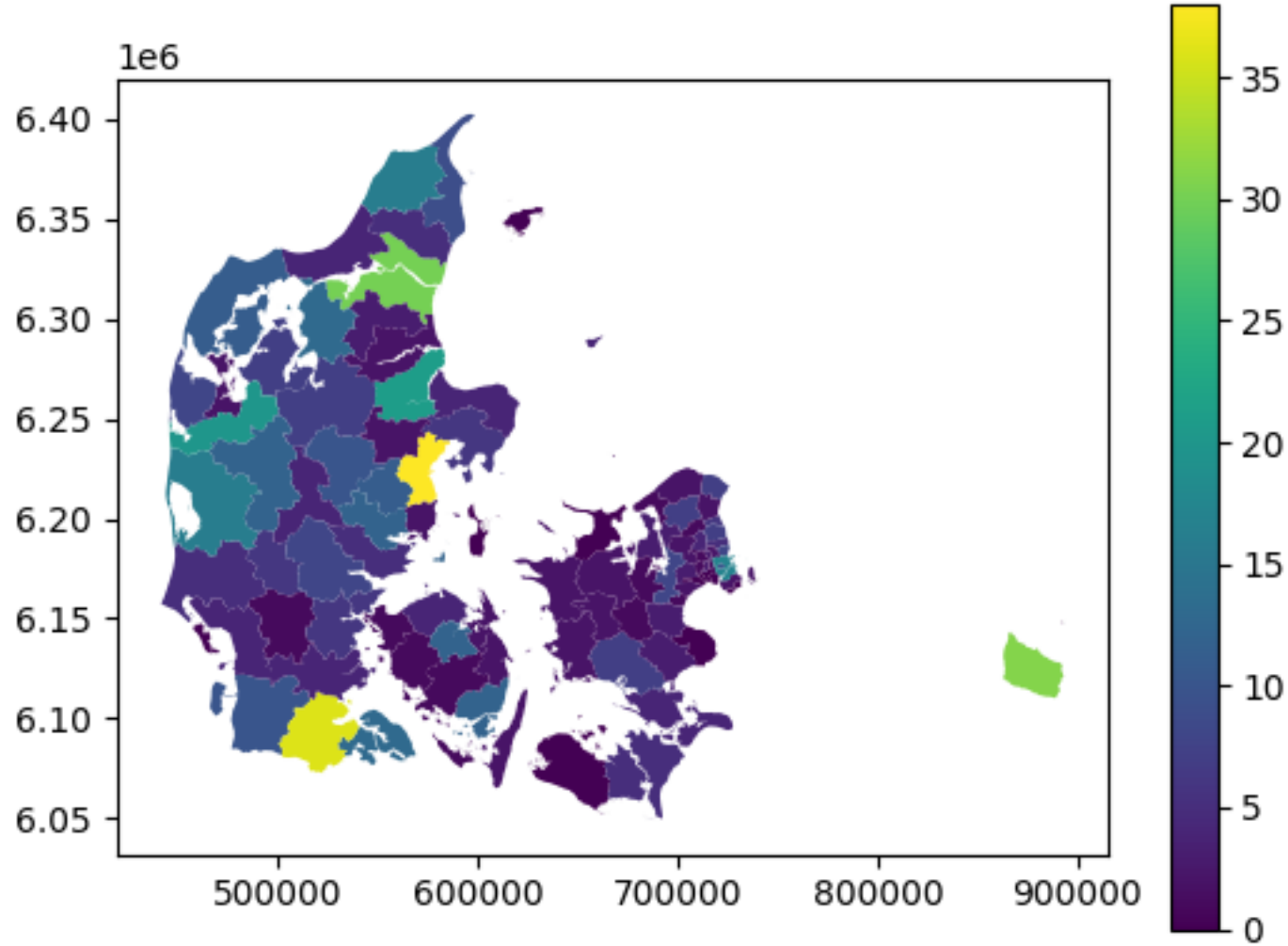


100%
Dækning af alle
dialektområder

DIALEKT-OMRÅDER I DANMARK

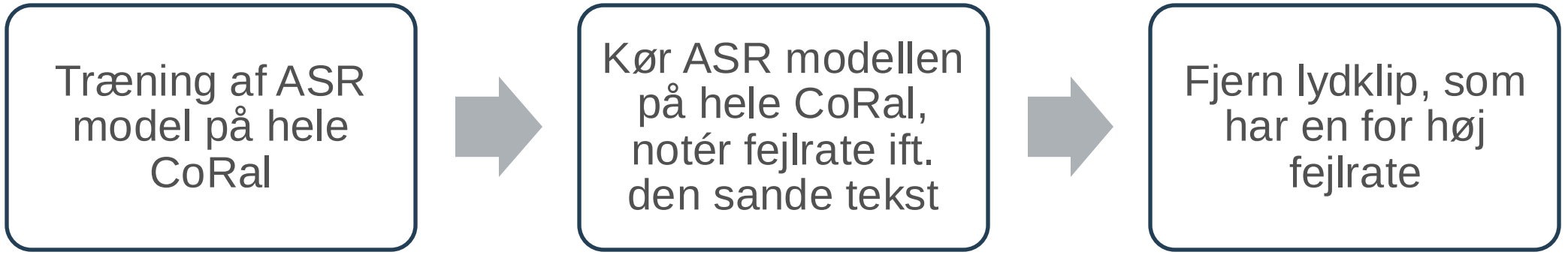


Nordisk Forskningsinstitut
Afd. for Dialektforskning



Frigivelse af første CoRaldatasæt

Automatisk validering



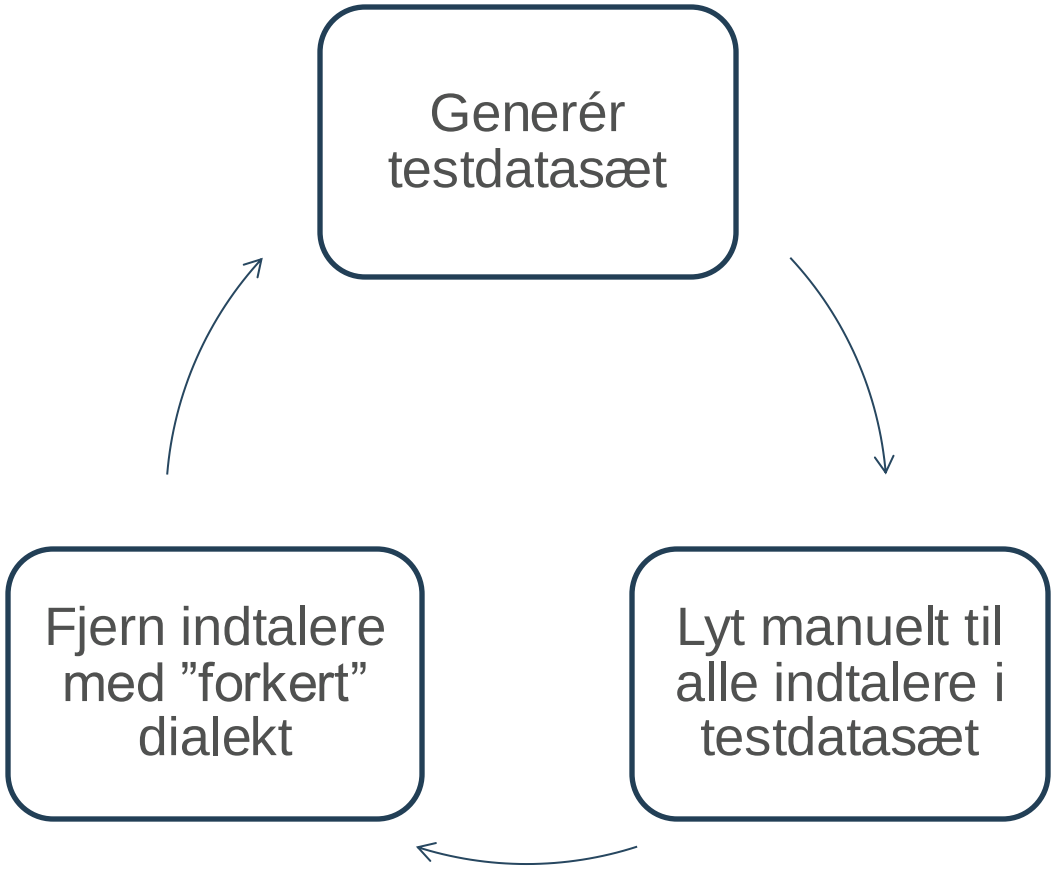
Frigivelse af første CoRaldatasæt

Testsæt (take 1)

- Testsæt automatisk udvalgt ud fra hele CoRal datasættet
- Ligeligt fordelt på tværs af køn, alder, dialekt og accent
- Intet overlap af indtalere mellem træning, validering og test datasættet
- Problem: flere indtalere har ikke den dialekt, som de har rapporteret
  - Dette gjorde testsættet for let, og vi ikke kunne stole på dialektperformance

Frigivelse af første CoRaldatasæt

Testsæt (take 2)



Frigivelse af første CoRaldatasæt

Licens

Datasættet vil blive frigivet under et **OpenRAIL-D** licens.

