

# Architecting Sovereign Norwegian AI

## and the European Legal Landscape

Javier de la Rosa •  **Språkmodeller**  
[versae@nb.no](mailto:versae@nb.no) • National Library of Norway

# Timeline

2005

## Digital Manifesto

Kickstarting massive digitization



## The Digital National Library of Norway

**Strategic platform 2005**  
The National Library of Norway, February 2005

### The Library Comes to the User

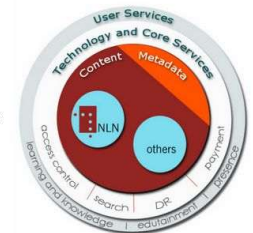
Mature information—and communication technology and the Internet—have given the library new possibilities when it comes to professional cooperation and a coordinated use of resources. The most distinctive opportunity is the possibility of providing access to a rich and varied content on everyone's desk 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.

# Timeline

2005

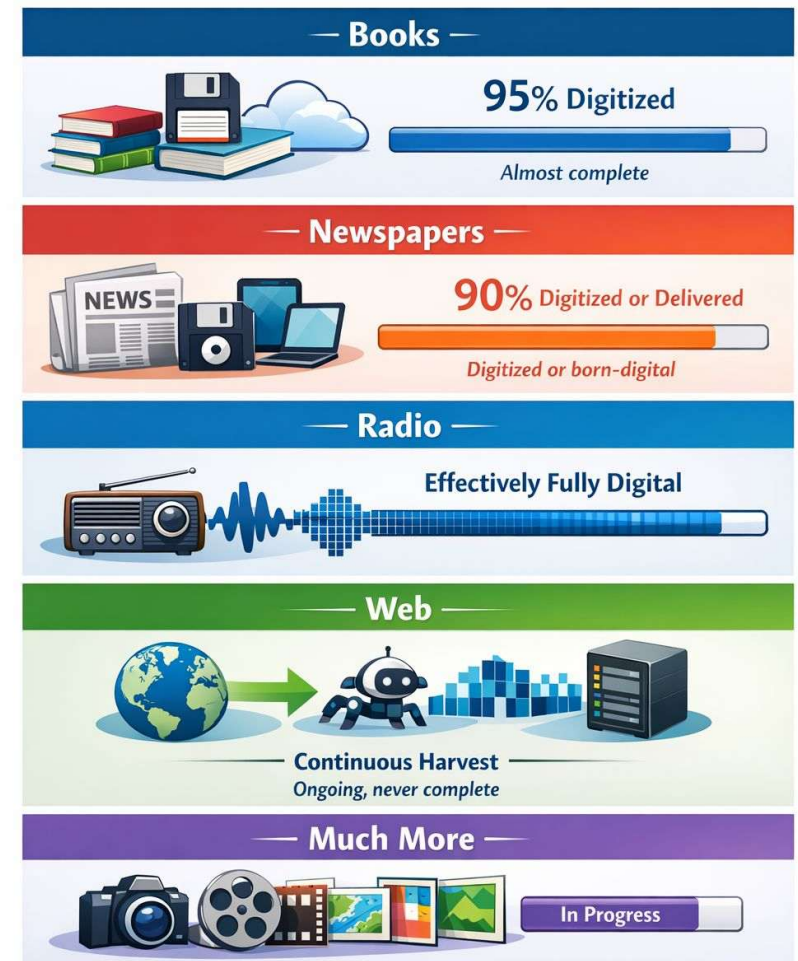
2018

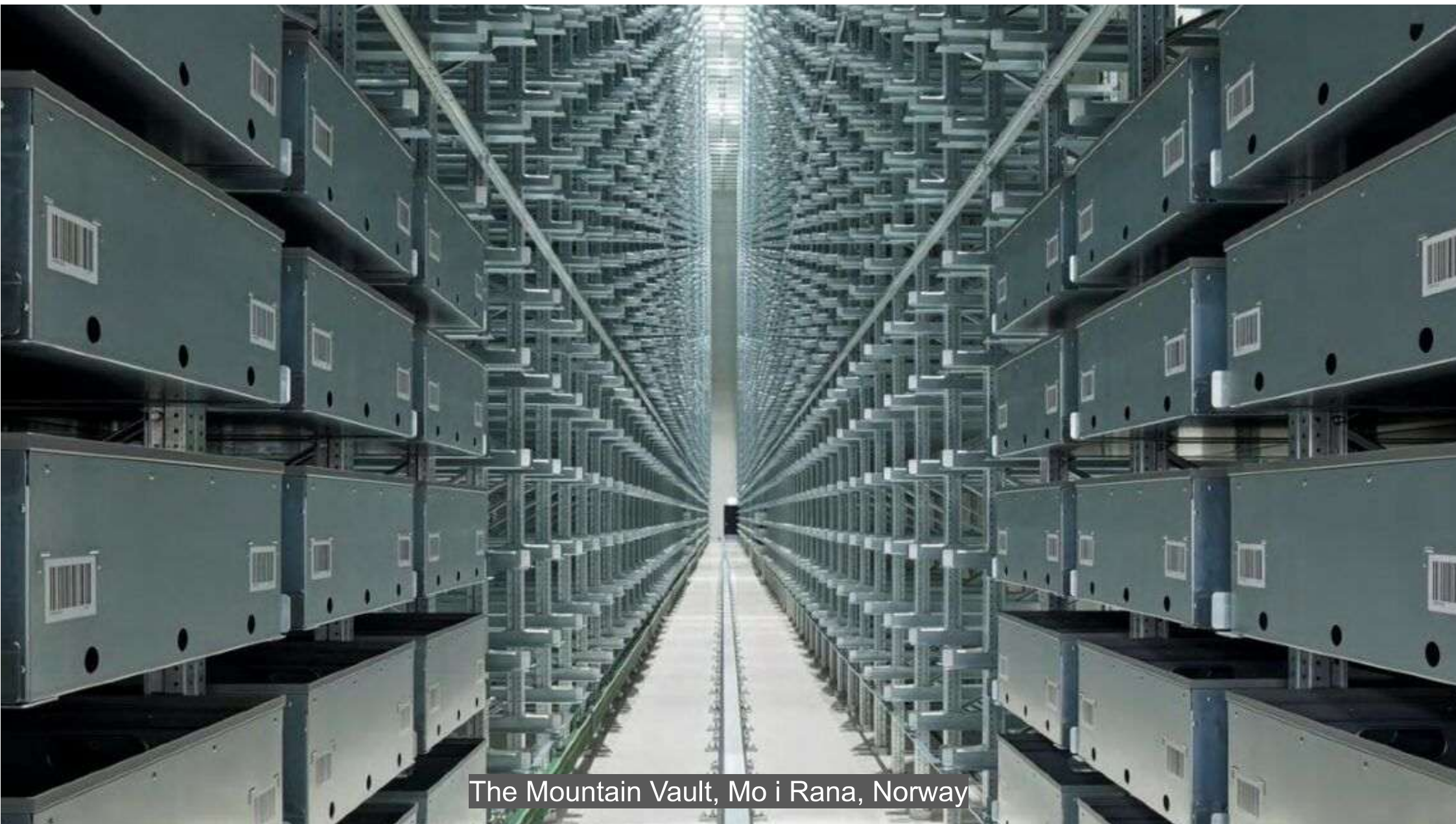
## Digital Manifesto

Kickstarting massive digitization

## AI Lab

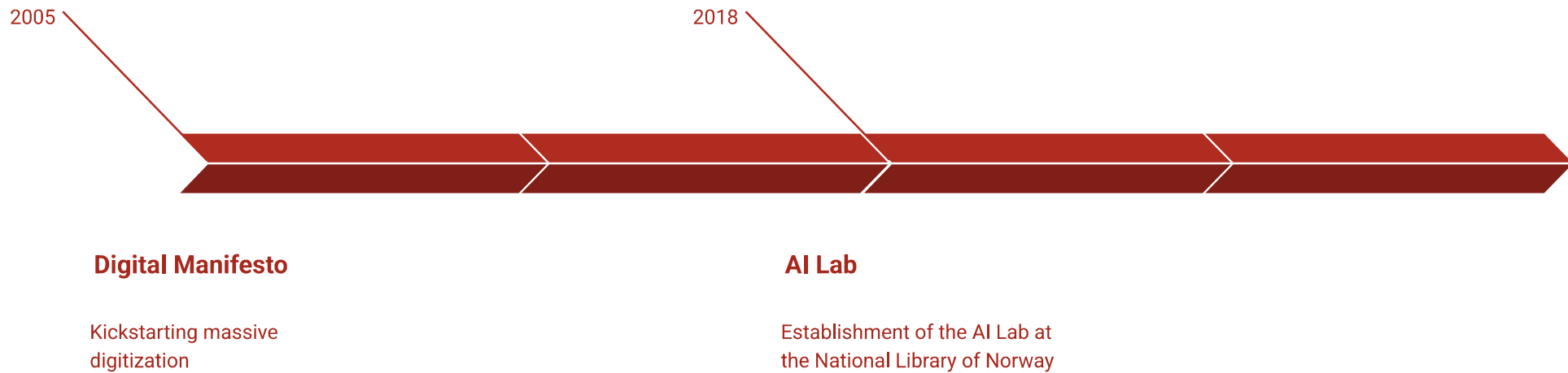
Establishment of the AI Lab at the National Library of Norway





The Mountain Vault, Mo i Rana, Norway

# Timeline



# Timeline

2005

*"The department hereby asks the to, if possible, examine the value language models. Relevant Norw Authors' and publishers' organiz*

*We ask that the National Library, possible compensation scheme for scheme."*



to Gov Mímir

*d research/development project  
ining of Norwegian generative  
ed to participate in the project.*

*rch project, assess the basis for a  
prepare a proposal for such a*

# Mímir project: Mandate from the Ministry of Culture

The aim of the assignment from Ministry of Culture is threefold:

1. Examining the value of copyright-protected material in training Norwegian generative language models
2. Assess the basis for a possible compensation scheme for Norwegian rights holders
3. Prepare proposals for a compensation scheme

Objective 1 was within the mandate of Mímir, while objectives 2 and 3 were carried out outside the project but still by the National Library.

