

# Artificial Intelligence at the National Library of Norway

Javier de la Rosa  
[versae@nb.no](mailto:versae@nb.no)



LUSTRE Workshop  
27-28 June 2024, London



The Mountain Vault, Mo i Rana, Norway



How did we get here?

The Mountain Vault, Mo i Rana, Norway



# Nasjonalbiblioteket

Established in 1989

Preserve published information for posterity

Give Norwegians access to its content



# Manifesto in 2005

## Recognition

"If it doesn't exist on the Internet, it almost doesn't exist at all"

## Ambition

"Access to content where the user is, when the user wants"

## Conclusion

"We must digitize the collection"

### The Digital National Library of Norway

Strategic platform 2005  
The National Library of Norway, February 2005

#### The Library Comes to the User

Mature information—and communication technology and the Internet—have given the library new possibilities when it comes to professional cooperation and a coordinated use of resources. The most distinctive opportunity is the possibility of providing access to a rich and varied content on everyone's desk and in every person's home. By focusing on the user in digital library development, the National Library of Norway will contribute to new and better services to benefit society and improve services offered by the library community.

The Norwegian Digital Library (NDL) is central in this effort, and both the content and the services in NDL will be strengthened by the Norwegian Library's (NL) Digital Library initiative. The NL represents the centre of gravity in the development of a strong Norwegian digital library. The Norwegian libraries will be strengthened as a consequence of the services established by the NL, services which other libraries can use as a base for their own activities in offering digital user services.

#### Knowledge and Experience

The NL will provide users with a plethora of content in digital form. Through a focused effort to develop digital library services, the NL will enable use and reuse of content from the NL in teaching and learning, in the production of knowledge, and also in the commercial arena. The collections will be available where and when it suits the user. The NL has chosen to focus its efforts along three main axes:

- **Presence:** Metadata and content must be presented in ways users prefer and in a context which is tailored for their needs. The user must be able to search both in metadata and content using his preferred tools<sup>1</sup>. NL will make it possible to regulate access by users via existing common platforms in society<sup>2</sup>.
- **Learning and knowledge:** The availability of resources within the NL for learning and research will be improved. Tools and mechanisms for e-learning and immediate access to high quality copies of originals are among the top priorities. This will also pave the way for commercial use of the contents of the NL.
- **Edutainment:** Information and knowledge must also be presented in a context that gives the users new experiences and knowledge. Services that make information and knowledge available, in the shape of exhibitions and adventure trips on the Internet, shall be developed continuously.



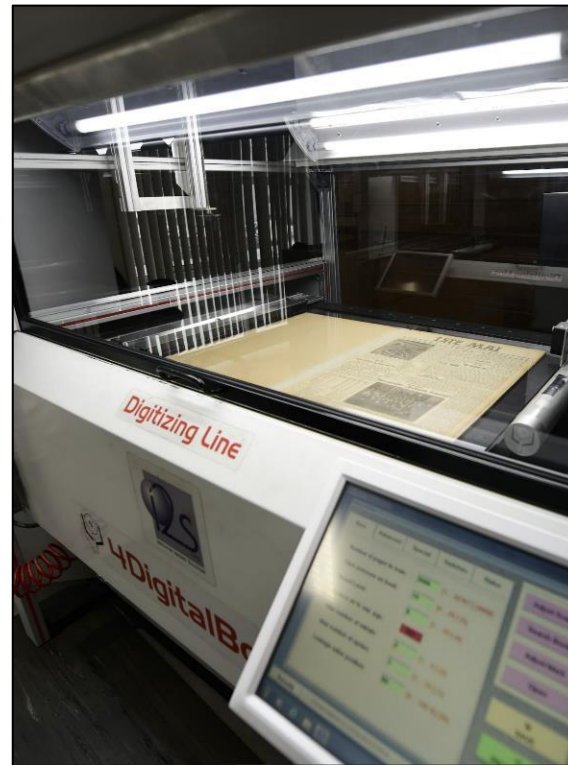
<sup>1</sup> Studies show that even researchers prefer to use general search engines like Google and Yahoo in stead of dedicated services when seeking information on the internet.

<sup>2</sup> Uninett ([www.uninett.no](http://www.uninett.no)) is establishing FEIDE, which is a common platform for authentication in the education and research sector in Norway.

# NB's digital programme: digital collection today

## Some examples

- Books: **90%** digitized
- Newspapers: **80%** are digitized or digitally delivered
- Radio: **Almost everything** is digital
- Web: **continuous** harvest
- Much more: photo, film, magazine, maps, etc.



A perspective view of a long, empty aisle in a server room. The aisle is flanked by rows of server racks on both sides, extending far into the distance. The racks are dark grey or black, and the floor is light-colored. The lighting is bright and even, creating a sense of depth and repetition.

What are we doing about it?

The Mountain Vault, Mo i Rana, Norway

# We at the NB AI-Lab

## Background

- Very large and varied digital collection in NB
- Own needs for improvement in production and user services
- Contribute knowledge about AI

## The people

- 4 full-time, possibility for "researchers in residence"
- Currently 1 researcher in residence

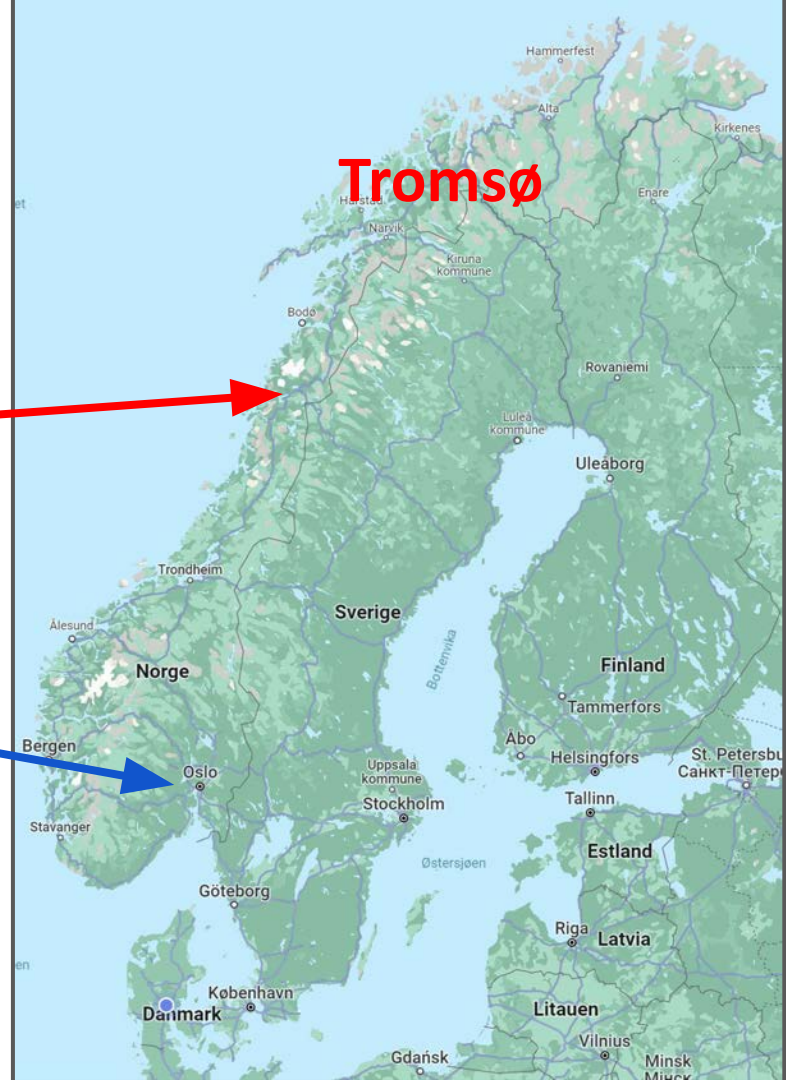


Where?

Mo i Rana



Oslo



Where?

Mo i Rana



Madrid



Oslo



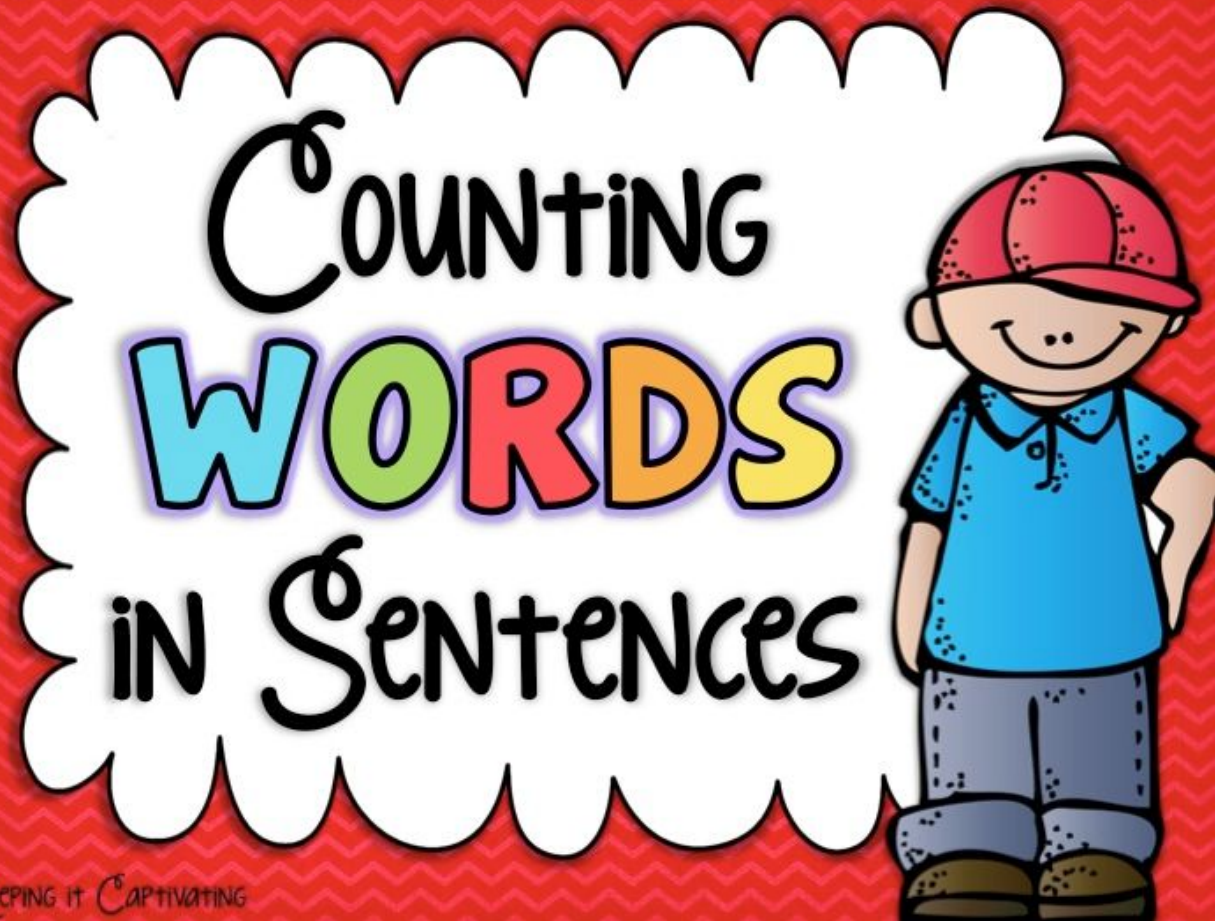
# What does the AI-Lab deliver?

- Internal and external knowledge
- Training datasets based on NB's collections
- Fully trained models for free use across modalities
- Services and APIs
- Support for internal workflows
- Experiments, pilots, and demonstrators
- Basic research

# How?



It all started with...

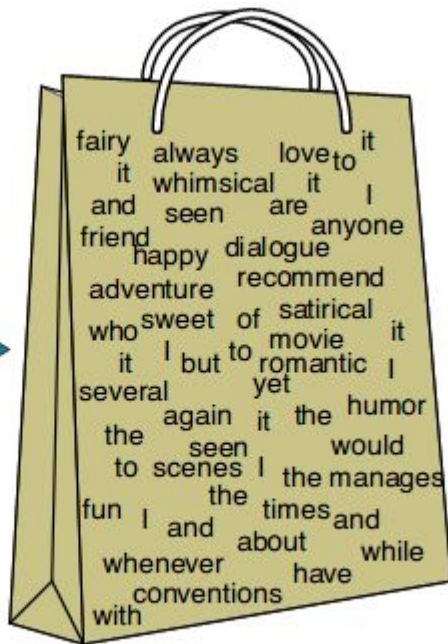


KEEPING IT CAPTIVATING

# From string of characters to numbers

## Bag of Words (BoW)

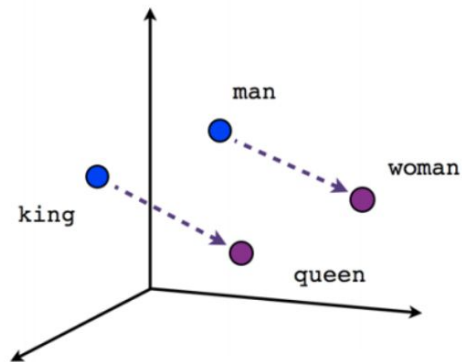
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



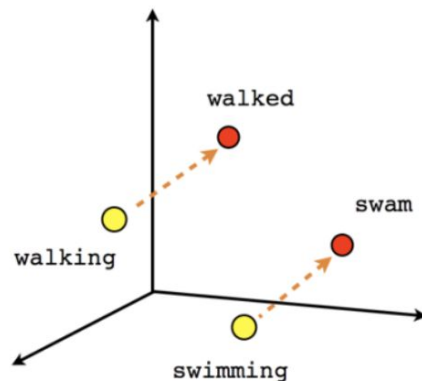
|           |     |
|-----------|-----|
| it        | 6   |
| I         | 5   |
| the       | 4   |
| to        | 3   |
| and       | 3   |
| seen      | 2   |
| yet       | 1   |
| would     | 1   |
| whimsical | 1   |
| times     | 1   |
| sweet     | 1   |
| satirical | 1   |
| adventure | 1   |
| genre     | 1   |
| fairy     | 1   |
| humor     | 1   |
| have      | 1   |
| great     | 1   |
| ...       | ... |

# From string of characters to numbers

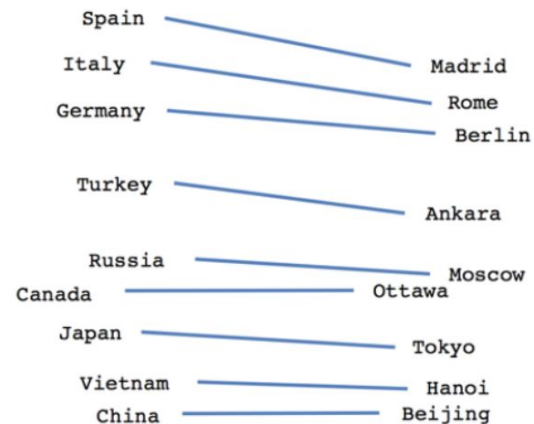
## Word2Vec



Male-Female



Verb tense



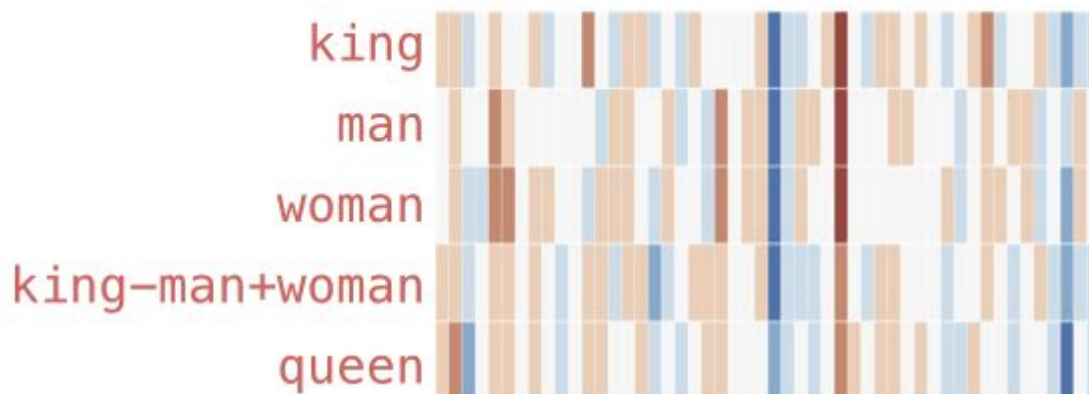
Country-Capital

Source: <https://thegradiant.pub/nlp-imagenet/>

# From string of characters to numbers

Word2Vec

king - man + woman  $\approx$  queen



Source: <http://jalamar.github.io/illustrated-word2vec/>



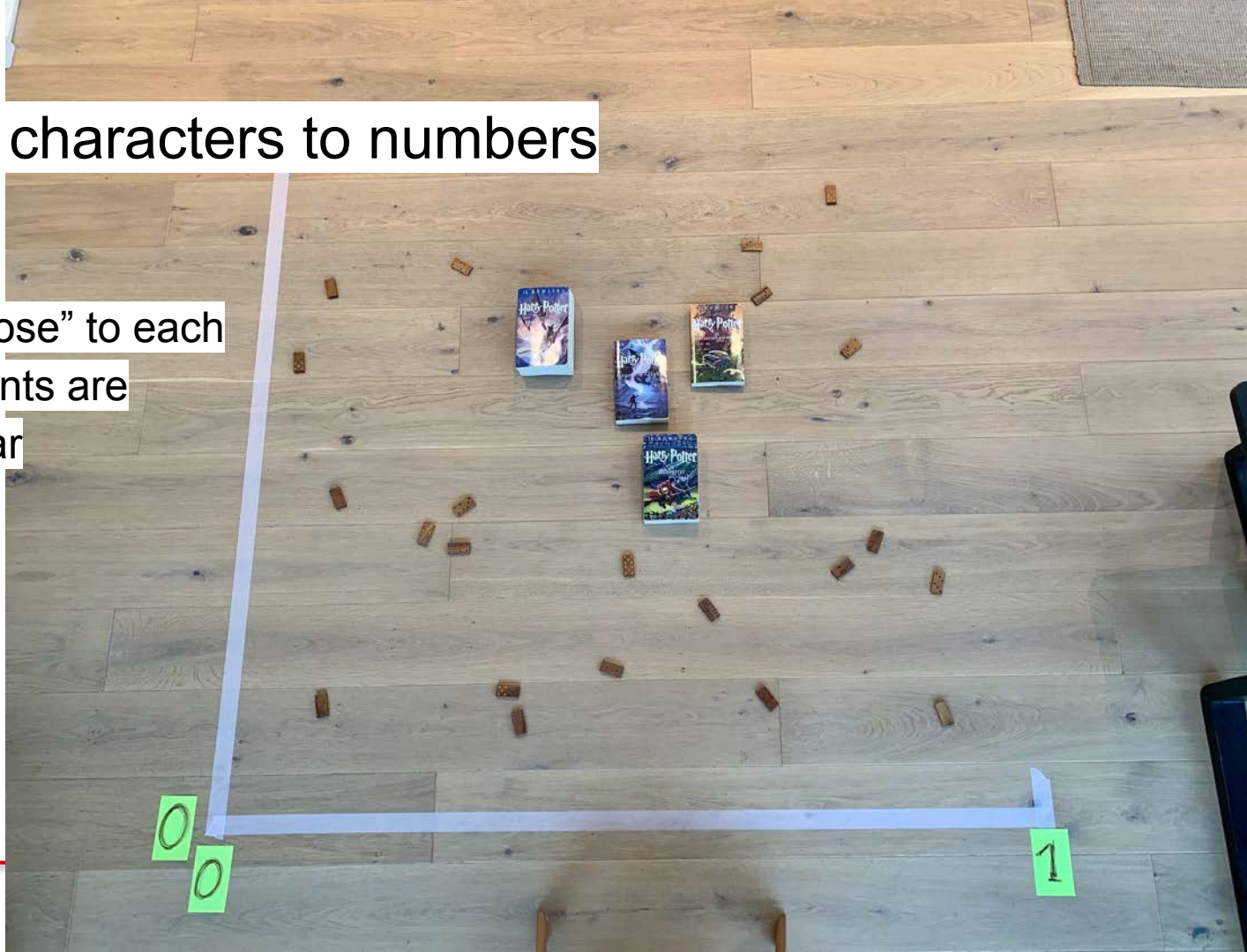
# From string of characters to numbers

## Doc2Vec

- Corpus of documents
- Model learns a numerical representation for each document (training)
- Model is able to generate representations for unseen documents (inference)
- Each representation contains values that “embed” the document
  - o Document embeddings!