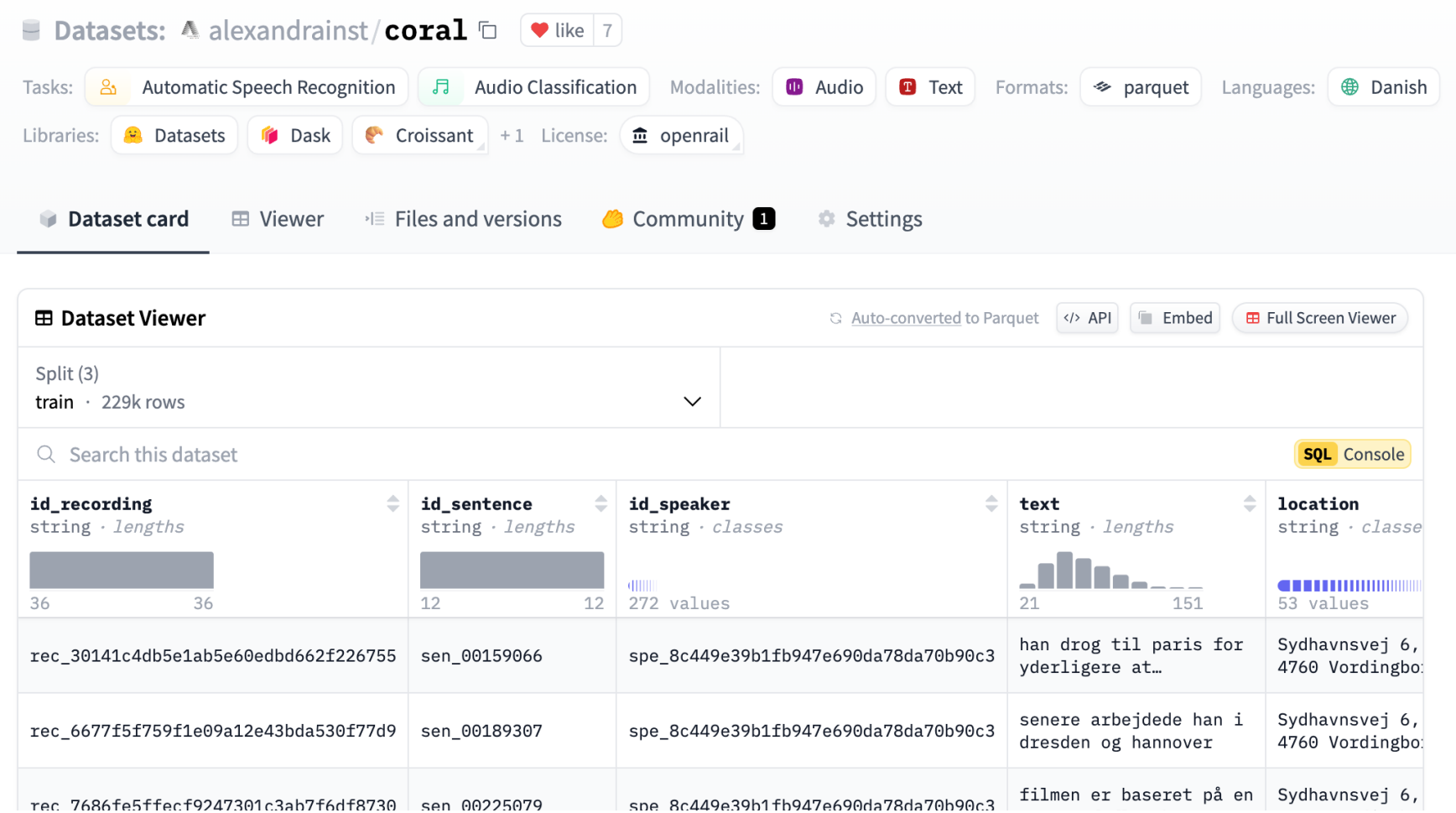
Open = Fri brug og distribution

RAIL = Responsible AI licens

D = Dataset

Frigivelse af første CoRaldatasæt

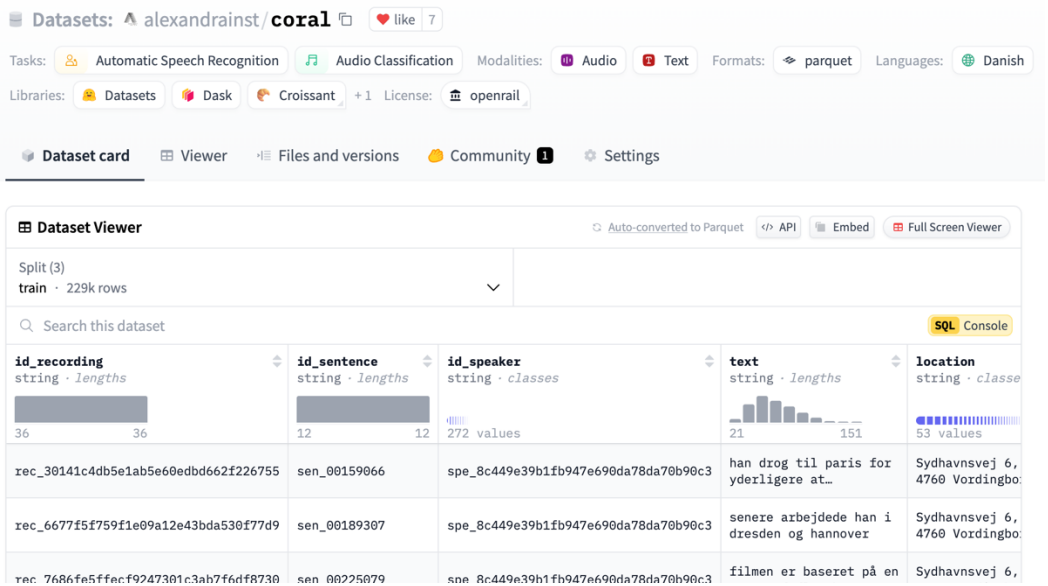
Adgang



Frigivelse af første CoRaldatasæt

Adgang

Kan findes på
https://huggingface.co/datasets/alexandrainst/coral



Frigivelse af første model

Navn

Røst

Frigivelse af første model

Navn

Røst



Frigivelse af første model

Navn

Røst





ALEXANDRA INSTITUTE

Frigivelse af første model

Hvad er modellen trænet på?

- Finjusteret version af wav2vec2 grundmodel fra META
  - Prætrænet på 436.000 timers ikke-transskriberet tale
  - Heraf 13.600 timers dansk tale
- Finjusteret på følgende data:
  - CoRal træningsdatasæt (377 timer)
  - Den danske del af Common Voice 17 (12 timer)