# From string of characters to numbers

Embeddings are “close” to each other when documents are “semantically” similar



# Searching for similarities

1. Find an object, book or image



# Searching for similarities

1. Find an object, book or image
2. Get its numerical representation



{0.9, 0.12, 0.45, ...}



# Searching for similarities

1. Find an object, book or image
2. Get its numerical representation
3. Compare to the numerical representation of all the other objects



{0.9, 0.12, 0.45, ...}

{0.9, 0.12, 0.45, ...}  
{0.1, 0.23, 0.27, ...}  
{0.2, 0.56, 0.87, ...}  
{0.9, 0.78, 0.62, ...}  
{0.6, 0.29, 0.32, ...}  
{0.9, 0.26, 0.79, ...}  
{0.9, 0.69, 0.16, ...}  
{0.8, 0.82, 0.87, ...}  
{0.5, 0.35, 0.10, ...}  
{0.3, 0.72, 0.97, ...}

...

# Searching for similarities

1. Find an object, book or image
2. Get its numerical representation
3. Compare to the numerical representation of all the other objects
4. Rank by the comparison metric



{0.9, 0.12, 0.45, ...}

{0.9, 0.12, 0.45, ...} → **0.9**  
{0.1, 0.23, 0.27, ...} → 0.1  
{0.2, 0.56, 0.87, ...} → 0.2  
{0.9, 0.78, 0.62, ...} → **0.8**  
{0.6, 0.29, 0.32, ...} → 0.4  
{0.9, 0.26, 0.79, ...} → **0.7**  
{0.9, 0.69, 0.16, ...} → 0.1  
{0.8, 0.82, 0.87, ...} → 0.1  
{0.5, 0.35, 0.10, ...} → 0.2  
{0.3, 0.72, 0.97, ...} → 0.4

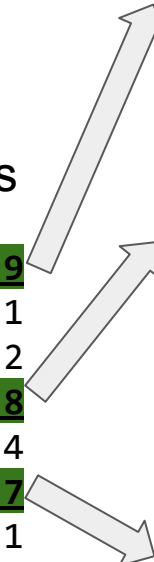
# Searching for similarities

1. Find an object, book or image
2. Get its numerical representation
3. Compare to the numerical representation of all the other objects
4. Rank by the comparison metric



{0.9, 0.12, 0.45, ...}

{0.9, 0.12, 0.45, ...} → 0.9  
{0.1, 0.23, 0.27, ...} → 0.1  
{0.2, 0.56, 0.87, ...} → 0.2  
{0.9, 0.78, 0.62, ...} → 0.8  
{0.6, 0.29, 0.32, ...} → 0.4  
{0.9, 0.26, 0.79, ...} → 0.7  
{0.9, 0.69, 0.16, ...} → 0.1  
{0.8, 0.82, 0.87, ...} → 0.1  
{0.5, 0.35, 0.10, ...} → 0.2  
{0.3, 0.72, 0.97, ...} → 0.4



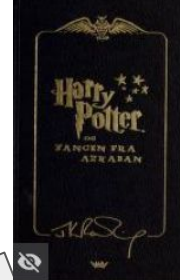
# Searching for similarities

[nb.no/maken](https://nb.no/maken)



{0.9, 0.12, 0.45, ...}

|                        |   |            |
|------------------------|---|------------|
| {0.9, 0.12, 0.45, ...} | → | <b>0.9</b> |
| {0.1, 0.23, 0.27, ...} | → | 0.1        |
| {0.2, 0.56, 0.87, ...} | → | 0.2        |
| {0.9, 0.78, 0.62, ...} | → | <b>0.8</b> |
| {0.6, 0.29, 0.32, ...} | → | 0.4        |
| {0.9, 0.26, 0.79, ...} | → | <b>0.7</b> |
| {0.9, 0.69, 0.16, ...} | → | 0.1        |
| {0.8, 0.82, 0.87, ...} | → | 0.1        |
| {0.5, 0.35, 0.10, ...} | → | 0.2        |
| {0.3, 0.72, 0.97, ...} | → | 0.4        |



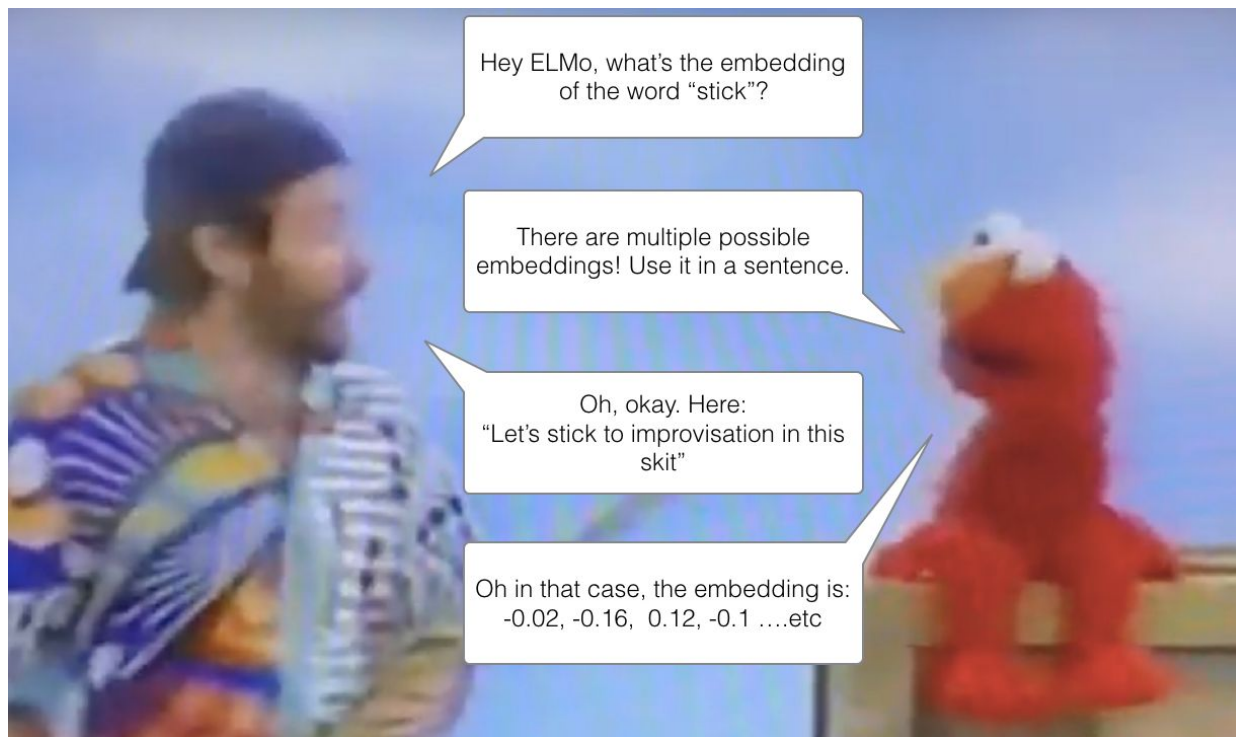


# From string of characters to numbers

## Word2Vec

- Sparsity ✓
- Semantics ✓
- Similarity ✓
- Polysemy ✗
- Language understanding ✗

# ELMo: context matters!



Source: [Illustrated BERT](#)

# ELMo: context matters!

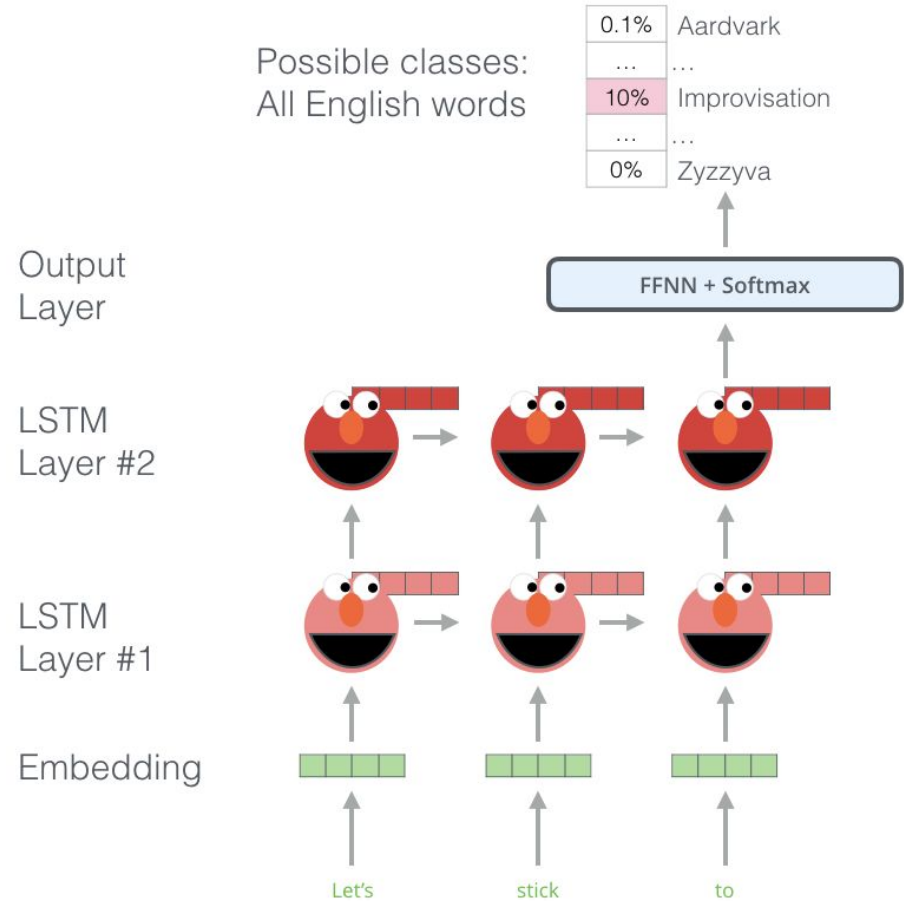
- Language modelling
  - What's the next word?

The cat

Source: [Illustrated BERT](#)

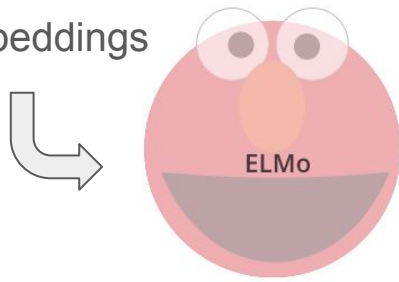
# ELMo: context matters!

- Language modelling
  - What's the next word?
- Massive dataset
- Reusable component



Source: [Illustrated BERT](#)

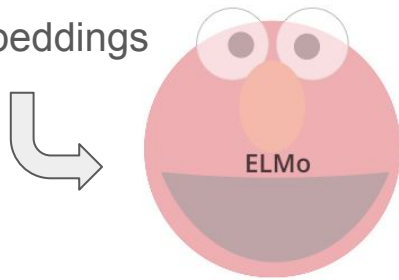
In-context embeddings



Source: [Illustrated BERT](#)



In-context embeddings



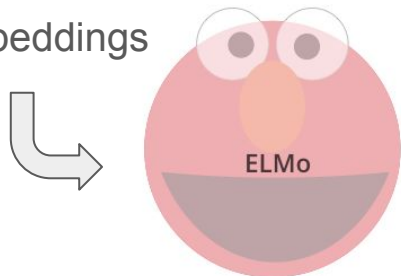
Attention mechanism

+



Source: [Illustrated BERT](#)

In-context embeddings



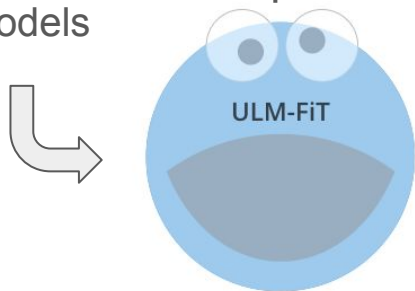
Attention is all you need

+



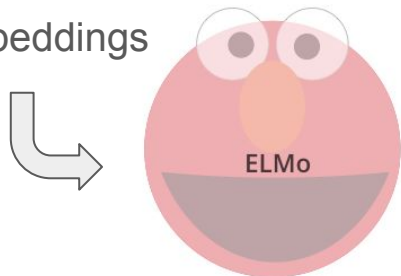
Pre-trained models

+



Source: [Illustrated BERT](#)

In-context embeddings

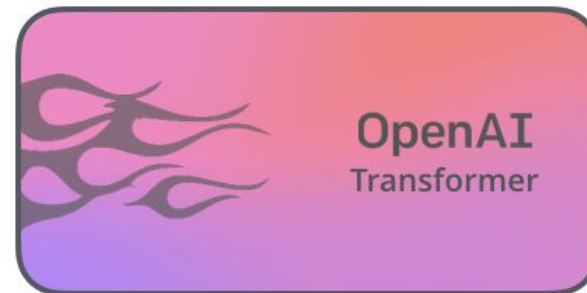
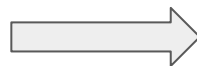


Attention is all you need

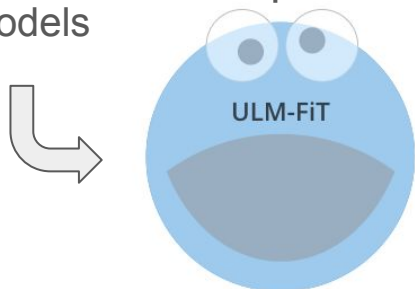


+

GPT-1

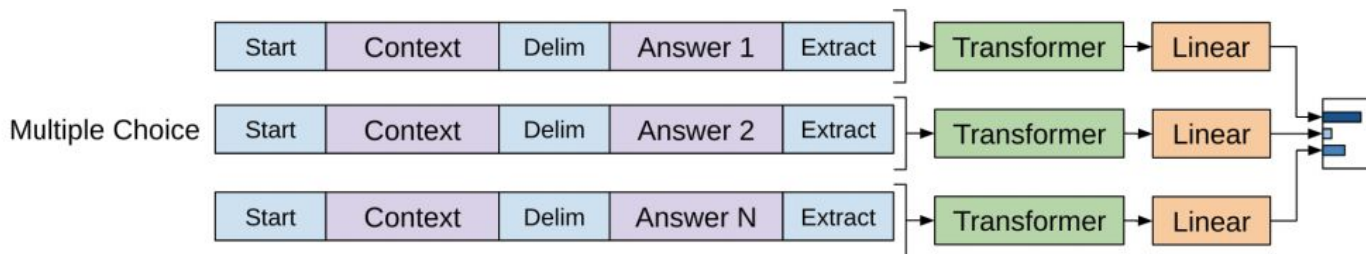
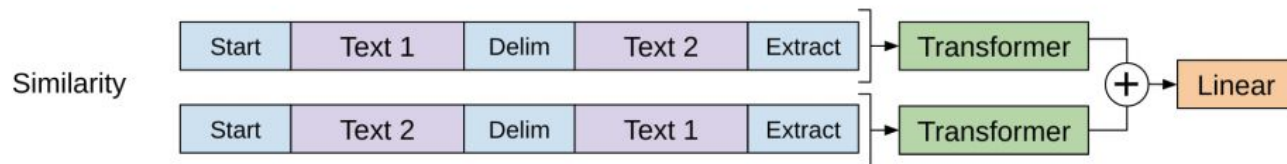
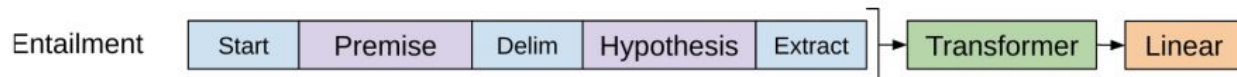
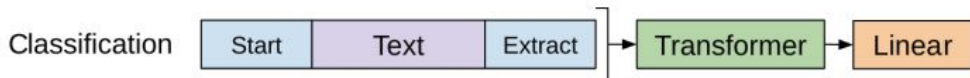


Pre-trained models



+

Source: [Illustrated BERT](#)



Source: [Illustrated BERT](#)