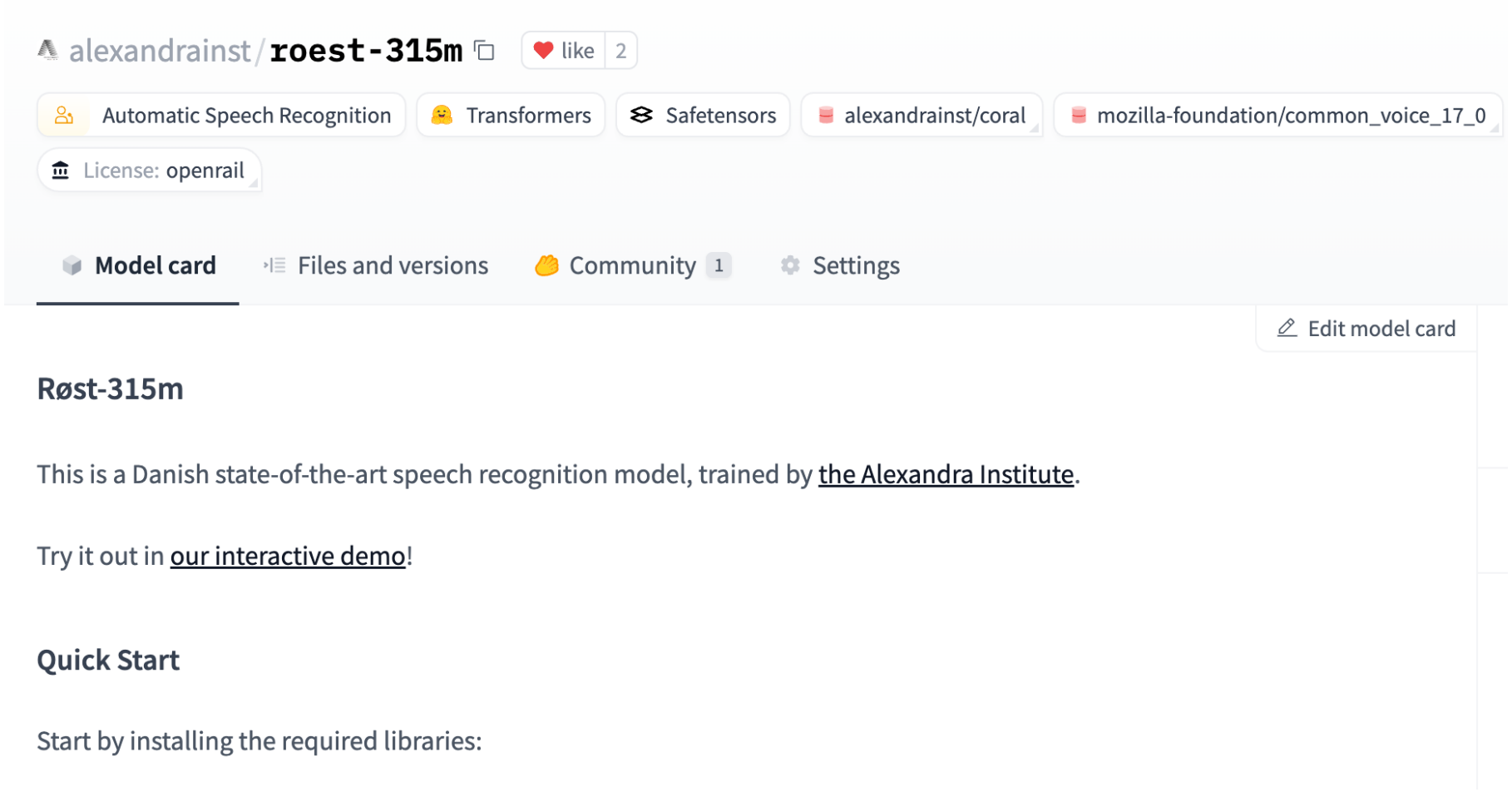
Frigivelse af første model

Performance



Frigivelse af første model

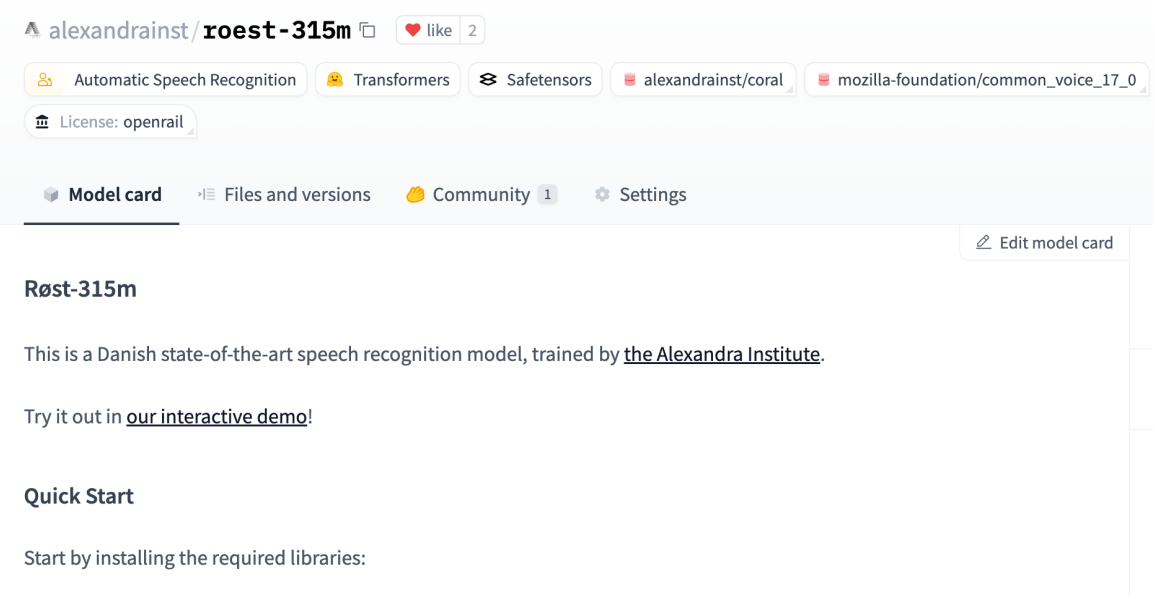
Adgang



Frigivelse af første model

Adgang

Kan findes på hf.co/alexandrainst/roest-315m



Frigivelse af første model

Adgang

st-315m like 2

nition Transformers Safetensors alexandrains

s and versions Community 1 Settings

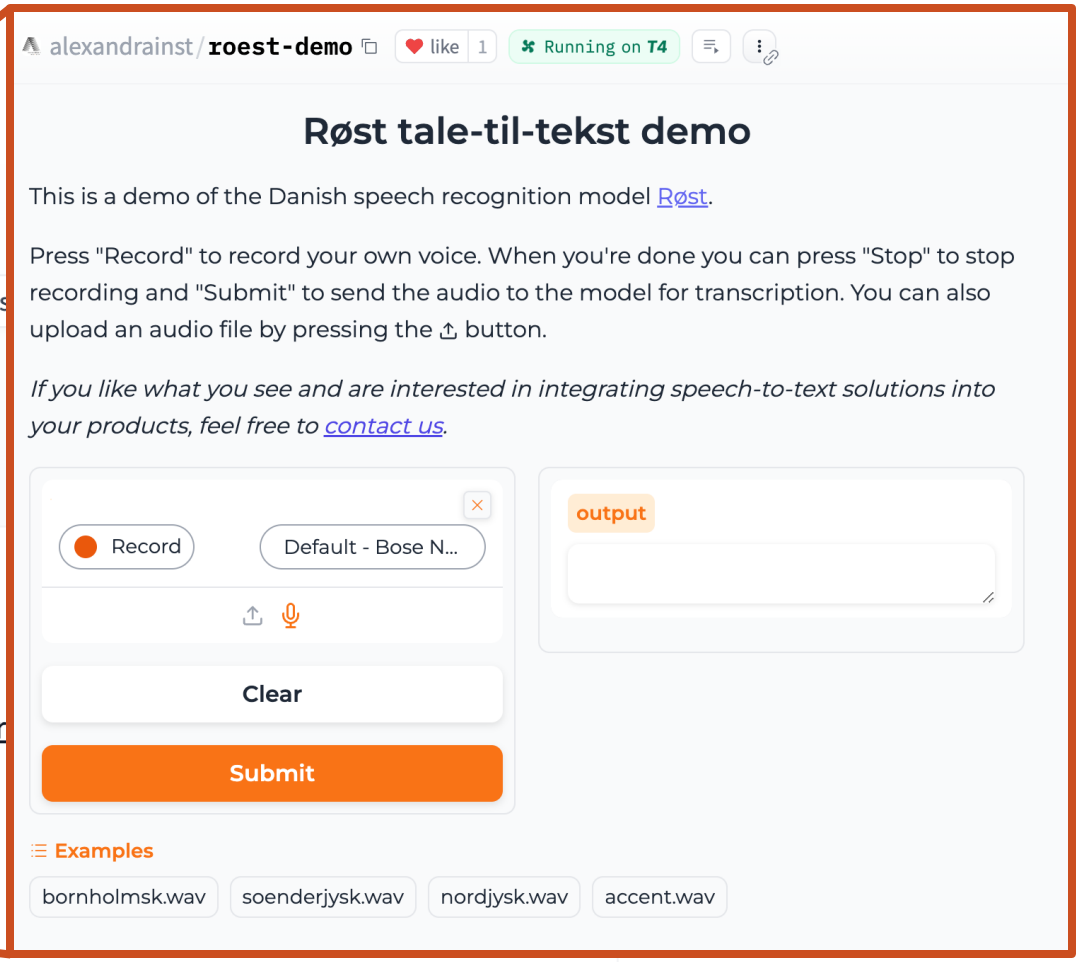
Røst-315m

This is a Danish state-of-the-art speech recognition model, trained by the Alexandr

Try it out in our interactive demo!

Quick Start

Start by installing the required libraries:



Et par dage senere...

Mads Henrichsen · Following
Building DanskGPT - LLM Senpai
1mo · Edited ·

NY MODEL ALARM

For et par dage siden lancerede Alexandra Instituttet deres nye Røst model, baseret på CoRal datasættet.

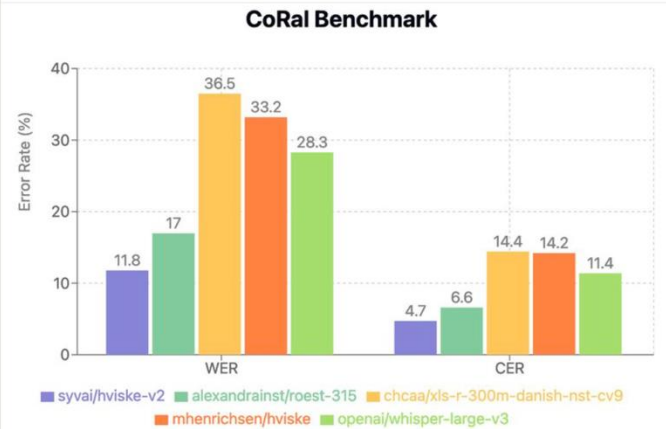
Hos syv.ai har vi også haft snablen nede i det nye datasæt og produceret en Hviske version 2, der nu er den nye state of the art model indenfor ASR.