# BERT

Bidirectional Encoder Representations from Transformers



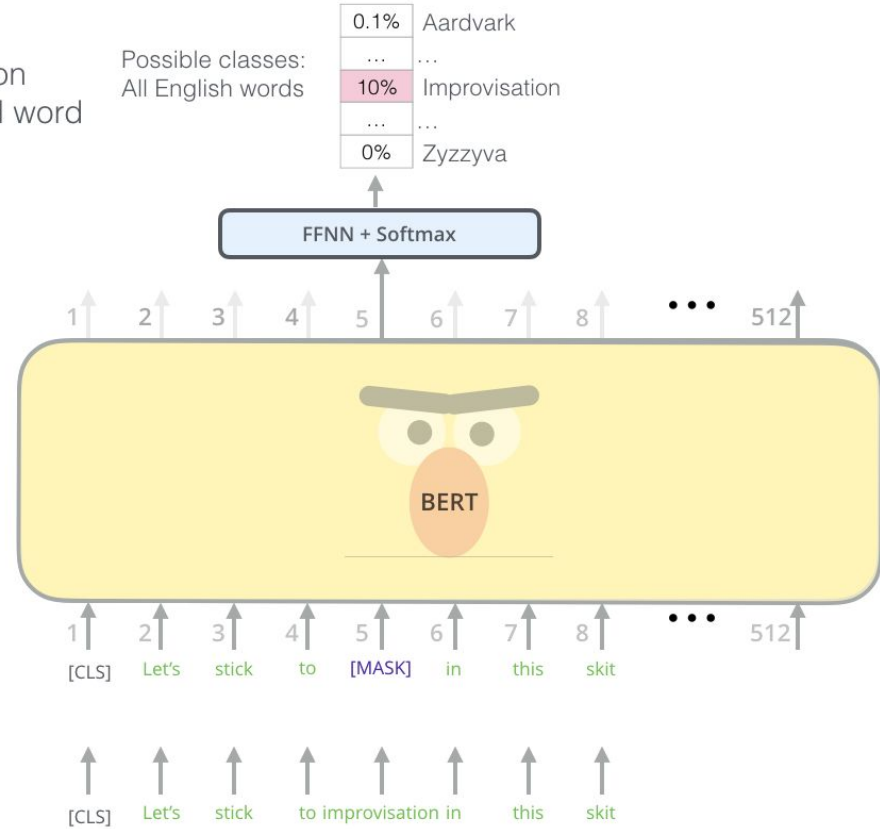
# BERT

## Masked Language Modeling

Use the output of the masked word's position to predict the masked word

Randomly mask 15% of tokens

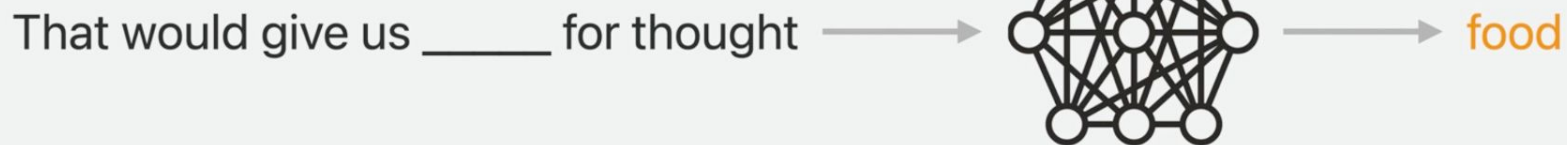
Input



Source: [Illustrated BERT](#)

# BERT

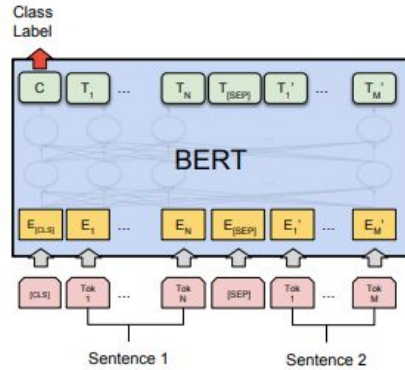
## Masked Language Modeling



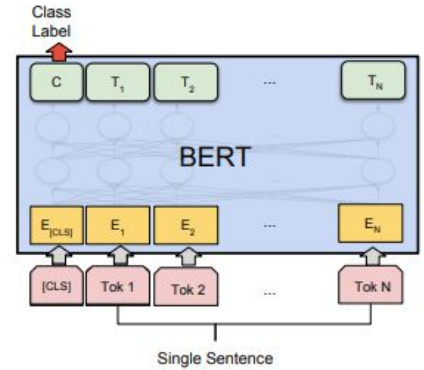
Source: [Illustrated BERT](#)

# BERT

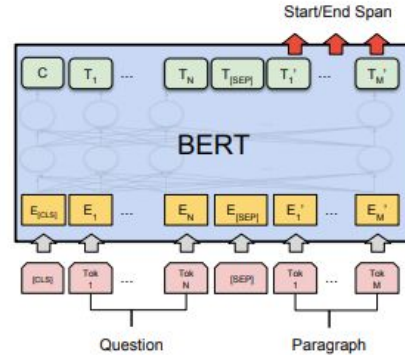
- Encoder only
- Fine-tuning
- Language understanding tasks



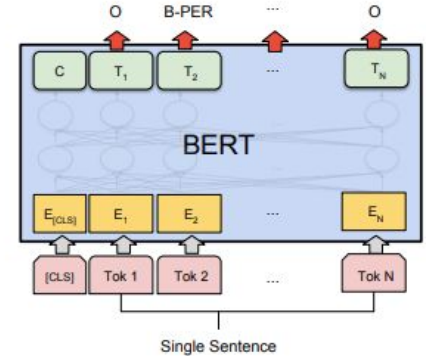
(a) Sentence Pair Classification Tasks:  
MNLI, QQP, QNLI, STS-B, MRPC,  
RTE, SWAG



(b) Single Sentence Classification Tasks:  
SST-2, CoLA



(c) Question Answering Tasks:  
SQuAD v1.1



(d) Single Sentence Tagging Tasks:  
CoNLL-2003 NER

Source: [Illustrated BERT](#)

# Original Pre-trained BERT Models

- bert-base-cased (English BERT)
- bert-base-uncased
- bert-large-uncased
- bert-base-multilingual-cased (mBERT)

# Original Pre-trained BERT Models

- bert-base-cased (English BERT)
- bert-base-uncased
- bert-large-uncased
- bert-base-multilingual-cased (mBERT)

|              | <b>Bokmål</b> | <b>Nynorsk</b> | <b>English</b> |
|--------------|---------------|----------------|----------------|
| English BERT | 75.1          | 77.8           | 91.3           |
| mBERT        | 83.8          | 85.6           | 90.8           |

|       | NER           |                | POS           |                | Sentiment                   | Political     |
|-------|---------------|----------------|---------------|----------------|-----------------------------|---------------|
|       | <b>Bokmål</b> | <b>Nynorsk</b> | <b>Bokmål</b> | <b>Nynorsk</b> | <b>Bokmål &amp; Nynorsk</b> | <b>Bokmål</b> |
| mBERT | 83.8          | 85.6           | 98.3          | 98.0           | 69.7                        | 78.4          |



# Original Pre-trained BERT Models

- bert-base-cased (English BERT)
- bert-base-uncased
- bert-large-uncased
- bert-base-multilingual-cased (mBERT)

|              | <b>Bokmål</b> | <b>Nynorsk</b> | <b>English</b> |
|--------------|---------------|----------------|----------------|
| English BERT | 75.1          | 77.8           | 91.3           |
| mBERT        | 83.8          | 85.6           | 90.8           |

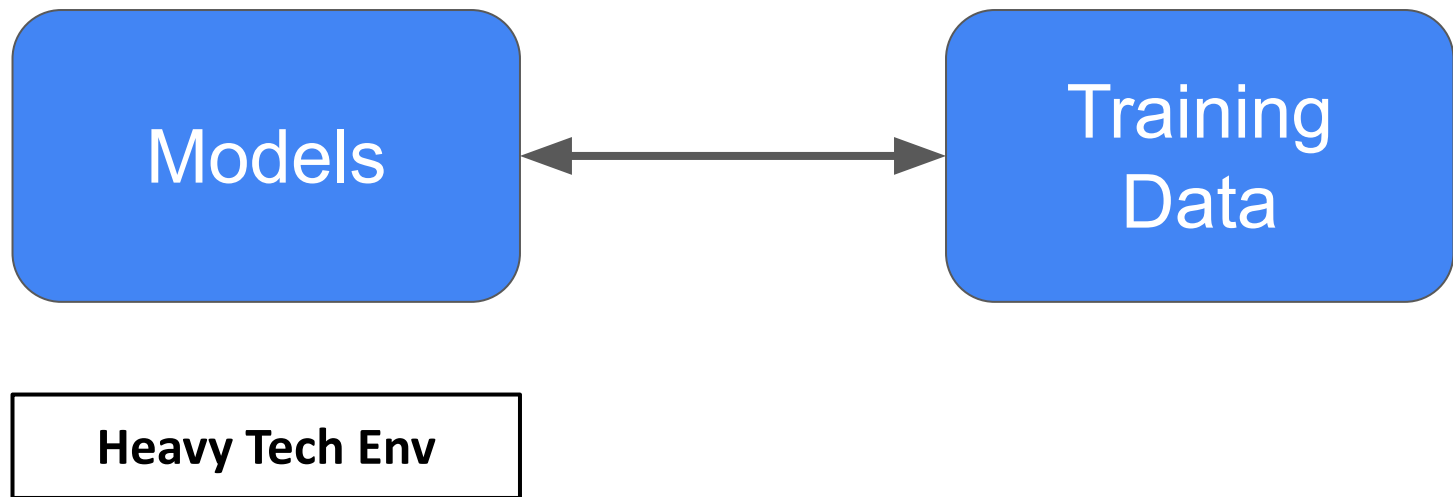
Could we build a BERT model for Norwegian as good as the original one was for English

|       | NER           |                | POS           |                | Sentiment                   | Political     |
|-------|---------------|----------------|---------------|----------------|-----------------------------|---------------|
|       | <b>Bokmål</b> | <b>Nynorsk</b> | <b>Bokmål</b> | <b>Nynorsk</b> | <b>Bokmål &amp; Nynorsk</b> | <b>Bokmål</b> |
| mBERT | 83.8          | 85.6           | 98.3          | 98.0           | 69.7                        | 78.4          |

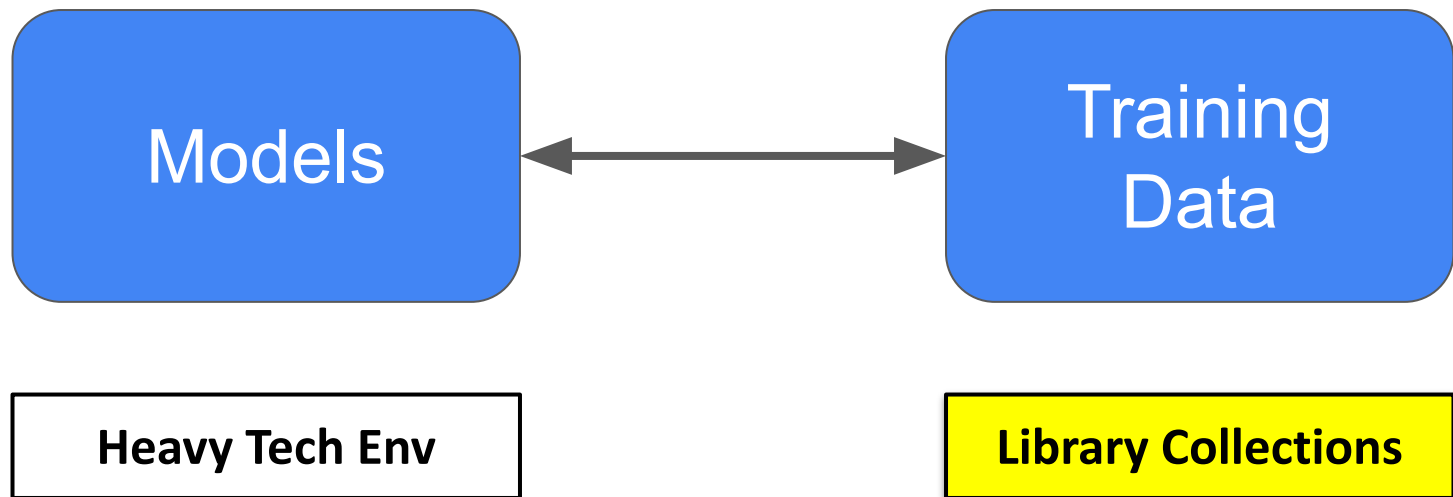
# Models: Software and Data



# Models: Software and Data



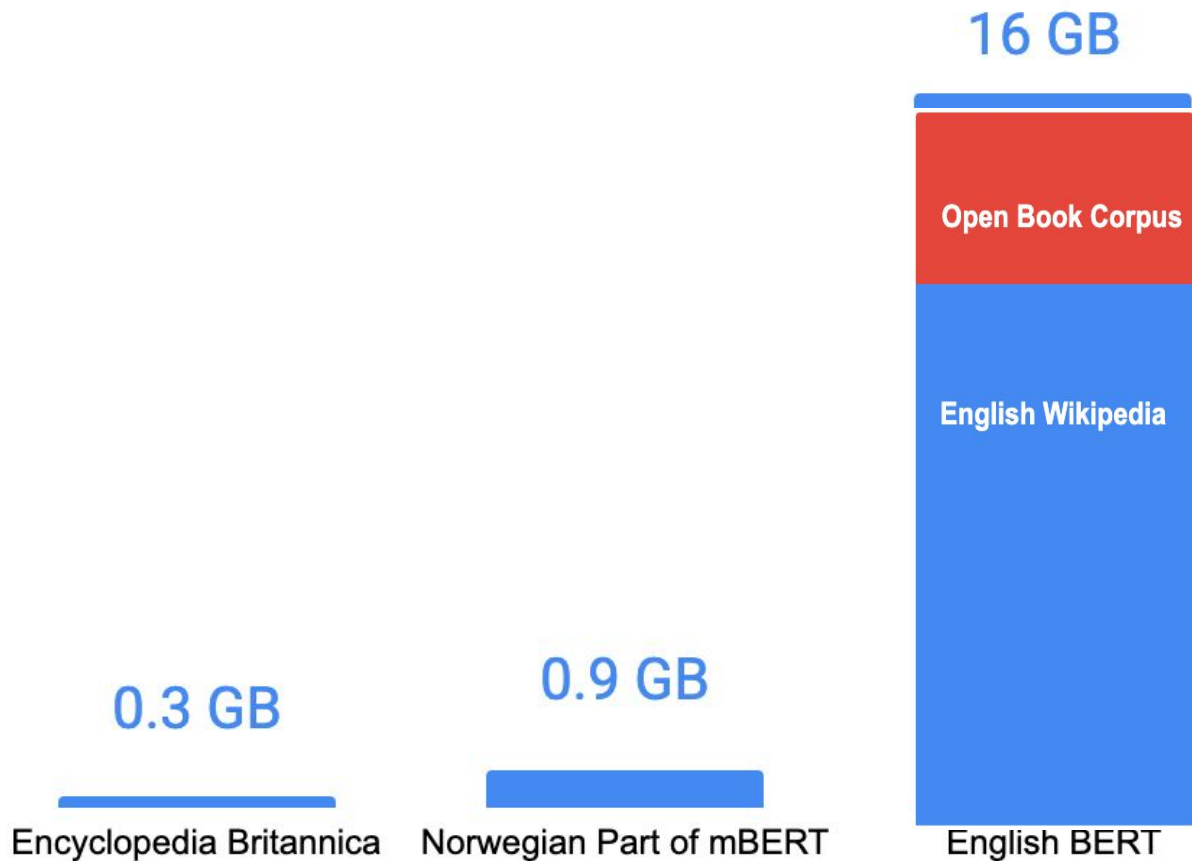
# Models: Software and Data



# NoTraM - Norwegian Transformer Models

1. Build a big corpus
2. Build the first Norwegian Transformer Model
  - NB-BERT (Base and Large)







# SIZE MATTERS

# QUALITY MATTERS MORE

# NB Collection

- Library collection:
  - Literature (OCR, PDF)
  - Newspapers (OCR, PDF)
  - Articles (OCR, HTML)
  - Public publications (OCR, PDF, XML)
  - Lovdata

# Web and other agreements

- CC100
- MC4
- Open Subtitles
- OSCAR
- Reddit
- Wikipedia
- Twitter
- VG Debatt
- Facebook

# Comparable Corpora

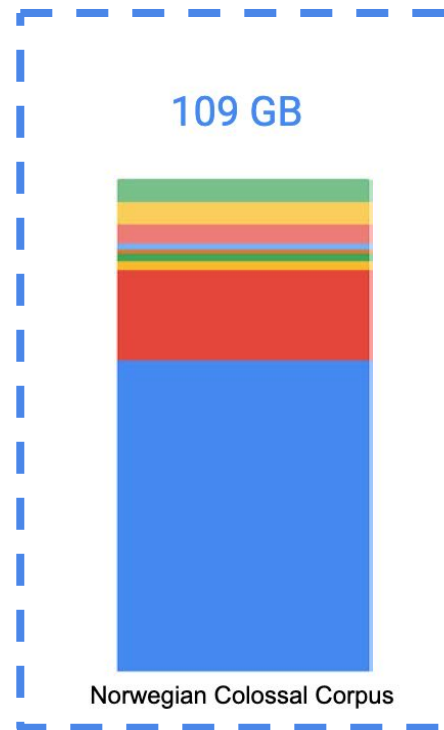


# Comparable Corpora

0.3 GB  
Encyclopedia Britannica

0.9 GB  
Norwegian Part of mBERT

16 GB  
English BERT



Norwegian Colossal Corpus

160 GB  
Open WeText  
Common Crawl Stories  
Common Crawl News  
Facebook RoBERTa



| Corpus                 | License                      | Size    | Words         | Documents  | Avg words per doc |
|------------------------|------------------------------|---------|---------------|------------|-------------------|
| Library Newspapers     | <a href="#">CC0 1.0</a>      | 14.0 GB | 2,019,172,625 | 10,096,424 | 199               |
| Library Books          | <a href="#">CC0 1.0</a>      | 6.2 GB  | 861,465,907   | 24,253     | 35,519            |
| LovData CD             | <a href="#">NLOD 2.0</a>     | 0.4 GB  | 54,923,432    | 51,920     | 1,057             |
| Government Reports     | <a href="#">NLOD 2.0</a>     | 1.1 GB  | 155,318,754   | 4,648      | 33,416            |
| Parliament Collections | <a href="#">NLOD 2.0</a>     | 8.0 GB  | 1,301,766,124 | 9,528      | 136,625           |
| Public Reports         | <a href="#">NLOD 2.0</a>     | 0.5 GB  | 80,064,396    | 3,365      | 23,793            |
| Målfrid Collection     | <a href="#">NLOD 2.0</a>     | 14.0 GB | 1,905,481,776 | 6,735,367  | 282               |
| Newspapers Online      | <a href="#">CC BY-NC 2.0</a> | 3.7 GB  | 541,481,947   | 3,695,943  | 146               |
| Wikipedia              | <a href="#">CC BY-SA 3.0</a> | 1.0 GB  | 140,992,663   | 681,973    | 206               |

# Norwegian Colossal Corpus (NCC)

Newspapers  
(2.56B words)

Målfrid  
(1.91B words)

Parliament  
(1.30B words)

Books  
(877.01M words)

Wikipedia  
(462.52M words)

Other  
(353.97M words)

# Original Pre-trained BERT Models

- bert-base-cased (English BERT)
- bert-base-uncased
- bert-large-uncased
- bert-base-multilingual-cased (mBERT)

|              | <b>Bokmål</b> | <b>Nynorsk</b> | <b>English</b> |
|--------------|---------------|----------------|----------------|
| English BERT | 75.1          | 77.8           | 91.3           |
| mBERT        | 83.8          | 85.6           | 90.8           |

|       | <b>NER</b>    |                | <b>POS</b>    |                | <b>Sentiment</b>            | <b>Political</b> |
|-------|---------------|----------------|---------------|----------------|-----------------------------|------------------|
|       | <b>Bokmål</b> | <b>Nynorsk</b> | <b>Bokmål</b> | <b>Nynorsk</b> | <b>Bokmål &amp; Nynorsk</b> | <b>Bokmål</b>    |
| mBERT | 83.8          | 85.6           | 98.3          | 98.0           | 69.7                        | 78.4             |

# NB-BERT

- bert-base-cased (English BERT)
- bert-base-uncased
- bert-large-uncased
- bert-base-multilingual-cased (mBERT)

|                | Bokmål      | Nynorsk     | English     | Swedish     | Danish      | Spanish     | Finnish     |
|----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| English BERT   | 75.1        | 77.8        | 91.3        | 82.5        | 73.9        | 81.8        | 82.9        |
| mBERT          | 83.8        | 85.6        | 90.8        | 85.3        | 83.4        | <b>87.6</b> | <b>88.7</b> |
| NB-BERT (ours) | <b>91.2</b> | <b>88.9</b> | <b>91.3</b> | <b>85.9</b> | <b>85.1</b> | 85.8        | 85.8        |

|                | NER         |             | POS         |             | Sentiment        | Political   |
|----------------|-------------|-------------|-------------|-------------|------------------|-------------|
|                | Bokmål      | Nynorsk     | Bokmål      | Nynorsk     | Bokmål & Nynorsk | Bokmål      |
| mBERT          | 83.8        | 85.6        | 98.3        | 98.0        | 69.7             | 78.4        |
| NB-BERT (ours) | <b>91.2</b> | <b>88.9</b> | <b>98.8</b> | <b>98.8</b> | <b>86.4</b>      | <b>81.8</b> |

# NB-BERT

— When Norwegians are Better than Danes at Danish: The State and Shortcomings of the Danish NLP, and How We Fix Them - Kenneth Enevoldsen, PhD Student, Aarhus University

Within the last decade language models like GPT3 or BERT have become the standard in Natural Language Processing (NLP) across a wide variety of tasks from translation to hate speech detection. Even low-resource languages like Danish have their own language model, but Norwegian models perform better than Danish model!? This talk will walk you through the current state and shortcoming of the Danish NLP and how we plan to improve them through a nationwide cross-sector collaboration.



**Kenneth Enevoldsen**

PhD Student

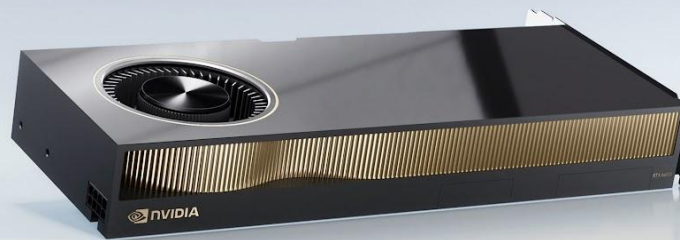
School of Culture and Society

Aarhus University

# Our models

## NB-BERT (encoder only)

- NB-BERT-base-ner, Named Entity Recognition
- NB-BERT-base-pos, Part of Speech
- NB-SBERT, Sentence Embeddings
- NB-BERT-mnli, Zero-shot Textual Entailment



- Multiple other models trained externally based on NCC, for instance T5-models

*Per i dag har Per og Dag startet firmaet Per og Dag*

P P O



# Text Analysis

- Powerful Norwegian language model
- Fine tuning for different purposes
- Analyzes text to find
  - Personal names
  - Different place names (city, country...)
  - Organization names

## NoTram Named Entity Recognition

Text

Israels statsminister Benjamin Netanyahu forsøker å framstille de siste ukenes bombing av G

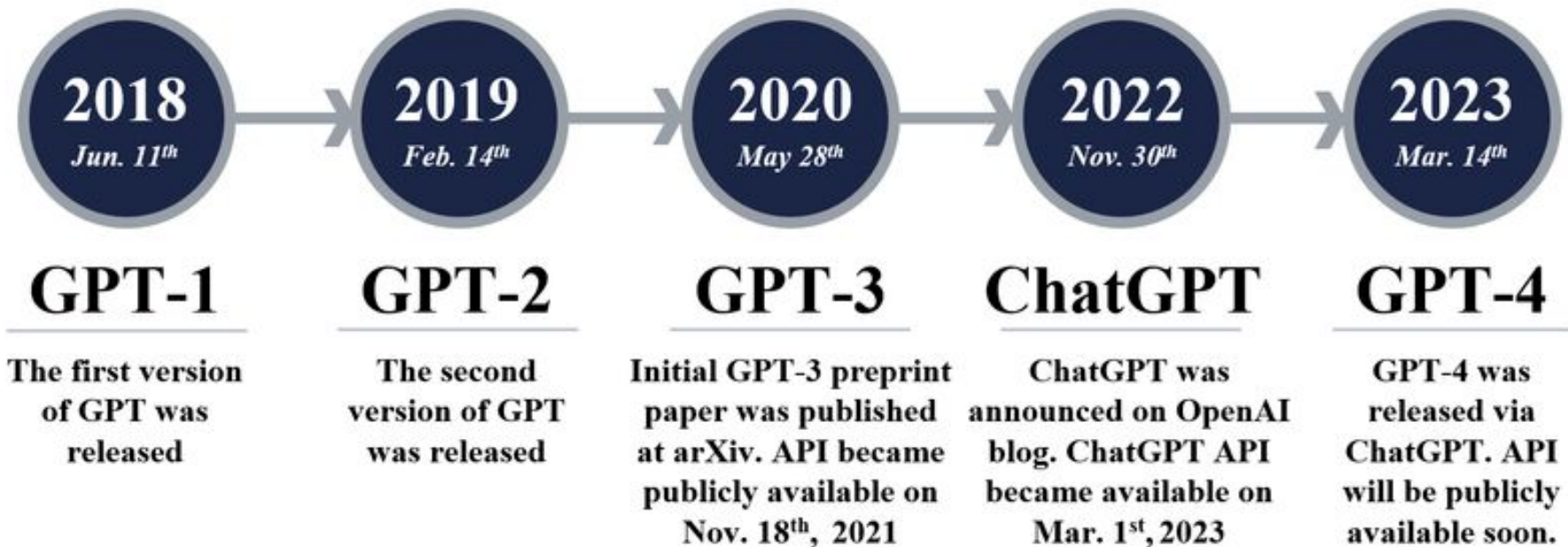
Israel GPE\_ORG ##s statsminister Benjamin Netanyahu PER forsøker å framstille de siste ukenes bombing av Gazastripa LOC som en seier, etter at en våpenhvile trådte i kraft natt til fredag. Net PER ##anya PER ##hu PER kaller bombingene « en eksepsjonell suksess ». Bombene drepte 243 palestinere, blant dem 66 barn, ifølge palestinske helsemyndigheter. FN ORG ##s hjelpeorganisasjon for palestinske flyktninger forteller at elever på deres skoler i Gaza GPE\_LOC har blitt drept på skoleveien, uten nærhet til noen åpenbare militære mål. Det blir likevel ikke noen rettergang om bombene mot skolebarn eller mediebygningen i Gaza GPE\_LOC, som blant annet huser nyhetsbyrået AP ORG og Al Jazeera ORG. Israel GPE\_ORG legger ikke fram bevis for at målene er militære, og ingen krever at de gjør det heller. Fordi Israel GPE\_ORG aldri blir stilt til ansvar i et internasjonalt rettssystem, vil uretten fortsette uten at landet legger noen bånd på egen framferd. Utskiftingen av den palestinske befolkningen med jødiske bosettere i Jerusalem GPE\_LOC og på Vestbredden LOC vil fortsette. Våpenhvile vil forbli midlertidige, med jevnlig oppbluss av motstand og bomber når okkupasjonens urett vekker folk til nye protester. Ifølge Israel GPE\_ORG handler det om en krig med islamistene i Hamas ORG, men den grunnleggende konflikten er den samme i dag som da Yassir Arafat PER ##s sekulære PL ORG ##O ORG var motparten. De som kjemper mot Israel GPE\_ORG ##s ulovlige okkupasjon, vil alltid være terrorister og legitime mål i Israel GPE\_ORG ##s øyne.

# What kind of tasks can BERT tackle?

## Classification

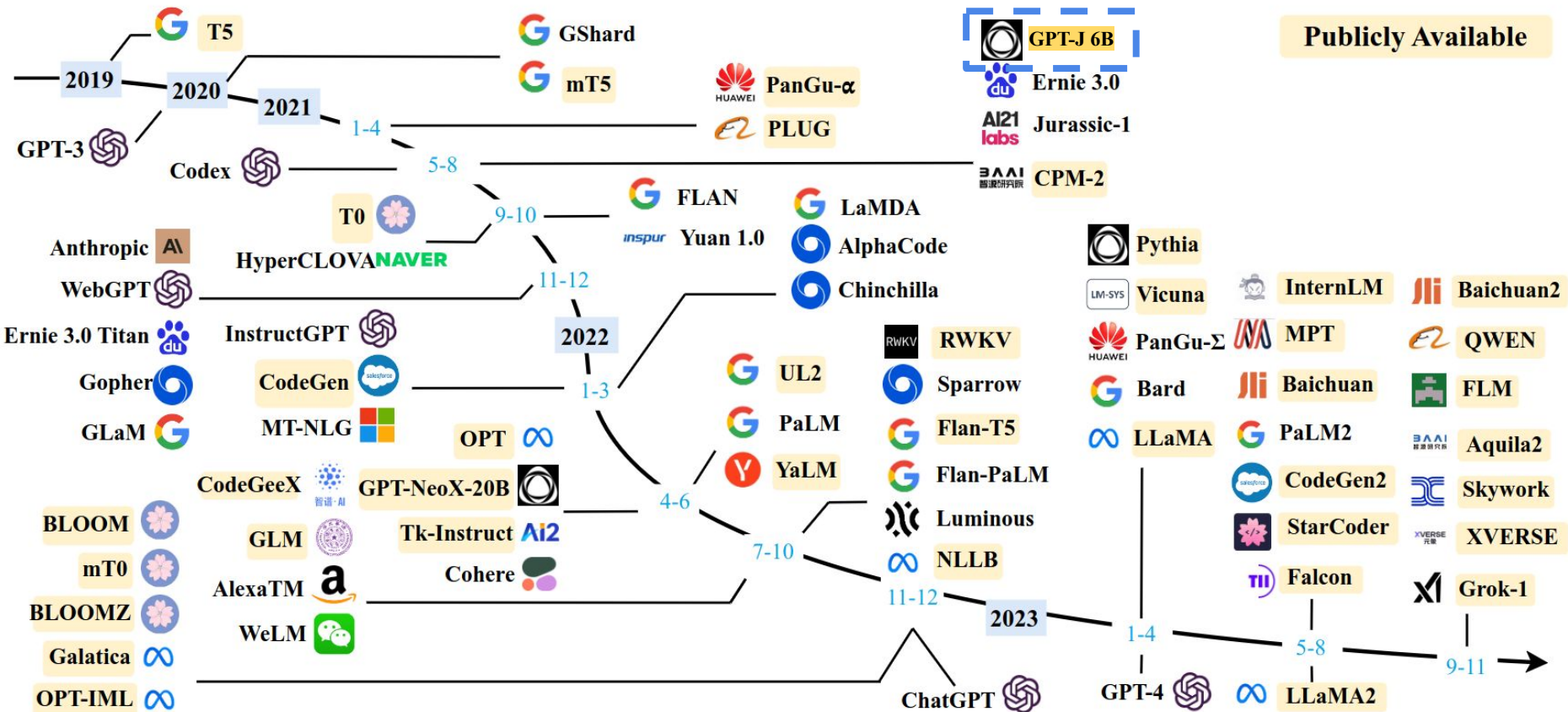
- At the word level (token)
- At the sentence level (sequence)

## *What about generation?*

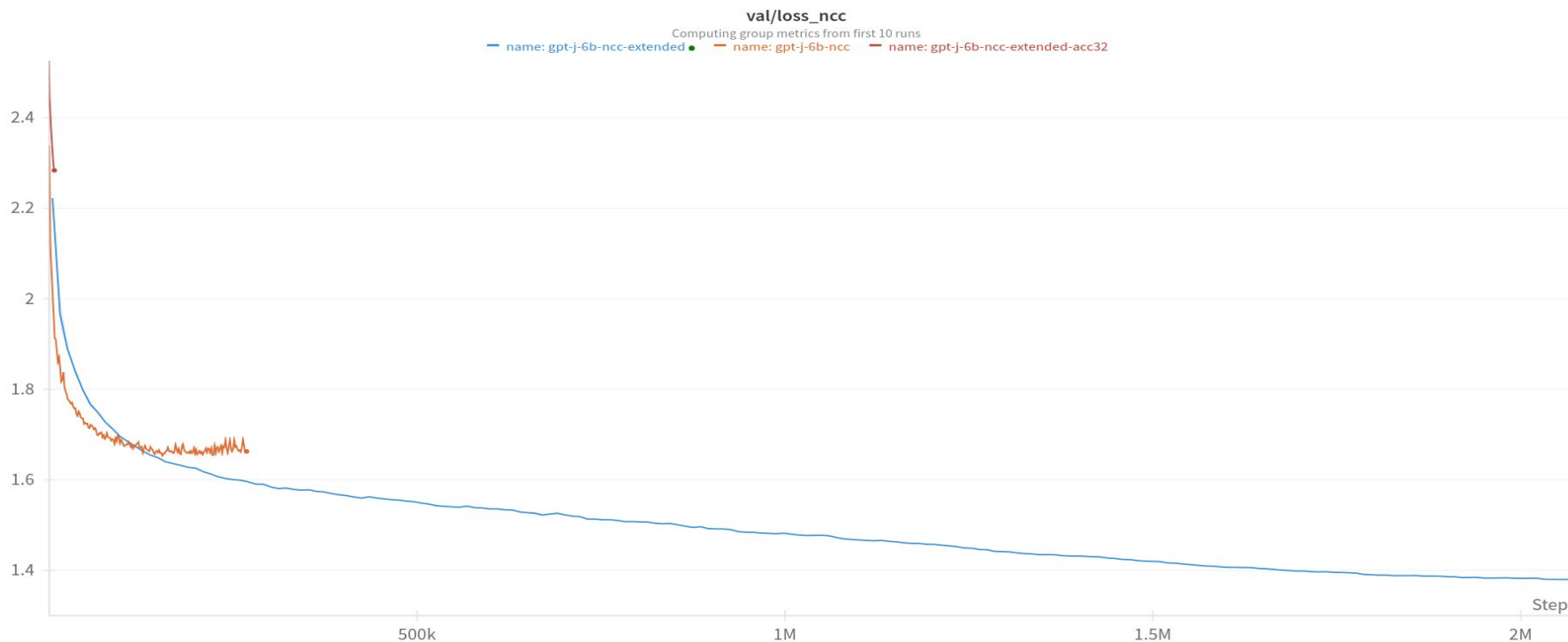


Source: [Evolution of GPT Models](#)