Highlights:
• Modellen er mere end 30% bedre end nr. Røst målt på WER (dog langsommere).
• Den er 64% bedre end Hviske v1.

Prøv den her:
https://ludwig.syv.ai

Du kan downloade Hviske v2 på HuggingFace:
https://lnkd.in/dDuuMTRy

Show translation



You and 84 others 30 comments · 2 reposts



DANSK SPROGMODEL KONSORTIUM



Large Language Models



I’m gonna

I’m gonna _____

85%	make
85%	think
75%	throw
20%	him
55%	paint
2%	can’t
60%	use

I’m gonna _____

85%	**make**
85%	think
75%	throw
20%	him
55%	offer
2%	can’t
60%	use

I’m gonna make _____

0%	make
10%	think
75%	throw
95%	**him**
55%	offer
2%	can’t
60%	use

I’m gonna make him ____

0%	make
98%	**an**
75%	throw
25%	him
55%	offer
2%	can’t
60%	use

I’m gonna make him an offer he can’t _____

98%	make
6%	an
75%	throw
2%	him
65%	offer
0%	can’t
99%	**refuse**

I’m gonna make him an offer he can’t refuse

77%	make
6%	an
75%	throw
2%	him
65%	offer
0%	can’t
99%	refuse

the probability of the next _____

the probability of the next token


> 1,000,000,000,000
tokens (trillion)

Where is all this data coming from?





COMMON CRAWL

Started 2008

6.103.515 GB
of “open” web data



Eric Schmidt,
ex-CEO Google

> *[…] if it took off, then you'd hire a whole bunch of lawyers to go clean the mess up, right? […]*

April 2024

IS IT OKAY TO STEAL?



Why not let the tech giants do the job for us



US CONTROLED & CLOSED













CRITICAL INFRASTRUCTURE



LLMs has an increasing role in Modern Society

- Security (Cybersecurity, Tele communication, Healthcare
- Economy (as driver for growth)
- Impacts on Public Information and Discourse