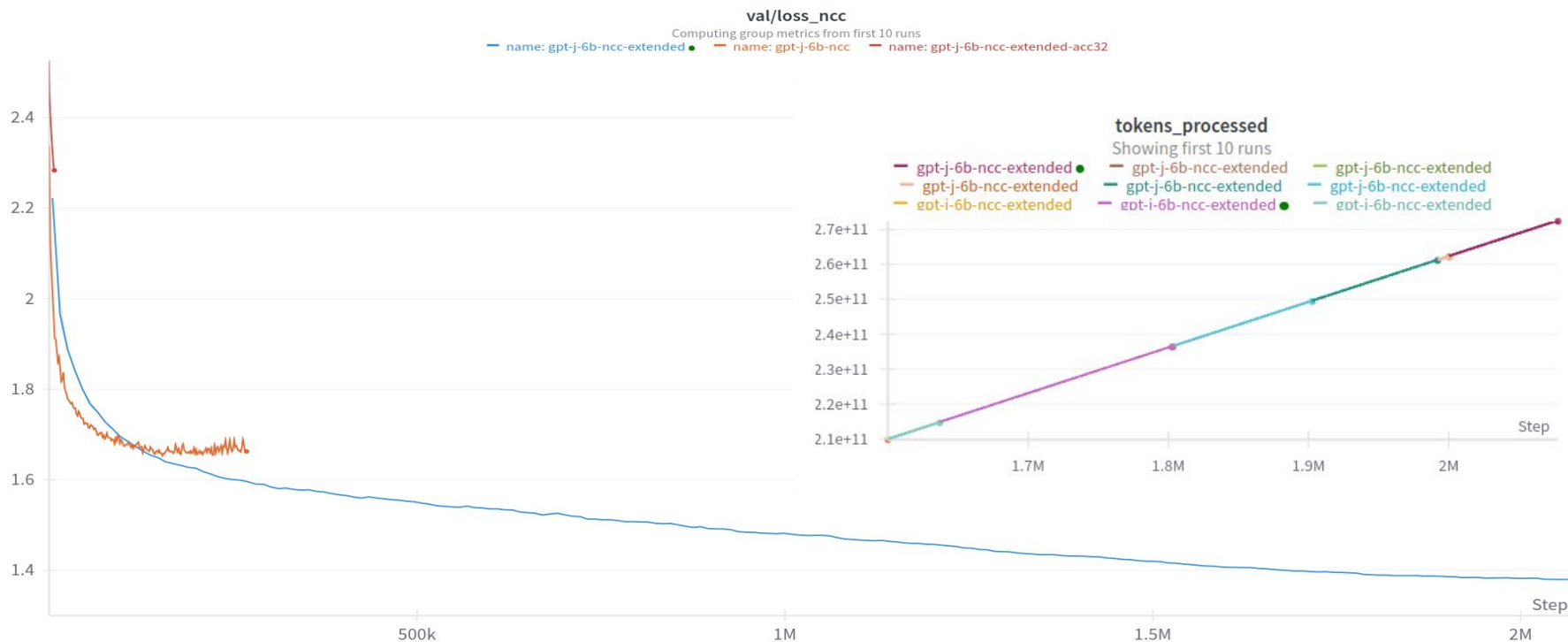
# NB-GPT-J 6B



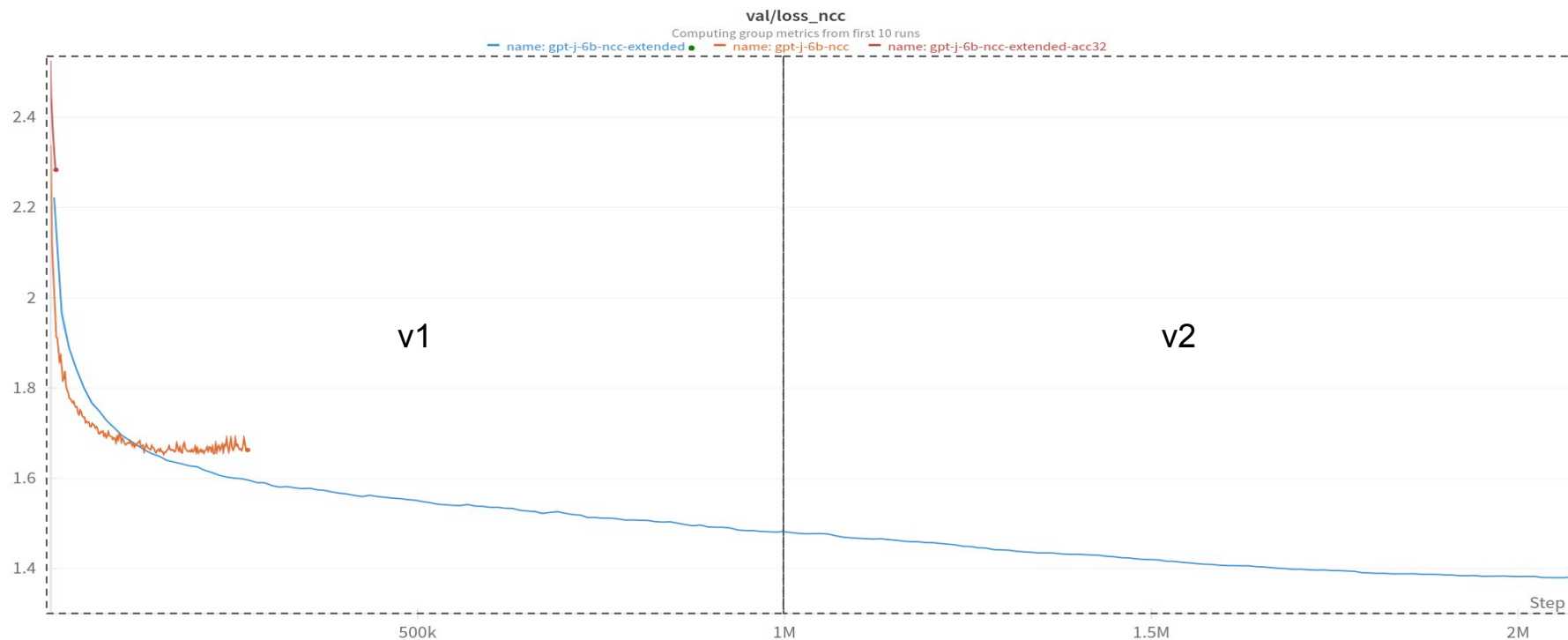
# NB-GPT-J 6B



# NB-GPT-J 6B

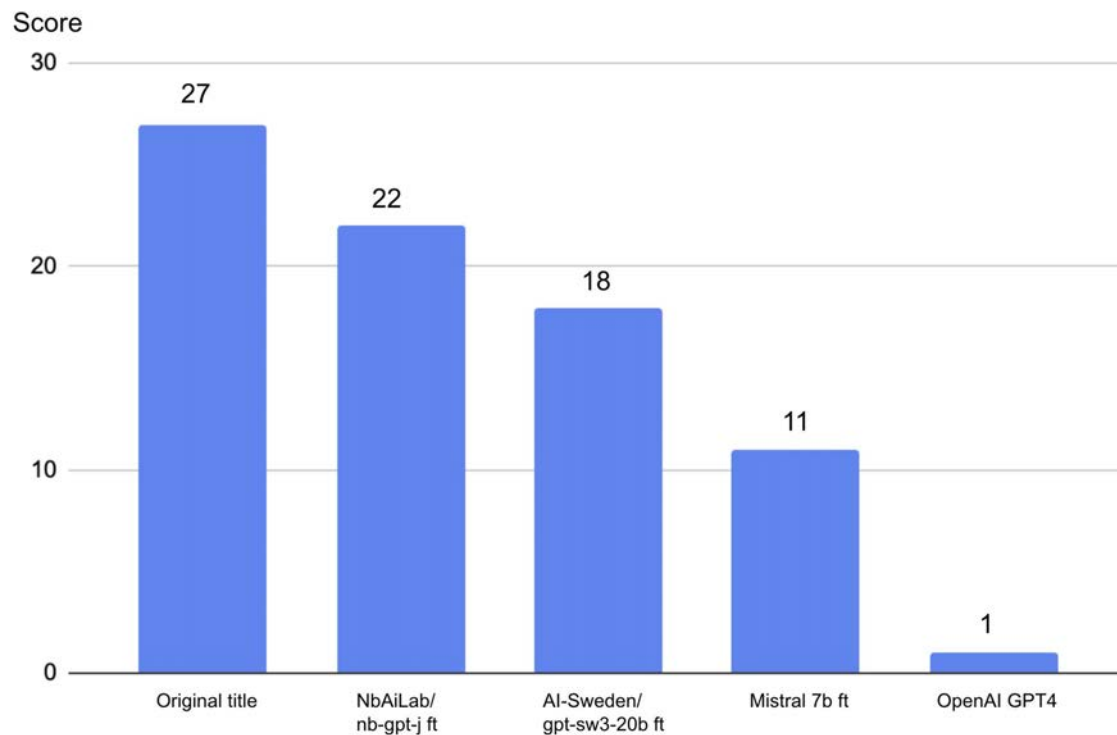


# NB-GPT-J 6B





# NB-GPT-J 6B







NbAiLab/nb-gpt-j-6B-alpaca

like 1

Text Generation Transformers PyTorch Safetensors NbAiLab/norwegian-alpaca Norwegian Norwegian Bokmål gptj alpaca License: openrail

Model card Files and versions Community 2 Settings

Train Deploy Use in Transformers

Edit model card

Gated model You have been granted access to this model

## NB GPT-J-6B Alpaca

This is a [NB GPT-J-6B Norwegian Bokmål](#) model fine-tuned on the [Norwegian Alpaca](#) dataset.

### Usage

```
from transformers import AutoModelForCausalLM, AutoTokenizer, GenerationConfig

base_model = "NbAiLab/nb-gpt-j-6B-alpaca"
tokenizer = AutoTokenizer.from_pretrained(base_model)
model = AutoModelForCausalLM.from_pretrained(base_model).cuda()
```

For generation, we can either use `pipeline()` or the model's `.generate()` method. Remember that the prompt needs a **Norwegian** template:

```
# Generate responses
def generate(instruction, input=None):
    if input:
        prompt = f"""Nedenfor er en instruksjon som beskriver en oppgave, sam

### Instruksjon:
{instruction}
```

Downloads last month

1,356



Safetensors Model size 6.06B params Tensor type F32 - BOOL

### Hosted inference API

Text Generation

Examples

Nedenfor er en instruksjon som beskriver en oppgave. Skriv et svar som fullfører forespørselen på riktig måte.

### Instruksjon: Skriv en e-post der du ønsker velkommen til en ny medarbeider ved navn Svein.

### Respons:

Compute

ctrl+Enter

0,5

This model can be loaded on the Inference API on-demand.

JSON Output

Maximize

### Dataset used to train NbAiLab/nb-gpt-j-6B-alpaca

NbAiLab/norwegian-alpaca

Viewer · Updated Jul 25 · 14 · 4

Spaces using NbAiLab/nb-gpt-j-6B-alpaca 10



NbAiLab / nb-gpt-j-6B-alpaca

♡ like 1

Text Generation

Transformers

PyTorch

Safetensors

NbAiLab/norwegian-alpaca

Norwegian

Norwegian Bokmål

gptj

alpaca

License: openrail

Model card

Files and versions

Community 2

Settings

⋮

Train

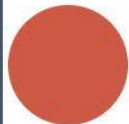
Deploy

Use in Transformers

Edit model card

Gated model You have been granted access to this model

Downloads last month

**You**

Kan du skrive et avsnitt som starter med setningen "Hun liker å sitte på en stein, med god utsikt over dalen."?

**Assistant**

Selvsagt. Hun liker å sitte på en stein, med god utsikt over dalen. Solstrålene filtrerer gjennom trærne og kaster skygger på bakken under henne. Fugler synger i greinene over henne, og vinden bringer en mild bris.

that the prompt needs a **Norwegian** template:

```
# Generate responses
def generate(instruction, input=None):
    if input:
        prompt = f"""Nedenfor er en instruksjon som beskriver en oppgave, samr

### Instruksjon:
{instruction}
```

JSON Output

Maximize

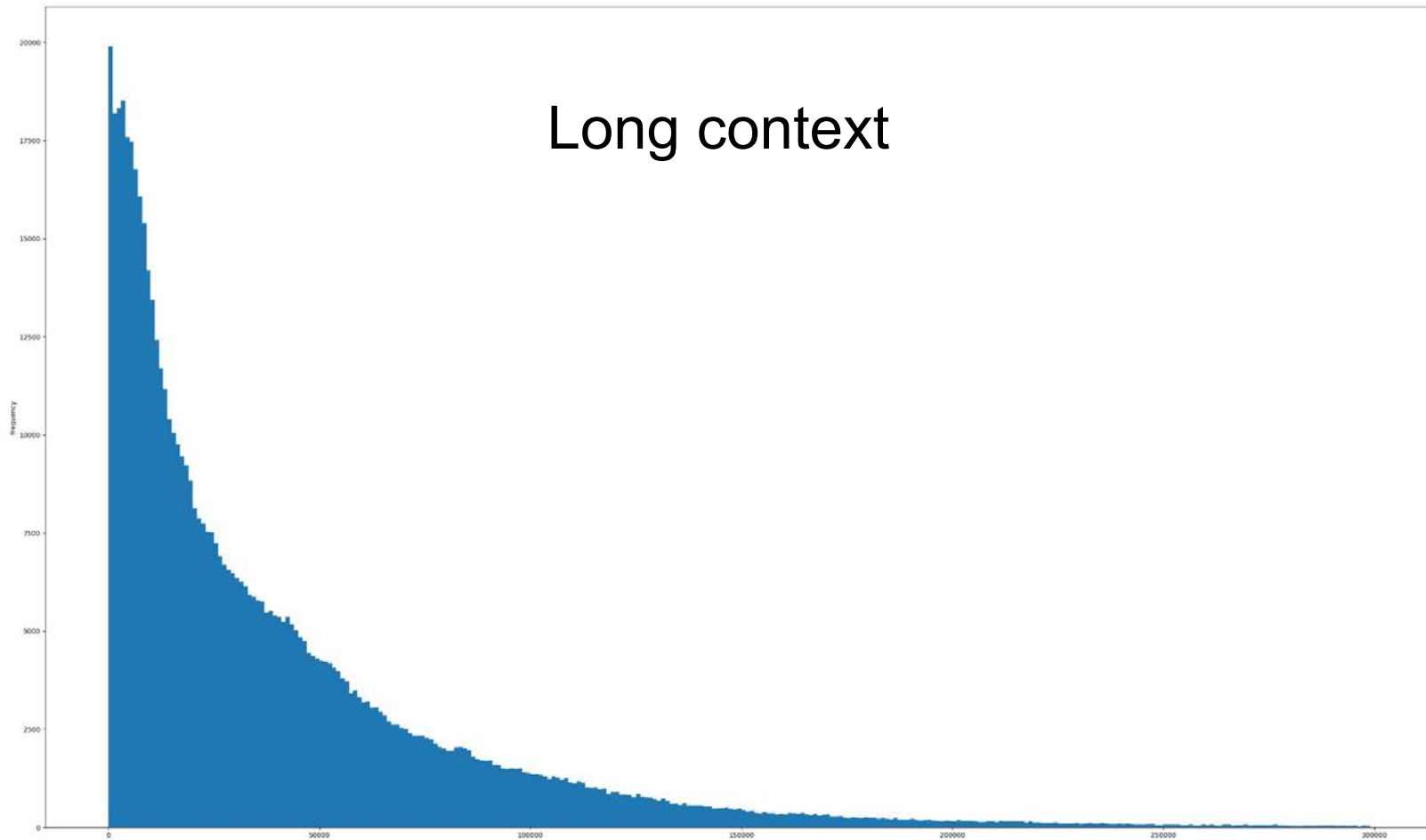
Dataset used to train NbAiLab/nb-gpt-j-6B-alpaca