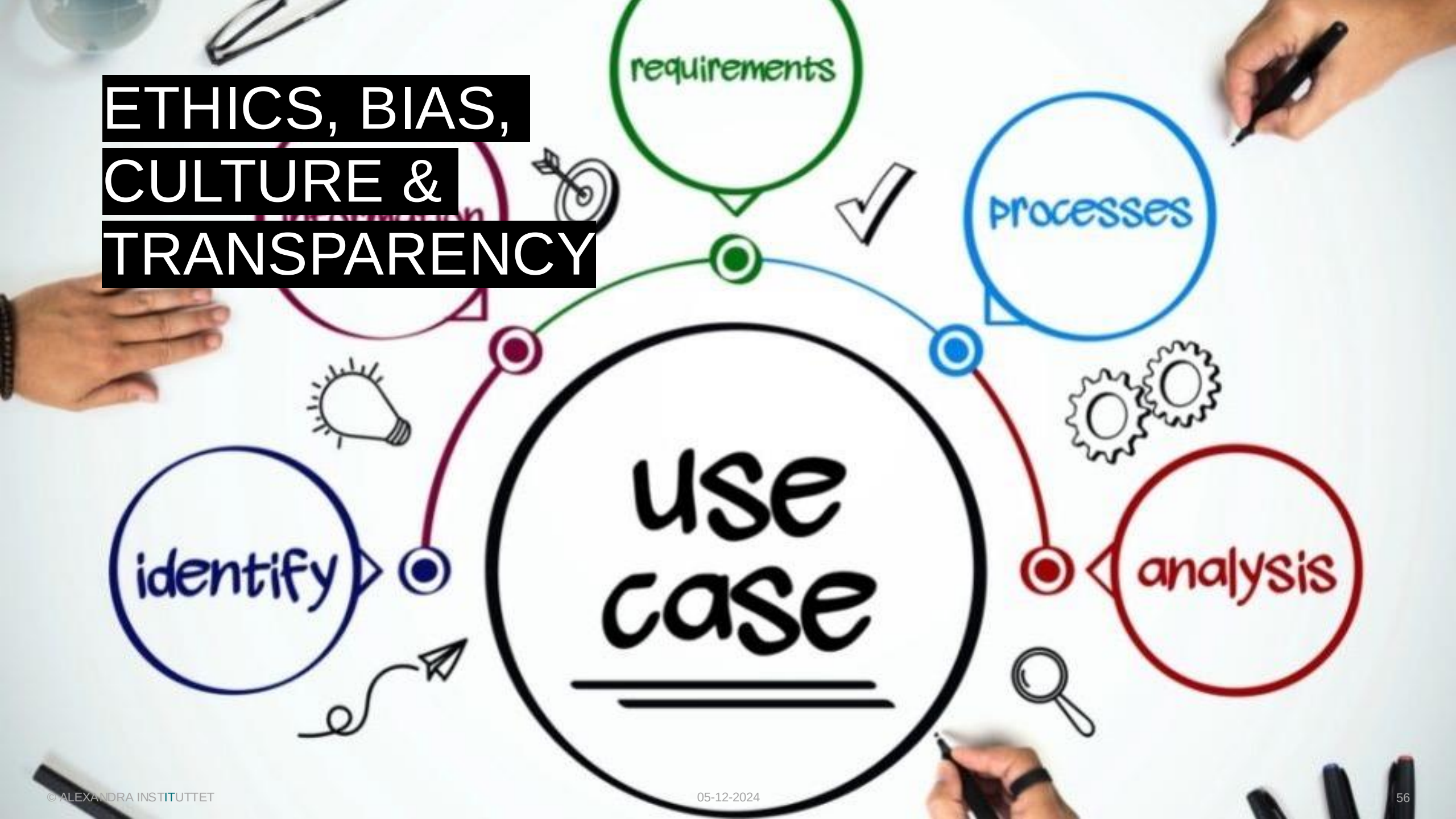






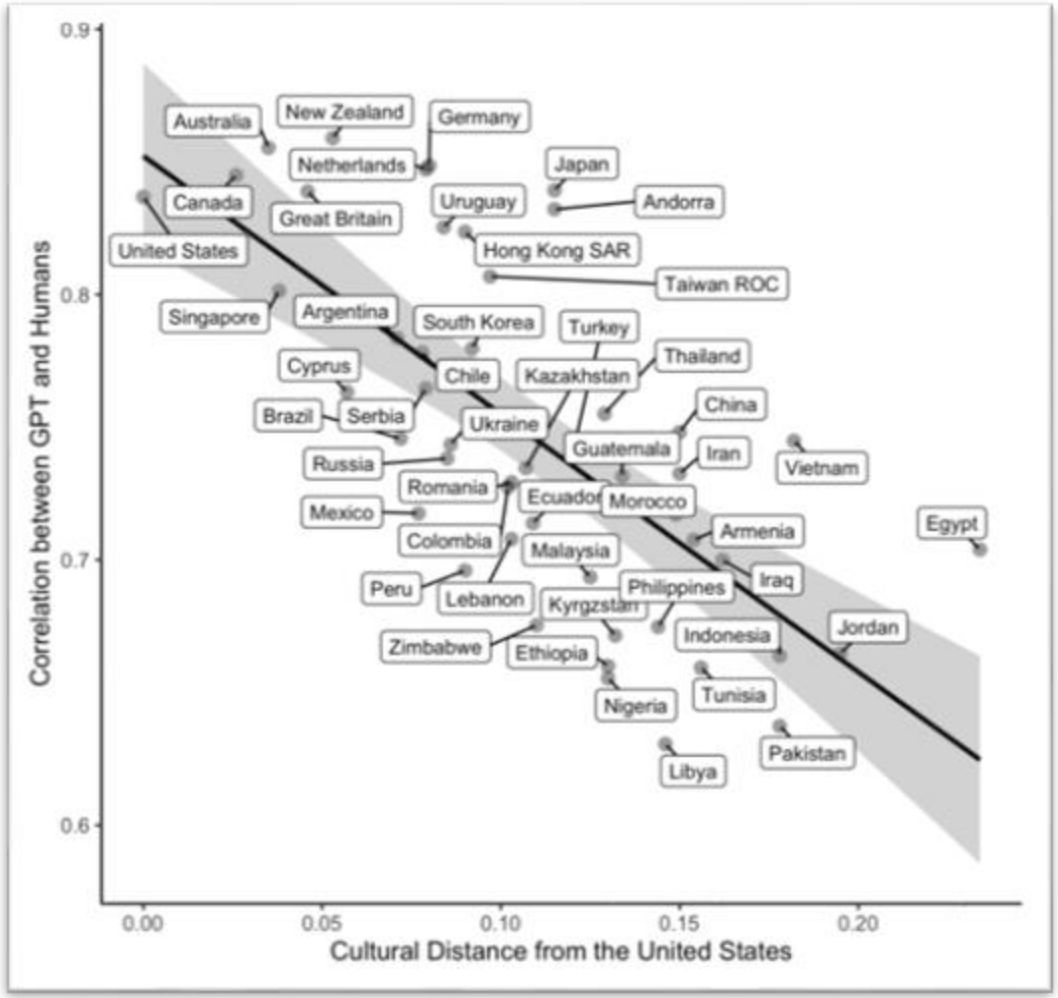


ETHICS, BIAS, CULTURE & TRANSPARENCY



A few comment on Bias ??

- **The *Pizza Battle***: “What’s the best pizza topping?”
- **The *Dog vs. Cat Fight***: "Are dogs better than cats?”
- **The *Mom Bias***: "What do moms do?"