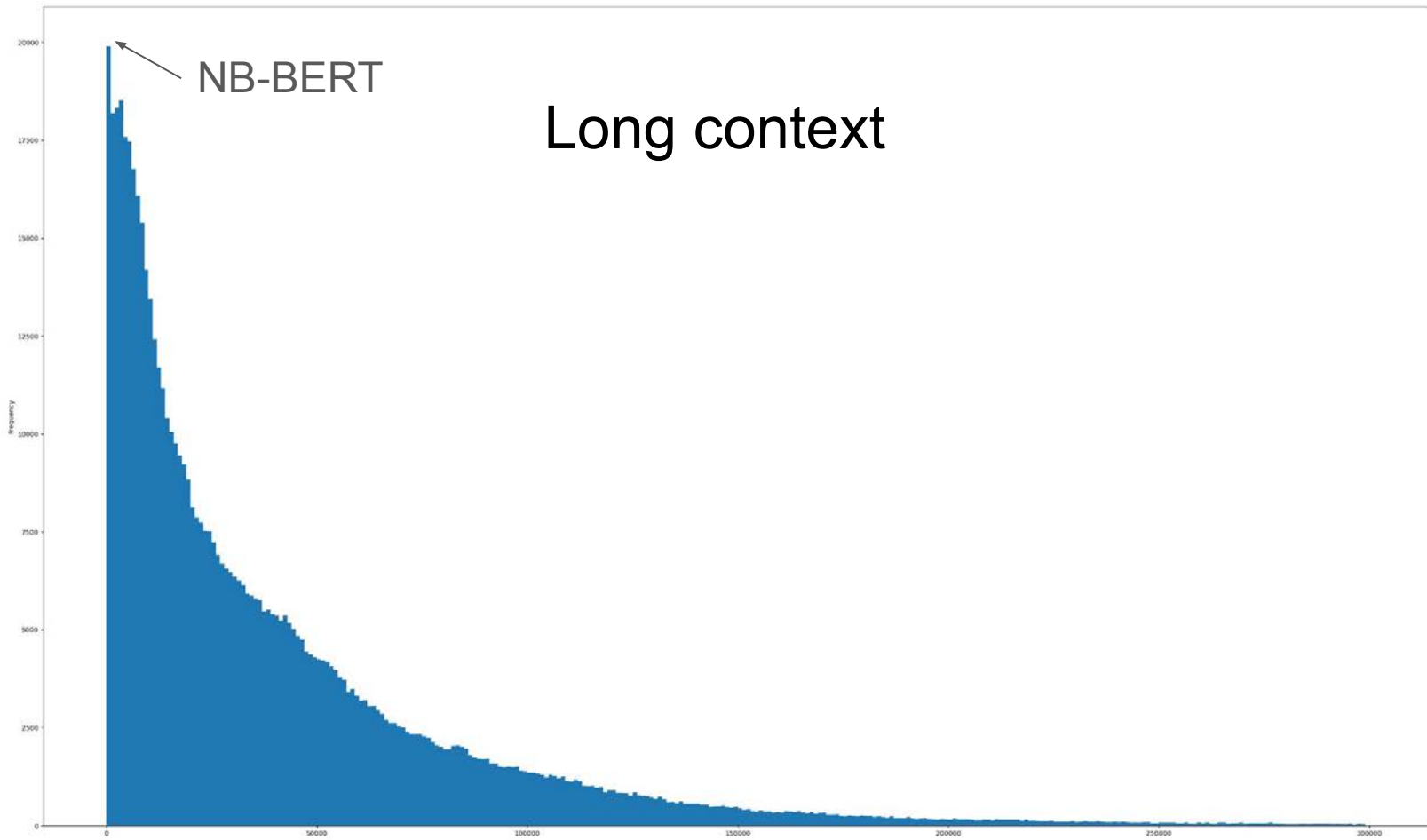
NbAiLab/norwegian-alpaca

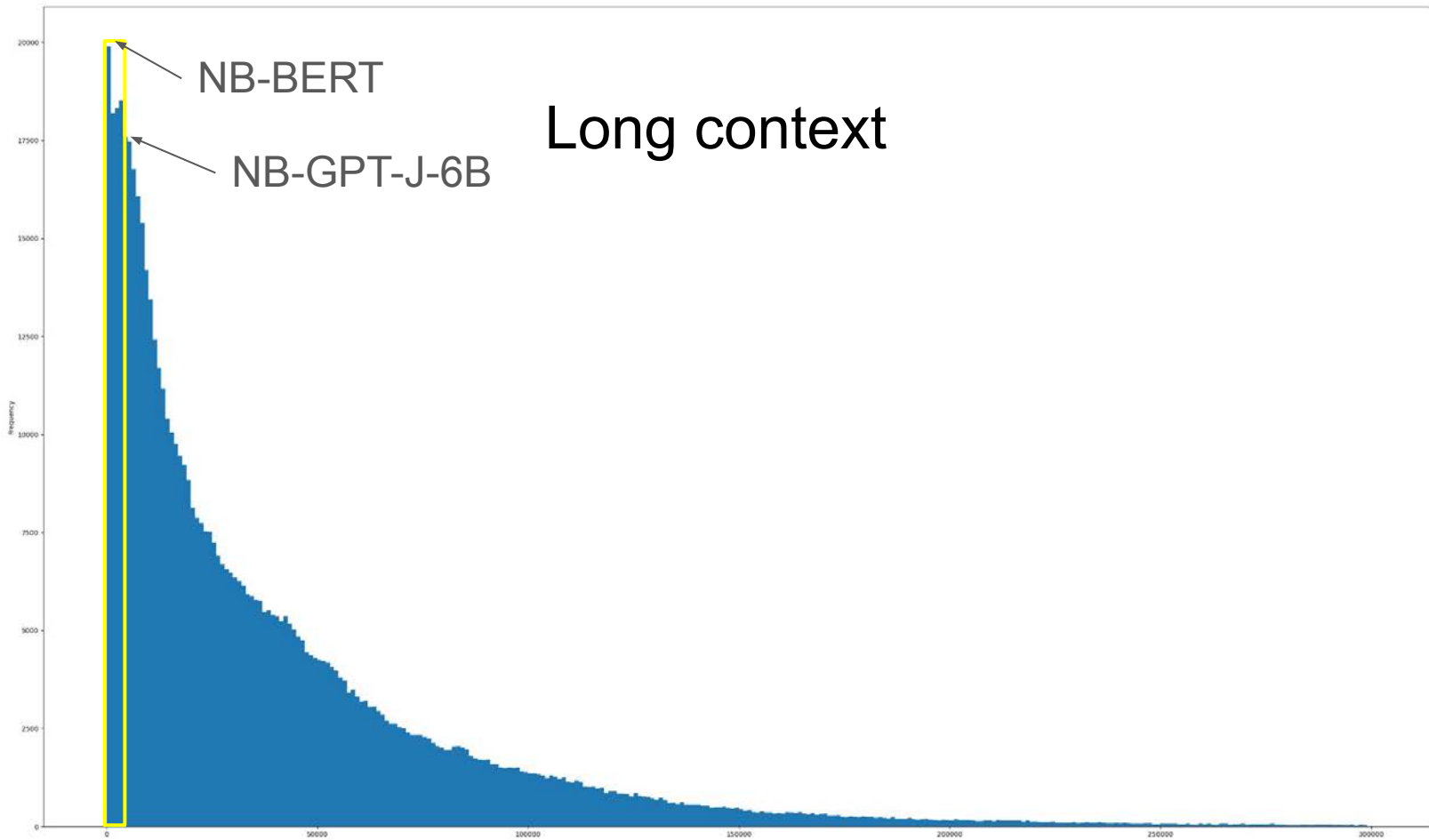
Viewer · Updated Jul 25 · ↓14 · ♥4

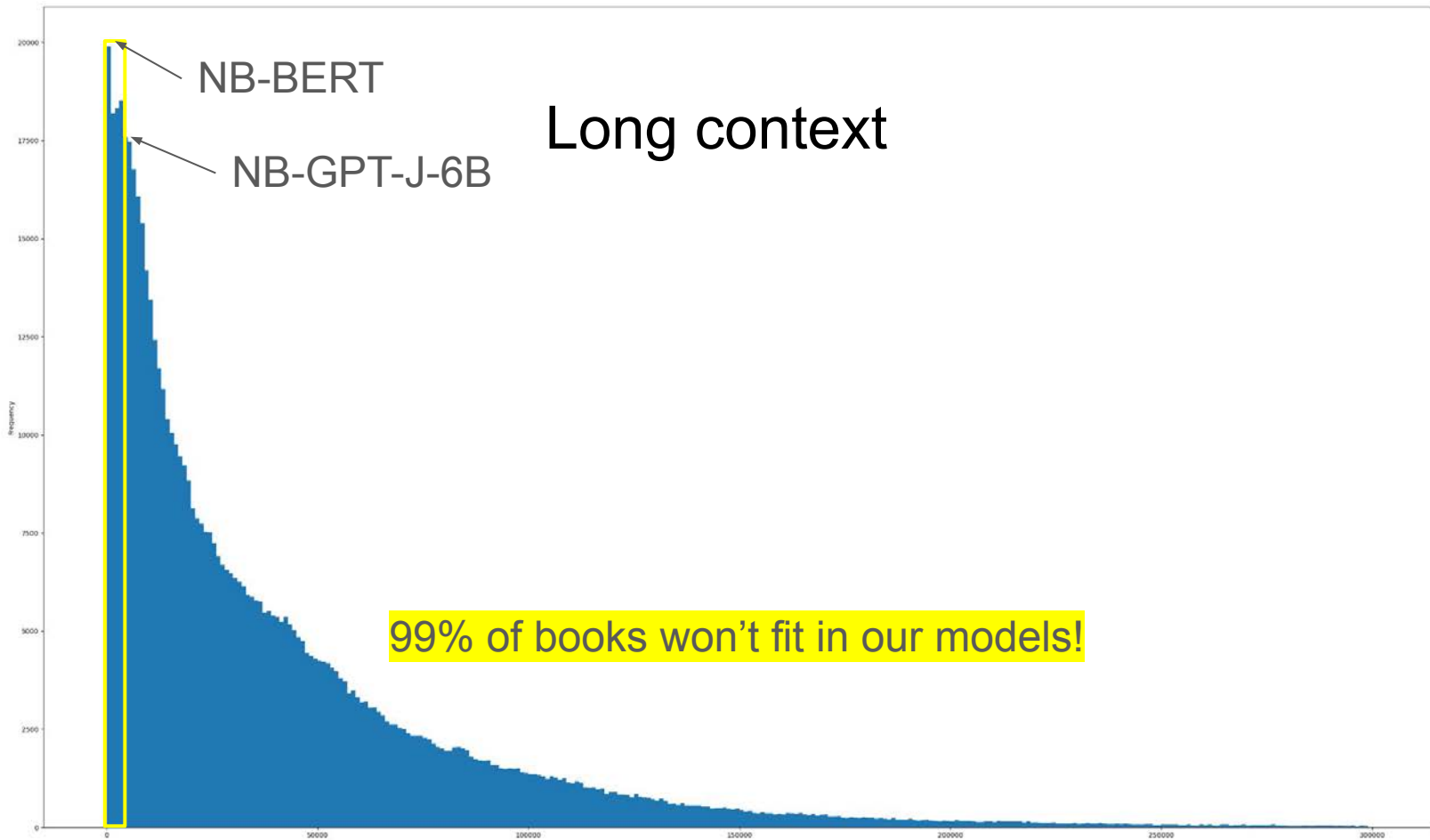
Spaces using NbAiLab/nb-gpt-j-6B-alpaca 10

# Long context









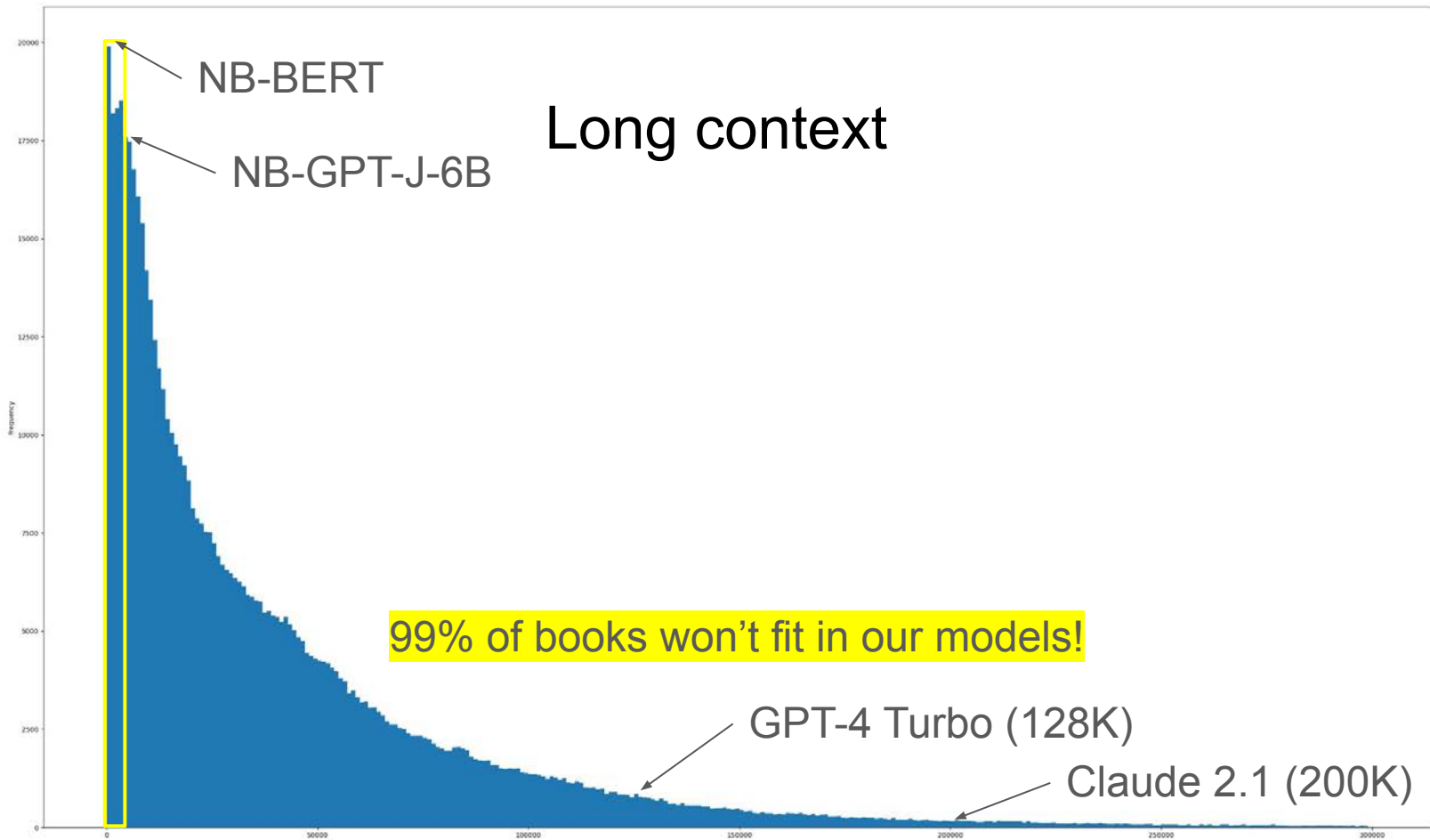
Long context

NB-BERT

NB-GPT-J-6B

99% of books won't fit in our models!





# Other Modalities

# Speech to Text

Can we create good open and free models for speech to text in Norwegian?  
Bokmål, Nynorsk? Dialects?

# Speech to Text

## Thora Meyer 1939

Fra patrisierhjem til turisthotell. Samtale med fru Thora Meyer, Meyergården, Mo i Rana. 17.08.1939

Fra patrisierhjem til turisthotell. Samtale med fru Thora Meyer, Meyergården, Mo i Rana. 00:13:01

**Tilgang**  
Tilgang for alle

**Tittel**  
Fra patrisierhjem til turisthotell. Samtale med fru Thora Meyer, Meyergården, Mo i Rana.

**Dato**  
17.08.1939

**Programleder**  
Jensson, Leif

**Publisert**  
NRK (the Norwegian Broadcasting Corporation)

**Andre opplysninger**  
<br>K948.441 - RANA Mo (eo) \<br><br>Fra patrisierhjem til turisthotell. Samtale med fru Thora MEYER (mv),<br>Meyergården, Mo i Ra na.<br><br>Thora MEYER (mv), Mo i Rana: Om den første Meyergården, som brant ned i 1883.<br>Det nye bygget ble så stort at det ga plass til svigermor, svigerfar og 15<br>barn, forrettningens personale, private gjester, forretningsforbindelser og<br>mange tjenere. I tillegg kom det et nytt hus på den gamle branntomten. To<br>damer drev dette som hotell, mens min mann og jeg hadde hele gården her for<br>oss selv. Herskapet og gjestene øverst ved bordet fikk sin egen kost <br>mens barna og personalet nederst fikk en annen kost. Gjest gjev<br>plikten<br>medførte en utstrakt gjestfrihet. Gjester kunne bli boende her i lange tider <br>uten å betale. Forretningsforbindelser fra Sverige bodde alle gratis i<br>"Kavaljerfløyen", andre svensker i "Svenskehotellet". Svigermor hadde 15 - 16<br>piker (hushjelpere). Hadde selv aldri under 8 piker.<br>Etter min mands død hadde jeg et par års opphold i tropene og i Amerika.<br>Kom tilbake til et stort, tomt hu s. Søynte med hotell drift i 1930. Ingen<br>større forandringer. Opphavet til de eksotiske navn på værelsene. Aldri<br>angret på at Meye rgården ble hotell. Av betydning at Mo hadde et sted hvor<br>tilreisende kunne ta inn og føle seg vel.<br>Malerisamlingen, blant annet a v kunstnere som bodde her, noen opptil flere år.<br>Min mann var særlig interessert i å støtte Nordlandsmalere.<br>(Gong-gong i bakgru nnen).<br>03<br>NRK gramm<br>80815<br>Jensson, Leif<br>

Les mindre ^

**Kilde for metadata**  
dra.nb.no (1990/03778.P)

**Varighet**  
00:13:01

Fra patrisierhjem til...  
Fra patrisierhjem til turis...

00:00:11 00:13:01



[02:50.000 --> rekke gamle ur  
[02:56.000 --> merker vi en c  
[03:02.000 -->  
[03:07.000 --> finner vi et r  
[03:13.000 -->  
[03:17.000 --> oppbevart på n  
[03:24.000 --> forteller her  
[03:31.000 --> heter for ekse  
[03:38.000 --> værelsenet nav  
[03:45.000 --> gjester.  
[03:51.000 --> kold pjolter p  
[03:57.000 --> jeg, for der k  
[04:03.000 -->

Tilbake i **Meiergården** **LOC** finnes enda en rikholdig våbensamling, en rekke gamle ur, gamle tresnitt, kobberstikk, pasteller og akvareller, og særlig merker vi en del kopier av presten **Heltsens** **PER** Rana bilder som oppbevares på **Bergens museum** **LOC**. En stor samling gammel porseleen er utstilt i skap, og dessuten finner vi et rikholdig bibliotek med litteratur fra slutten av forrige århundredet opptil i dag. Omtrent en tusen bind fra det 18 århundredet og tidligere er oppbevart på museet. Nå er heim blitt **Meiergårdens** **LOC** turisthotell, med mange om eget forteller her, mer om privathjemmet enn om hotellet. Værelsenet har navn, ikke nummer. Telefonkiosken heter det for eksempel **Ringebu** **LOC**. Toalettene, **Victoria Falls** **LOC** og **Niagara Falls** **LOC**, forøvrig bærer værelsenet navn fra alle jordens kanter. Og bo her er som en jordomseling når den går på visit til andre gjester. Skal vi drikke kaffe i **Cairo** **GPE\_LOC** i dag, eller hva sier de til den kold pjolter på **Himalaya** **LOC**? Og dratt i **Blocksberg** **GPE\_LOC**, svarer kanskje de. Ja, gjerne det, svarer jeg, for der bor fruen selv og mottar sine gjester med den største gjestfrihet.

ing, en  
særlig  
s museum.  
essuten  
i dag.  
e er  
eget  
onkiosken  
rer  
L andre  
L den  
svarer

# Norwegian wav2vec 2.0?

- XLS-R Multilingual wav2vec 2.0 (Baevski et al., 2020)
- Swedish VoxRex (Malmsten, Haffenden, Börjeson, 2022)
- NPSC-based models (Solberg and Ortiz, 2022)

# Norwegian wav2vec 2.0?

- XLS-R Multilingual wav2vec 2.0 (Baevski et al., 2020)
- Swedish VoxRex (Malmsten, Haffenden, Börjeson, 2022)
- NPSC-based models (Solberg and Ortiz, 2022)
- HuggingFace Challenge results

| Model                            | Final WER |
|----------------------------------|-----------|
| NbAiLab/nb-wav2vec2-1b-bokmaal   | 6.33      |
| NbAiLab/nb-wav2vec2-300m-bokmaal | 7.03      |
| NbAiLab/nb-wav2vec2-300m-nynorsk | 12.22     |

# Norwegian ASR: Wav2Vec2

- Data
  - NST: Norwegian Speech Synthesis
  - NPSC: Norwegian Parliamentary Speech Corpus
- Bokmål and Nynorsk
- VoxRex and XLS-R

| Size | Model                        | NPSC        | NPSC (Bokmål) | NST         |
|------|------------------------------|-------------|---------------|-------------|
| 300M | <i>No language model</i>     |             |               |             |
|      | NPSC-Bokmål                  | 11.76       | 9.79          | 21.46       |
|      | NST                          | 24.50       | 22.45         | 5.52        |
|      | NST-NPSC                     | 9.58        | 8.86          | 5.44        |
|      | NST-NPSC-Bokmål              | 10.37       | 8.33          | 5.49        |
|      | <i>5-gram language model</i> |             |               |             |
|      | NPSC-Bokmål                  | 9.07        | 7.14          | 19.19       |
|      | NST                          | 19.41       | 17.33         | <b>4.38</b> |
|      | NST-NPSC                     | <b>7.60</b> | <b>6.92</b>   | 4.39        |
|      | NST-NPSC-Bokmål              | 10.05       | 7.96          | 4.42        |
| 1B   | <i>No language model</i>     |             |               |             |
|      | NPSC-Bokmål                  | 9.49        | 7.51          | 17.64       |
|      | NST                          | 25.07       | 22.94         | 5.08        |
|      | NST-NPSC                     | 8.99        | 7.14          | 5.25        |
|      | NST-NPSC-Bokmål              | 8.69        | 6.46          | 4.93        |
|      | <i>5-gram language model</i> |             |               |             |
|      | NPSC-Bokmål                  | 8.37        | 6.41          | 14.94       |
|      | NST                          | 21.47       | 19.36         | 4.39        |
|      | NST-NPSC                     | 8.03        | 6.15          | 4.54        |
|      | NST-NPSC-Bokmål              | <b>8.02</b> | <b>5.81</b>   | <b>4.30</b> |
|      | Ortiz and Burud (2021)       | 20.64       |               |             |
|      | Solberg and Ortiz (2022)     | 17.10       |               |             |

De la Rosa, Javier, Rolv-Arild Braaten, Per Egil Kummervold, and Freddy Wetjen (2023). "Boosting Norwegian Automatic Speech Recognition." In: *Proceedings of the 24rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Faroe Islands: Linköping University Electronic Press, Sweden



# Whisper

## Multitask training data (680k hours)

### English transcription

- 👤 "Ask not what your country can do for ..."
- 📄 Ask not what your country can do for ...

### Any-to-English speech translation

- 👤 "El rápido zorro marrón salta sobre ..."
- 📄 The quick brown fox jumps over ...

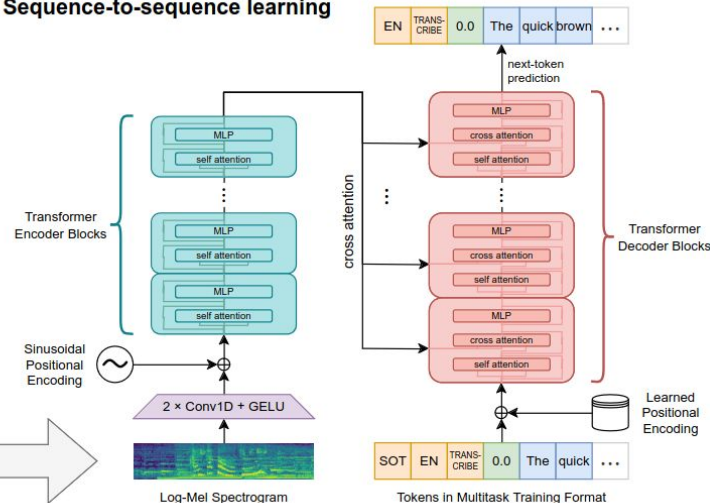
### Non-English transcription

- 👤 "언덕 위에 올라 내려다보면 너무나 넓고 넓은 ..."
- 📄 언덕 위에 올라 내려다보면 너무나 넓고 넓은 ...