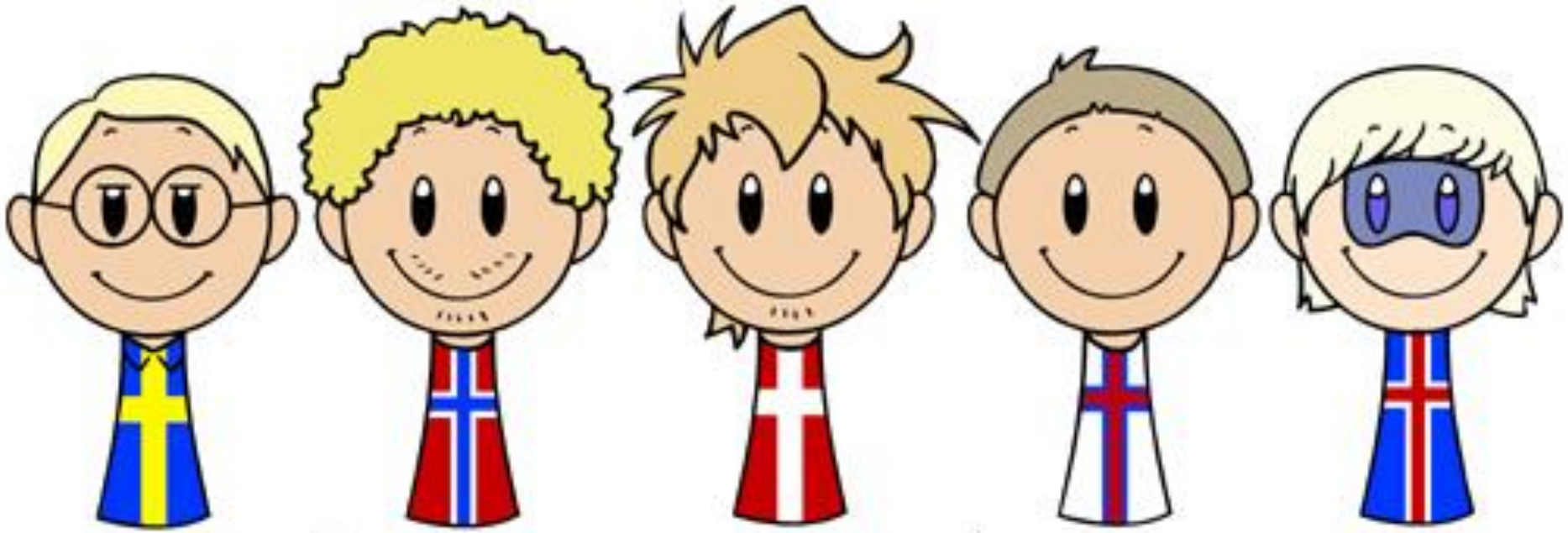


Kilde: J Bossman, WEF


RESEARCH
ECOSYSTEM

What about Big Tech and the Danish language





ScandEval

© ALEXANDRA INSTITUTTET



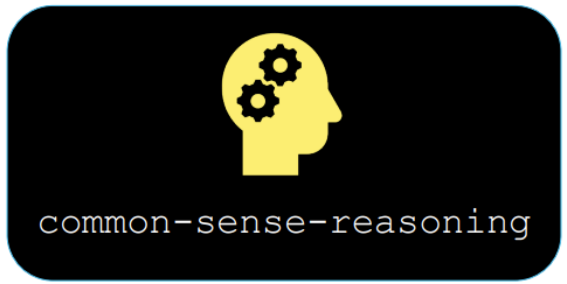




ScandEval currently supports 7 different tasks

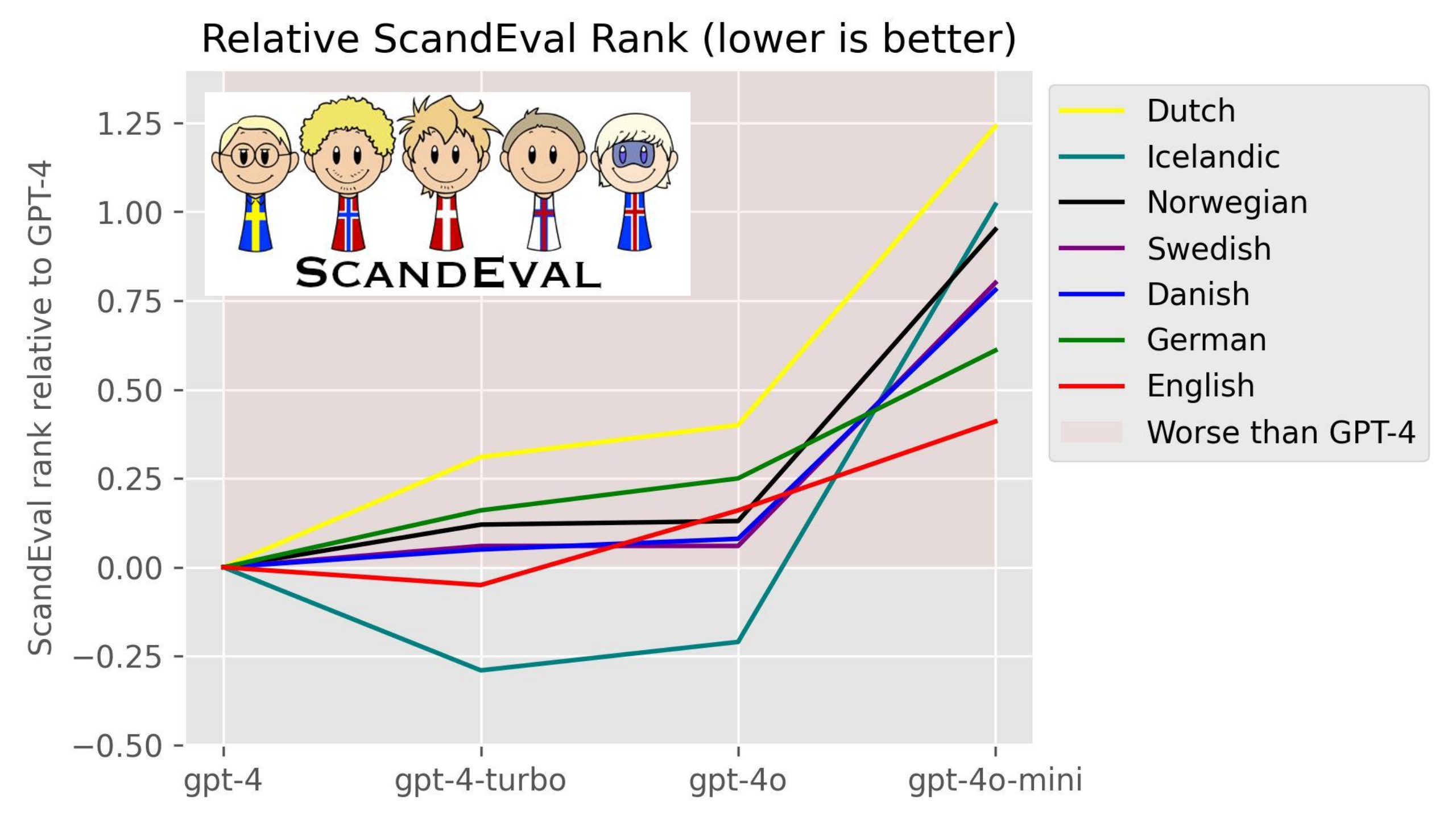








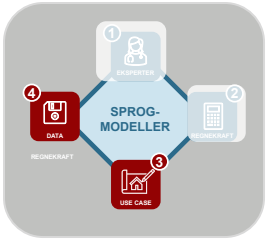
© ALEXANDRA INSTITUTTET


Relative ScandEval Rank (lower is better)
ScandEval rank relative to GPT-4
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
gpt-4
gpt-4-turbo
gpt-4o
gpt-4o-mini
SCANDEVAL
Dutch
Icelandic
Norwegian
Swedish
Danish
German
English
Worse than GPT-4

Are we doomed in Denmark



DANSK
SPROGMODEL
KONSORTIUM



IBM

ALEXANDRA
INSTITUTTET

DANSK
ERHVERV


> 1,000,000,000,000
tokens (trillion)

DSK – DFM - MUNIN

Dansk Sprogmodel Konsortium

Deling af relevante usecases samt deling af data til træning af DFM modellerne



USE CASES & DATA

Danish Foundation Models

Udvikling af åbne danske sprogmodeller til det danske samfund



LARGE LANGUAGE MODELS

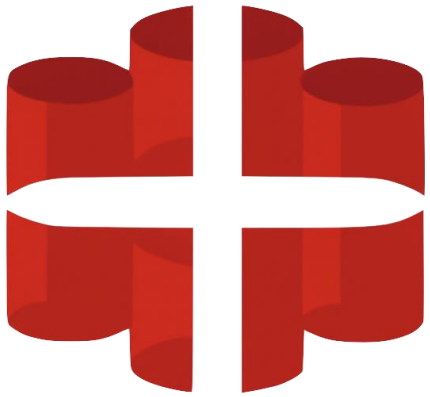
MUNIN Sprogmodeller

MUNIN

Frigivet open source Sprogmodeller til gratis benyttelse

© ALEXANDRA INSTITUTTET

Danish Foundation Models



Danish Foundation Models

AARHUS UNIVERSITET

KØBENHAVNS UNIVERSITET

SDU University of Southern Denmark

ALEXANDRA INSTITUTTET

© ALEXANDRA INSTITUTTET

Source: https://www.foundationmodels.dk/

VALUE ADDING USE-CASES




2021.AI
AARHUS KOMMUNE
ALEXANDRA INSTITUTTET
aeven
atp
BAGGER·SØRENSEN
cBRAIN The Process Company
DANSK ERHVERV
DLA PIPER
FALCK
DANSKE REVISORER FSR
HOFOR
IBM
JN DATA
KL
KOMBIT
NORLYS
ordbogen.com
salling group
SDC
Topdanmark
VISMA | Enterprise
REGION NORDJYLLAND
DK Medier.


.OUTSIDE.
.INSIDE.