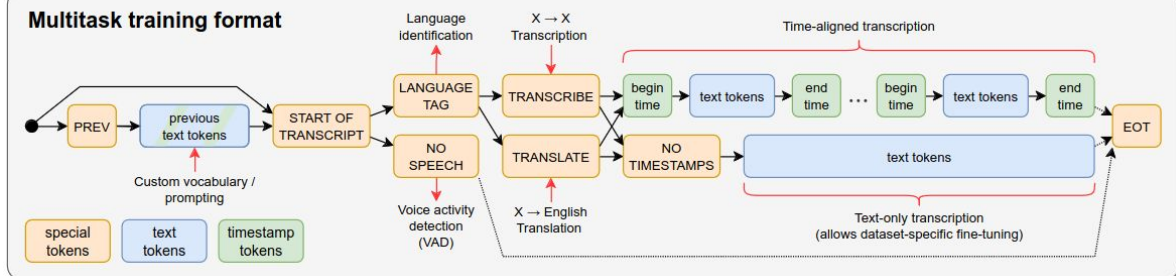
### No speech

- 🔊 (background music playing)
- 📄 ∅

## Sequence-to-sequence learning



## Multitask training format



Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision.," 2022.

# Norwegian ASR: Wav2Vec2 vs Whisper

- Data
  - NST: Norwegian Speech Synthesis
  - NPSC: Norwegian Parliamentary Speech Corpus
- Bokmål and Nynorsk
- VoxRex and XLS-R

| Size | Model                        | FLEURS      |
|------|------------------------------|-------------|
| 300M | <i>No language model</i>     |             |
|      | NPSC-Bokmål                  | 18.51       |
|      | NST                          | 13.94       |
|      | NST-NPSC                     | 12.43       |
|      | NST-NPSC-Bokmål              | 12.51       |
|      | <i>5-gram language model</i> |             |
|      | NPSC-Bokmål                  | 12.98       |
|      | NST                          | 11.27       |
|      | NST-NPSC                     | 9.93        |
|      | NST-NPSC-Bokmål              | <b>9.88</b> |
|      | Whisper small (244M)         | 24.20       |
| 1B   | <i>No language model</i>     |             |
|      | NPSC-Bokmål                  | 16.26       |
|      | NST                          | 13.05       |
|      | NST-NPSC                     | 11.17       |
|      | NST-NPSC-Bokmål              | 11.53       |
|      | <i>5-gram language model</i> |             |
|      | NPSC-Bokmål                  | 13.03       |
|      | NST                          | 11.53       |
|      | NST-NPSC                     | <b>9.87</b> |
|      | NST-NPSC-Bokmål              | 10.00       |
|      | Whisper large (1.55B)        | 11.4        |
|      | Whisper large-v2 (1.55B)     | 9.5         |

De la Rosa, Javier, Rolv-Arild Braaten, Per Egil Kummervold, and Freddy Wetjen (2023). "Boosting Norwegian Automatic Speech Recognition." In: *Proceedings of the 24rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Faroe Islands: Linköping University Electronic Press, Sweden

# Sámi Whisper?

- No support for Sámi in Whisper
- Very small supervised dataset
- Use of closest language in Whisper for transfer: Finnish

# Sámi Whisper?

- No support for Sámi in Whisper
- Very small supervised dataset
- Use of closest language in Whisper for transfer: Finnish
- Experimental version: <https://huggingface.co/NbAiLab/whisper-large-sme>
  - WER: 24.91

# Sámi Whisper

Read speech:



Ja de bosui davvebiegga nu garrasit go sáhtii, muhto mađi eanet son bosui, dađi čavgadeappot vánddardeaddji giesai jáhka iežas birra.

*And then the North Wind started blowing as hard as it could, but the harder the wind blew down the road, the tighter the man clung to his coat.*

- **Wav2Vec2:** ja de bosui davvebiegga nu garrasiid go sáhii muhto mađi eanes son bosui dađi čávga deappo vánddardeaddji gieasaid jáhka eižas birra
- **Whisper:** ja de bosui davvebiegga nu garrasit go sáhtii muhto mađi eanet son bosui dađi čávga lea eambo go vánddardeaddji geasái jáhke iežas birra

# Sámi Whisper

## Spontaneous speech:



Gal han dat lea áhkku nu ahte... áhkku lea nu rábmái, ahte dat muitala buot gili áhkkuide, dat ahte doppe lea olmborranrásiid levsset badjin, ahte son gártá juohke beavve viehkat mihcamáraid maŋŋá viežžamin daid olmborranrásiid, gal han son... gal han son ii viečča daid.

*So grandmother is like... grandmother is so bragg, that she tells all the grandmothers in the village, that there are garden angelicas [...], that after the midsummer she has to go harvesting garden angelicas every day, but she... she does not (really) harvest them.*

- **Wav2Vec2:** galhan dat lea áhkku nu ahte áhkku lea nu rábmái ahte dat muitala buot gili áhkui dat doppe lea olbmo buorrá rási levšsebadjin ahte son geartá juohkebeai li biehkat mihžamárai maŋá veažžimin daid olbmo borra rási gal han suon galhan son ii viežža daid
- **Whisper:** gal han dat lea áhkku nu ahte áhkku lea nu rábmái ahte dat muitala buot gili áhkkuid dat doppe lea olbmo borra rásiid leavssobadjin ahte son gártá juohke beavve viehkat mihcamat leat maŋŋá viežžamin daid olbmo borra rásiid gal han son gal han son ii vieččan daid

# NB-Whisper

- Goal: Train a Seq2Seq model for Norwegian ASR
  - Full focus since August 2022
  - Change in direction October 2022
  - From unsupervised + supervised to weak-supervision
- Aligned 60.000 hours of legally deposited material
  - 7 million - 30 second snippets
  - Subtitles, parliamentary speeches, audio books
- Developed the first training scripts for Whisper using Jax on TPU
  - Collaboration with HuggingFace and Google TPU Research Cloud

# NB-Whisper

- Norwegian speech to Nynorsk, Bokmål or English
- “Any language” to Nynorsk or Bokmål
- Model sizes:
  - Tiny (39M)
  - Base (74M)
  - Small (244M)
  - Medium (769M)
  - Large (1550M)
- Improving the quality of Norwegian transcription with “at least two model classes” compared to OpenAIs models.
- Able to control the style of the transcripts.





# NorHand\_v0.7

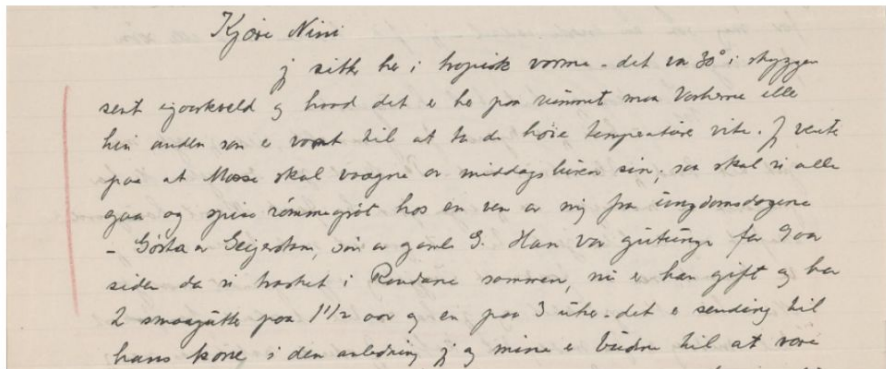
# HTR?

Free Public AI Model for Handwritten Text Recognition with [Transkribus](#)

[Transkribus](#) > [Public Models](#) > [NorHand\\_v0.7](#)

[« Back to all public models](#)

This model is a general model for Norwegian handwriting from approx. 1820 to 1940. The training data set includes letters and journals of 15 Norwegian writers, including Camilla Collett (1813-1895), Henrik Ibsen (1828-1906), Knut Hamsun (1859-1952) and Sigrid Undset (1882-1949), among others.



## Model Overview

**Name:**

NorHand\_v0.7

**Creator:**

National Library of Norway

**Model ID:**

37762

**Century:**

19th, 20th

**Languages:**

Norwegian

**Script:**

Latin alphabet

**Engine:**

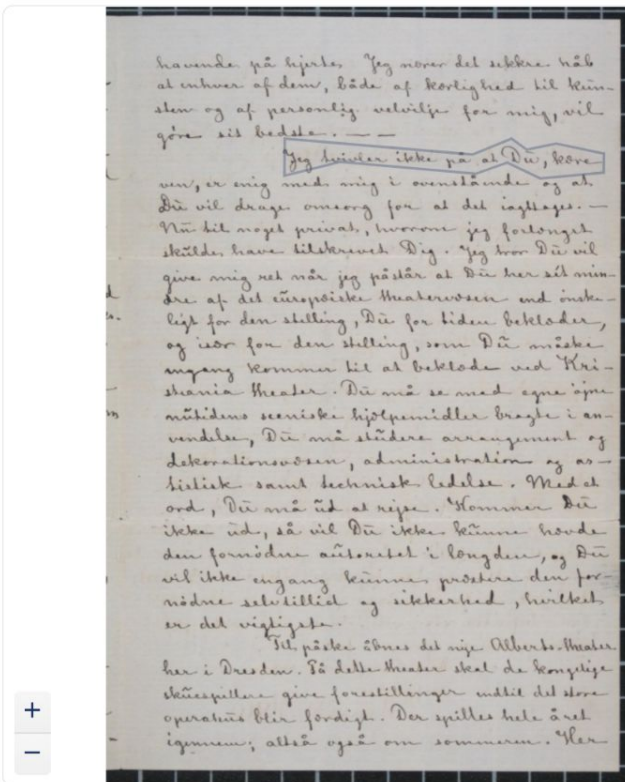
HTR+

**Material:**

Handwritten

**CER on validation set:**

3.74 %



havende på hjerte Jeg nærer det sikke håb  
at enhver af dem både af kærlighed til kun-  
sten og af personlig velvilje for mig, vil  
gøre sit bedste.-

Jeg tvivler ikke på at Du, kære

ven, er enig med mig i ovenstående og at  
Du vil drage omsorg for at det iagttages.  
Nu til noget privat, hvormed jeg forlængst  
skulde have tilskrevet Dig. Jeg tror Du vil  
give mig ret når jeg påstår at Du her sit min-  
dre af det europæiske theatervæsen end ønske-  
ligt for den stilling, Du for tiden beklæder  
og især for den stilling, som Du måske  
engang kommer til at beklæde ved Kri-  
stiania theater. Du må se med egne øjne  
nutidens sceniske hjælpemidler bragte i an-  
vendelse, Du må studere arrangement og

Vis alle objekter  ?

Søk på avsender/forfatter

Søk på mottaker

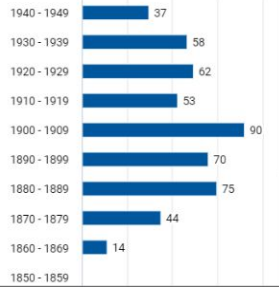
Søk etter tittel

Søk etter tema/emne

DATO

Fra dato

Til dato



Filter 515 Treff i brev og manuskripter

**Brevs. 140 Brev fra Backer, Harriet (1845-1932) til...**  
Backer, Harriet  
Tilgang for alle

Side 3 «...Det gjør ikke noget, om Du synes, jeg er dum. Kjærlig Hilsen til Hulda. **Din hengivne** Harriet Backer Glem for Guds Skyld ikke H hans Eventyrstue paa Holmenkollen. 4de Juledag...»

[Flere treff](#)

**Brevs. 177 Brev fr... Backer, H... Tilgang for alle**

Side 3 «...I gode gamle Dage! Gratulerer Kjære trofaste Ven og ønsker Dig alt Velsignet og Godt. **Din hengivne** Harriet Kjære Hulda. Gratulerer Dig hjertelig og ønsker jeg var hos Dig! Du vet, jeg...»

[Flere treff](#)

**Brevs. 563 Brev fra Backer, Harr... Tilgang for alle**

Side 6 «...Vi tænker Begge to idag paa Ellif. Velsignet er hans Mindel! Jeg er saa bedrøvet. **Din hengivne** gamle Ven Harriet B: Kjære Signe Meyer. Tak at De husket paa mig! Det var...»

[Flere treff](#)

**Brevs. 39 Brev fra Ibsen, Henrik (1828-1906) til Hansen,...**  
Ibsen, Henrik  
09.01.1896  
Tilgang for alle

Side 2 «...eller i den anden retning. Med mange venskabelige hilsener og gode nytårsønsker tegner jeg mig **din hengivne** og forbundne Henrik Ibsen. kan give vedkommende direktør, herr Abrahams, noget afgørende svar hverken...»

[Flere treff](#)

**Brevs. 915 Brev fra Thiis, Jens (1870-1942) til Vigeland,...**  
Thiis, Jens  
29.11.1911  
Tilgang for alle

Side 2 «...men - allikevel! - vil du ikke overveie det? Ta mig ialfall ikke spørsmålet ilde op. **Din hengivne** Jens Thiis 29 nov. 1911. Hr. billedhugger Gustav Vigeland Sorgenfribakken 2 Her. Kjære Gustav...»

[Flere treff](#)

**Brevs. 915 Brev fra Thiis, Jens (1870-1942) til Vigeland...**  
Thiis, Jens  
01.02.1920  
Tilgang for alle

Side 3 «...fordeles mellem "Fontænen" og din ældre produktion. Jeg håper på nogen linjer fra dig snarest. **Din hengivne** Jens. H. I sin Staffeldtsgate 6 1 Febr. 1920. Kjære Gustav, Jeg sender dig her korrekturavtryk av...»

[Flere treff](#)

**Brevs. 915 Brev fra Thiis, Jens (1870-1942) til...**  
Thiis, Jens  
06.06.1902  
Tilgang for alle

Side 5 «...på det imødkommende svar, og når det vi har det, garanterer jeg kjøbet af Ibsenbysten! **Din hengivne** ven Jens Dit ønske om forskud på urnen fremlegger jeg i første arbeidsudvalgs møde...»

[Flere treff](#)

**Brevs. 915 Brev fra Thiis, Jens (1870-1942) til...**  
Thiis, Jens  
03.02.1903  
Tilgang for alle

Side 7 «...nemlig endnu igjen at forberede mig til et foredrag som jeg skal holde om to timer. **Din hengivne** ven Jens Ragna ber hilse dig. I Bergen ifald du ikke får den foran universitet og...»

[Flere treff](#)

**Brevs. 39 Brev fra Ibsen, Henrik 21.11.1898 Tilgang for alle**

Side 1 «...Henrik Ibsen. Kære ven! Herved sender jeg dig den ønskede udtalelse. Undskyld forsiknelsen **Din hengivne** H. I...»

[Flere treff](#)

**Brevs. 305 Brev fra Undset, Sigrd (1882-1949) til Kamstrup...**  
Undset, Sigrd  
23.12.1935  
Tilgang for alle

Side 1 «...Kunsterforbundet Kjeld Stubs gate Oslo. Kjære Kitt, Glædelig jul og godt nytta! ønskes dig av **din hengivne** Sigrd Undset...»

[Flere treff](#)

**Brevs. 742 Brev fra Undset, Sigrd (1882-1949) til Møller,...**  
Undset, Sigrd  
21.12.1926  
Tilgang for alle

**Brevs. 140 Brev fra Backer, Harriet (1845-1932) til...**  
Backer, Harriet  
24.01.1920  
Tilgang for alle

**Brevs. 664 Brev fra Undset,...**  
Undset, Sigrd  
Sep. 1932  
Tilgang for alle

**Brevs. 742 Brev fra Undset, Sigrd...**  
Undset, Sigrd  
25.11.1928  
Tilgang for alle

**Brevs. 305 Brev fra Undset, Sigrd (1882-1949) til Kamstrup...**  
Undset, Sigrd  
22.12.1924  
Tilgang for alle

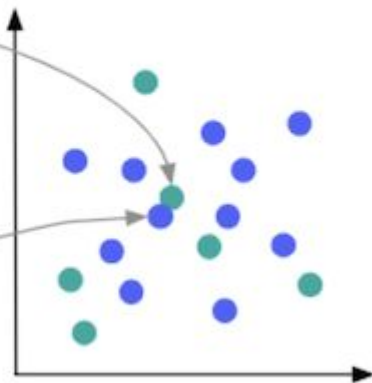
Sortering: Relevans

# Embedding Space



“woodblock print of the Edo period depicting three boats moving through a storm-tossed sea with a large wave forming a spiral in the centre and Mount Fuji visible in the background”

Joint embedding space  
(typically a vector space of  
dimension 512 or 768)



kids playing in the snow





kids play



output 0

horse in a livingroom

horse in a livingroom ..... 10000%

output 1

output 0

woman playing the guitar

woman playing the guitar ..... 10000%

output 1





### Søk med tekst

- Frase søk
- And søk
- Or søk
- Wildcard søk
- Fuzzy søk

Fråse:

Postkort

Risikå Bilad

Sommarsposten

Svalbardposten

Utvid liste



roger jøseveld



| Bilde_id                         | WEB   |
|----------------------------------|---|
| 1) NB_RABL_P_AE2000065534_0011   | 1) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0011   |
| 2) NB_RABL_P_AE2000065534_0001   | 2) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0001   |
| 3) NB_RABL_P_AE20000651106_0001  | 3) nb.no/items/URN:NBN:no-nb_digifoto_AE20000651106_0001  |
| 4) NB_RABL_P_AE2000065534_0014   | 4) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0014   |
| 5) NB_RABL_P_AE2000065534_0002   | 5) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0002   |
| 6) NB_RABL_P_AE2000065534_0012   | 6) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0012   |
| 7) NB_RABL_P_AE2000065534_0003   | 7) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0003   |
| 8) NB_RABL_P_AE2000065534_0005   | 8) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0005   |
| 9) NB_RABL_P_AE20000651106_0001  | 9) nb.no/items/URN:NBN:no-nb_digifoto_AE20000651106_0001  |
| 10) NB_RABL_P_AE20000661135_0001 | 10) nb.no/items/URN:NBN:no-nb_digifoto_AE20000661135_0001 |
| 11) NB_RABL_P_AE2000065534_0013  | 11) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0013  |
| 12) NB_RABL_P_AE2000065534_0007  | 12) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0007  |
| 13) NB_RABL_P_AE2000065534_0015  | 13) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0015  |
| 14) NB_RABL_P_AE2000065534_0010  | 14) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0010  |
| 15) NB_RABL_P_AE2000065534_0008  | 15) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0008  |
| 16) NB_RABL_P_AE2000065534_0009  | 16) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0009  |
| 17) NB_RABL_P_AE2000065534_0006  | 17) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0006  |
| 18) NB_RABL_P_AE2000065534_0004  | 18) nb.no/items/URN:NBN:no-nb_digifoto_AE2000065534_0004  |
| 19) NB_RABL_P_AE20000548119_0011 | 19) nb.no/items/URN:NBN:no-nb_digifoto_AE20000548119_0011 |
| 20) NB_RABL_P_AE20000801129_0003 | 20) nb.no/items/URN:NBN:no-nb_digifoto_AE20000801129_0003 |

Total number of tuples: 10000 Start tuple number: 1 Size of set: 24

# Postcards



Flere bilder av personer i det bildet du har valgt





Maps?





Maps?







# Mímir Project

Evaluating the Impact of Copyrighted Materials on Generative Large Language Models for Norwegian Languages

# Data

| <b>Configurations</b>  | <b>Dataset</b>                    |
|--|-----------------------------------|
| mimir-base   | mimir-base                        |
| mimir-extended   | mimir-extended                    |
| mimir-base + all other newspapers                              | mimir-newspapers                  |
| mimir-base + all other newspapers + all nonfiction             | mimir-factual                     |
| mimir-base + all nonfiction                                    | mimir-nonfiction                  |
| mimir-base + all fiction                                       | mimir-fiction                     |
| mimir-base + all books   | mimir-books                       |
| mimir-base + all other newspapers + all books - all translated | mimir-untranslated-withnewspapers |
| mimir-base + all books - all translated                        | mimir-untranslated                |
| mimir-base + all books - all untranslated                      | mimir-translated                  |
| mimir-base + all books - all other newspapers                  | mimir-rightholders                |

# Models trained

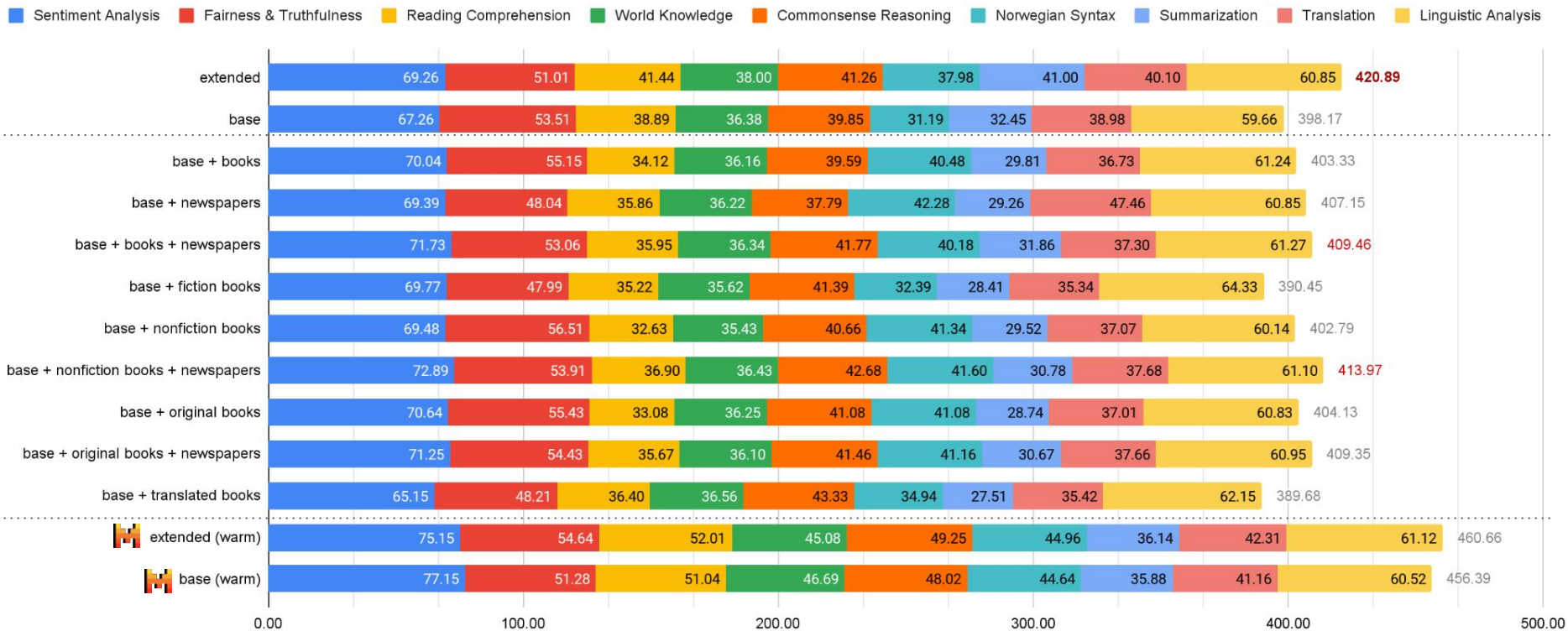
| Status | Initialization                                | Data                              | Name   |
|--------|---|-----------------------------------|--|
| ✓      | From scratch                                  | mimir-base                        | <a href="#">mimir-mistral-7b-base-scratch</a>        |
| ✓      | From scratch                                  | mimir-extended                    | <a href="#">mimir-mistral-7b-extended-scratch</a>    |
| ✓      | Pre-existing                                  | mimir-base                        | <a href="#">mimir-mistral-7b-base</a>                |
| ✓      | Pre-existing                                  | mimir-extended                    | mimir-mistral-7b-extended                            |
| ✓      | <a href="#">mimir-mistral-7b-base-scratch</a> | mimir-fiction                     | <a href="#">mimir-7b-fiction</a>                     |
| ✓      | <a href="#">mimir-mistral-7b-base-scratch</a> | mimir-nonfiction                  | <a href="#">mimir-7b-nonfiction</a>                  |
| ✓      | <a href="#">mimir-mistral-7b-base-scratch</a> | mimir-factual                     | <a href="#">mimir-7b-factual</a>                     |
| ✓      | <a href="#">mimir-mistral-7b-base-scratch</a> | mimir-newspapers                  | <a href="#">mimir-7b-newspapers</a>                  |
| ✓      | <a href="#">mimir-mistral-7b-base-scratch</a> | mimir-books                       | <a href="#">mimir-7b-books</a>                       |
| ✓      | <a href="#">mimir-mistral-7b-base-scratch</a> | mimir-rightholders                | mimir-7b-rightholders                                |
| ✓      | <a href="#">mimir-mistral-7b-base-scratch</a> | mimir-untranslated-withnewspapers | <a href="#">mimir-7b-untranslated-withnewspapers</a> |
| ✓      | <a href="#">mimir-mistral-7b-base-scratch</a> | mimir-untranslated                | <a href="#">mimir-7b-untranslated</a>                |
| ✓      | <a href="#">mimir-mistral-7b-base-scratch</a> | mimir-translated                  | <a href="#">mimir-7b-translated</a>                  |

# Evaluation

1. Sentiment Analysis
2. Fairness & Truthfulness
3. Reading Comprehension
4. World Knowledge
5. Commonsense Reasoning
6. Norwegian Syntax
7. Summarization
8. Translation
9. Linguistic Analysis

# Evaluation

## Aggregated Scores per Model Skill

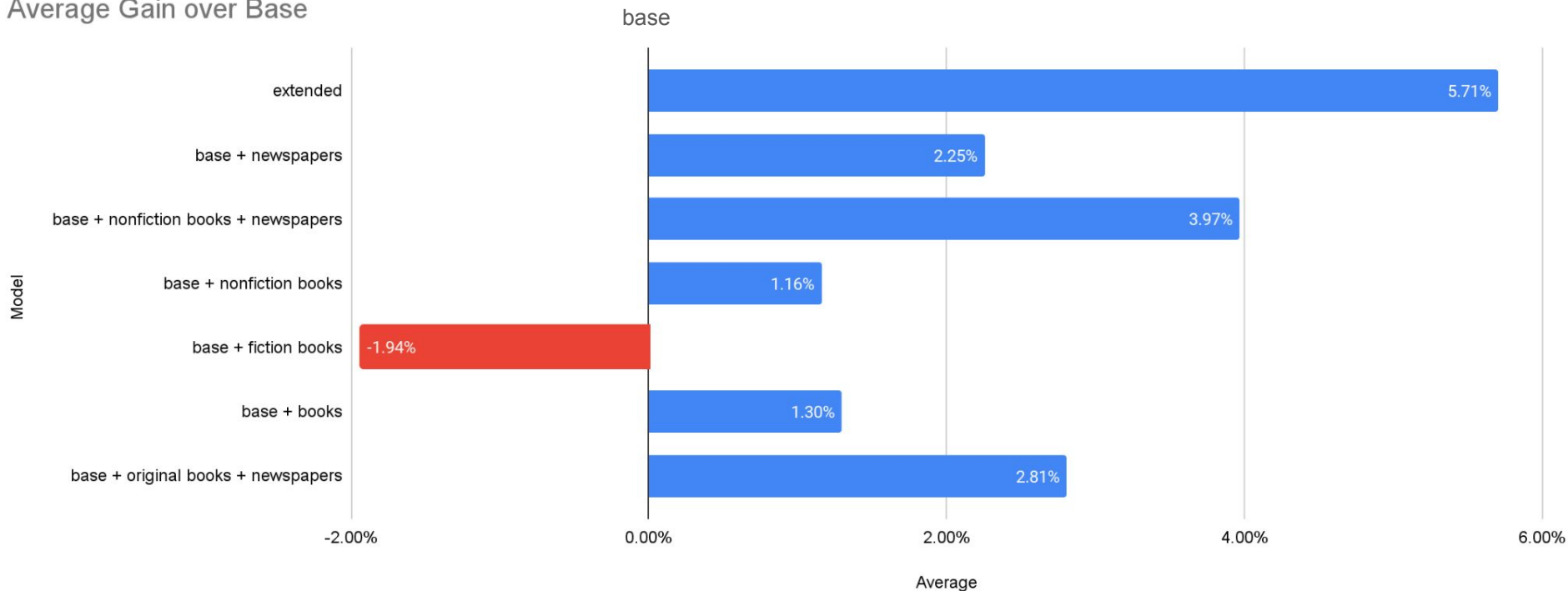


# Evaluation: Ablations from Scratch

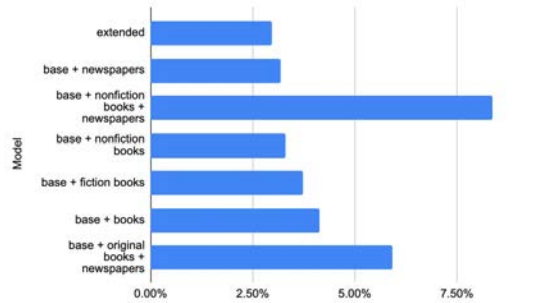


# Evaluation: Ablations from Scratch

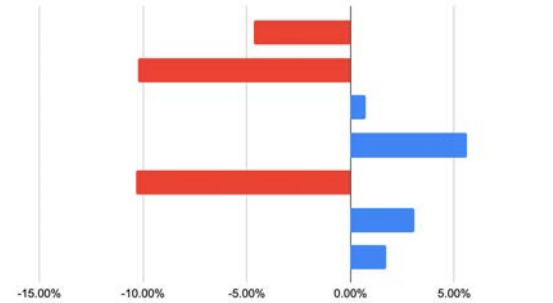
Average Gain over Base



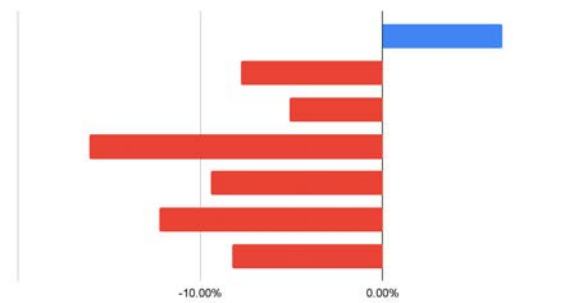
Sentiment Analysis



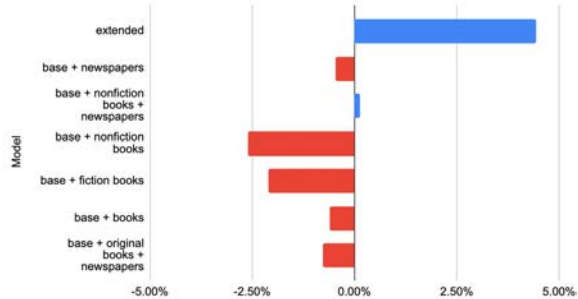
Fairness &amp; Truthfulness



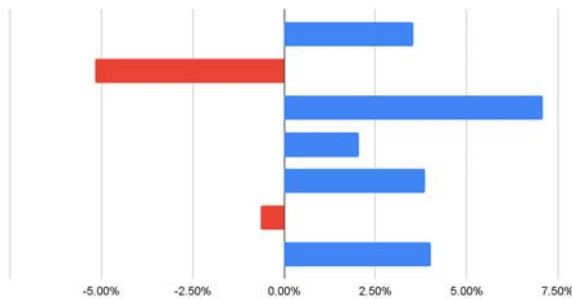
Reading Comprehension



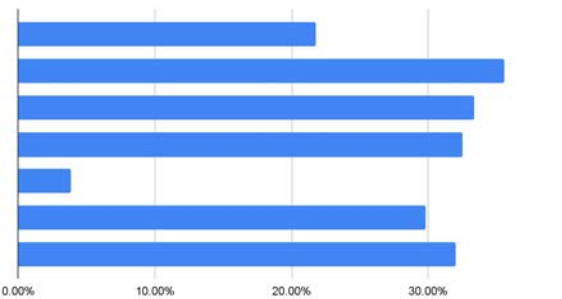
World Knowledge



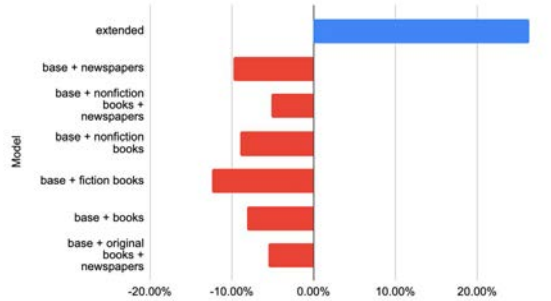
Commonsense Reasoning



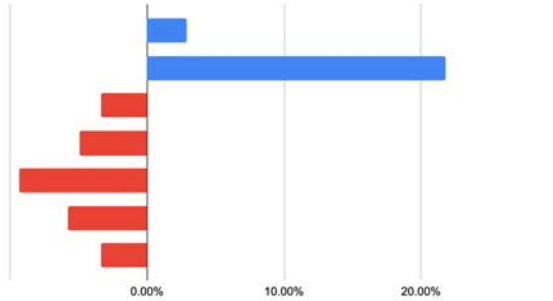
Norwegian Syntax



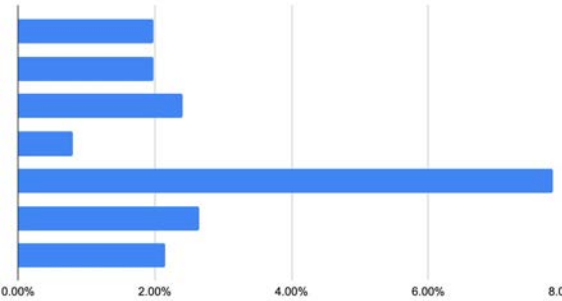
Summarization



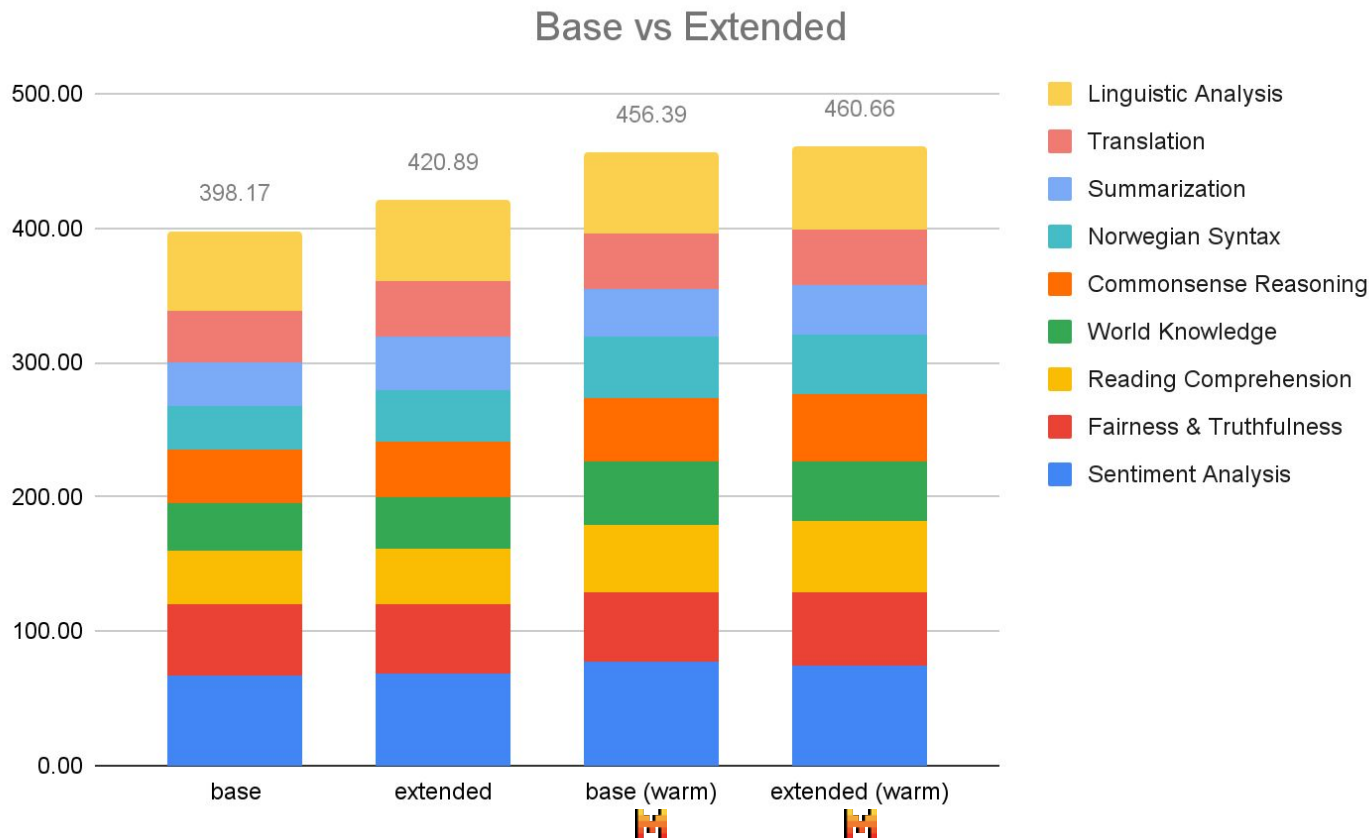
Translation



Linguistic Analysis



# Evaluation: Pre-Training



# Benefits of all of this?

## Obvious

- Better access to content for more people
- Easier to find relevant content
- Greater possibility of use and reuse
- New ways of working internally
- Good contribution to society

## Not-so-obvious

- Greater focus on the content rather than the carrier of the content
- Changed view of the collection
- Digital tools of course

# Thanks!

## Questions?

Javier de la Rosa  **AI-lab**  
[versa@nb.no](mailto:versa@nb.no) National Library of Norway

[www.nb.no](http://www.nb.no)

[ai.nb.no](http://ai.nb.no)