DANSK
SPROGMODEL
KONSORTIUM

Anyone can become a member if you want to donate data

Join the Consortium

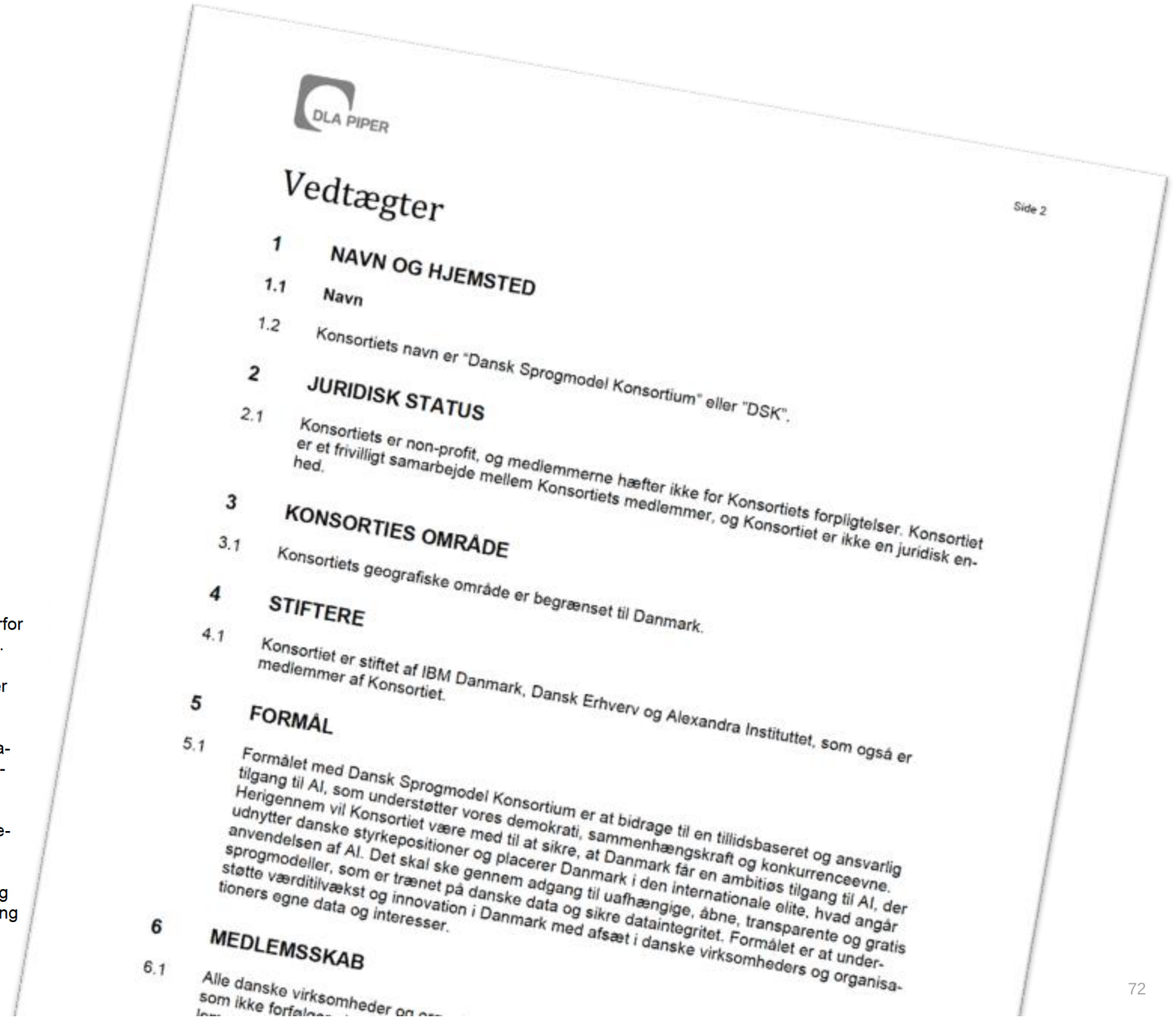
1. Åbne og gratis: De udviklede danske sprogmodeller skal være open source og derfor til enhver tid være fri og gratis tilgængelige for alle, og de kan anvendes kommercielt.

2. Overholde lovgivning: De anvendte data og udviklede sprogmodeller skal til enhver tid overholde dansk lov og EU-lovgivning, herunder GDPR og AI Forordningen.

3. Data: Modellerne skal trænes på datasæt, som er godkendt til formålet af opretshaverne. Data skal filtreres for personinformation, inden de må bruges til træning af modellerne. Data må ikke forlade EU.

4. Transparens og etik: Der skal være transparens i forhold til datakilder og modeltræning, samt sikkerhed for at danske værdier og etik er indarbejdet.

5. Sikkerhed: Der implementeres robuste sikkerhedsprotokoller, der beskytter data og modellerne mod potentielle sårbarheder, samt sikrer, at data kun anvendes til udvikling af danske sprogmodeller.


DLA PIPER

Vedtægter

Side 2

1 NAVN OG HJEMSTED

1.1 Navn

1.2 Konsortiets navn er "Dansk Sprogmodel Konsortium" eller "DSK".

2 JURIDISK STATUS

2.1 Konsortiets er non-profit, og medlemmerne hæfter ikke for Konsortiets forpligtelser. Konsortiet er et frivilligt samarbejde mellem Konsortiets medlemmer, og Konsortiet er ikke en juridisk enhed.

3 KONSORTIES OMRÅDE

3.1 Konsortiets geografiske område er begrænset til Danmark.

4 STIFTERE

4.1 Konsortiet er stiftet af IBM Danmark, Dansk Erhverv og Alexandra Instituttet, som også er medlemmer af Konsortiet.

5 FORMÅL

5.1 Formålet med Dansk Sprogmodel Konsortium er at bidrage til en tillidsbaseret og ansvarlig tilgang til AI, som understøtter vores demokrati, sammenhængskraft og konkurrenceevne. Herigennem vil Konsortiet være med til at sikre, at Danmark får en ambitiøs tilgang til AI, der udnytter danske styrkepositioner og placerer Danmark i den internationale elite, hvad angår anvendelsen af AI. Det skal ske gennem adgang til uafhængige, åbne, transparente og gratis sprogmodeller, som er trænet på danske data og sikre dataintegritet. Formålet er at understøtte værditilvækst og innovation i Danmark med afsæt i danske virksomheders og organisationers egne data og interesser.

6 MEDLEMSSKAB

6.1 Alle danske virksomheder og
som ikke forfølger

ALEXANDRA INSTITUTTET

Brancher Services Teknologier Referencer Bliv klogere Om os Kontakt

DANSK SPROGMODEL KONSORTIUM

Kom med i **Dansk Sprogmodel Konsortium**

Vil I være med til at sætte skub i udviklingen af ansvarlige, danske sprogmodeller?



til danske data, så sprogmodellerne trygt kan anvendes. I dag gør mange allerede brug af sprogmodeller, der dog ofte primært er trænet på store sprog som engelsk. De modeller kan derudover være underlagt andre værdisæt og anden regulering.

*Principper og forpligtelser

Bliv en del af konsortiet

TILMELD JER HER

© ALEXANDRA INSTITUTTET

5-12-2024

alexandra.dk/dsk/

ALEXANDRA INSTITUTTET

ABOUT

*The Alexandra Institute is a non-profit commercial **consultant** company that works with **applied IT research and development**. Our mission is to create value, growth and welfare in our society with innovative, IT-based products and services.*

CLIENT EXAMPLES

TORBEN BLACH
Senior Projekt- og Programleder
AI Lab, Århus

Torben.blach@alexandra.dk
+45 9155 8799

Sammen kommer vi #forandigitalt

ALEXANDRA INSTITUTTET