# Mímir Project

The data from the collection of the National Library of Norway that is copyrighted seems to be of better quality (i.e., yields better LLMs) than the data that is not protected by copyright.

## The Impact of Copyrighted Material on Large Language Models: A Norwegian Perspective

Javier de la Rosa<sup>1</sup> Vladislav Mikhailov<sup>2</sup> Lemei Zhang<sup>3</sup> Freddy Wetjen<sup>1</sup> David Samuel<sup>2</sup>

Peng Liu<sup>3</sup> Rolv-Arild Braaten<sup>1</sup> Petter Mæhlum<sup>2</sup> Magnus Breder Birkenes<sup>1</sup>

Andrey Kutuzov<sup>2</sup> Tita Enstad<sup>1</sup> Hans Christian Farsethås<sup>2</sup> Svein Arne Bryggfeldt<sup>1</sup>

Jon Atle Gulla<sup>3</sup> Stephan Oepen<sup>2</sup> Erik Velldal<sup>2</sup> Wilfred Østgulen<sup>1</sup> Lilja Øvrelid<sup>2</sup>

Aslak Sira Myhre<sup>1</sup>

<sup>1</sup>National Library of Norway

<sup>2</sup>University of Oslo

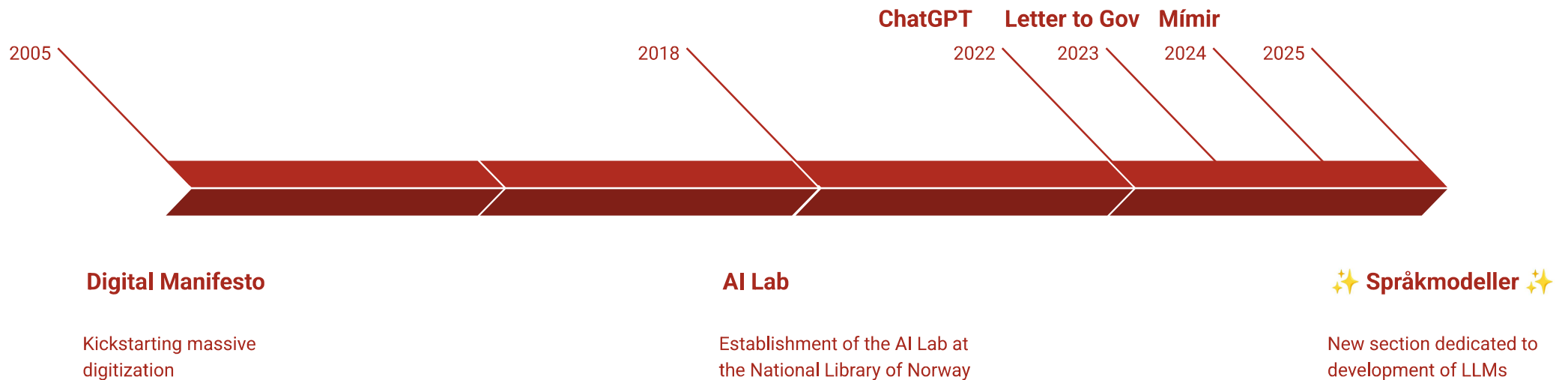
<sup>3</sup>Norwegian University of Science and Technology

Correspondence: versae@nb.no

Aggregated Scores per Model Skill



# Timeline



## Tildelingsbrev 2025 ("Letter of allocation")

*The National Library **has been tasked with training, updating, and making available Norwegian and Sámi language models**, and for this purpose, its allocation has been increased by NOK 20 million. As a general rule, the National Library shall offer **base models** that actors in the public sector or business can fine-tune and further develop to perform more specific tasks.*

*Through the Ministry of Education's budget, an additional NOK 20 million has been allocated to the state high-performance computing company Sigma2, for financing computing power for training the language models.*

# Tildelingsbrev 2025 ("Letter of allocation")

## ✨ Språkmodeller ✨

Angelina, Freddy, Ines, Javier, Katri, Lucas, Markus,  
Marthe, Pawel, Per Egil, Phoebe, Rolv-Arild, Victoria

+ Arne, Magnus, Marie, Marie, Thea, Tita (Språkbanken)

*Through the Ministry of Education's budget, an additional NOK 20 million has been allocated to the state high-performance computing company Sigma2, for financing computing power for training the language models.*

# Tildelingsbrev 2025 ("Letter of allocation")

## ✨ Språkmodeller ✨

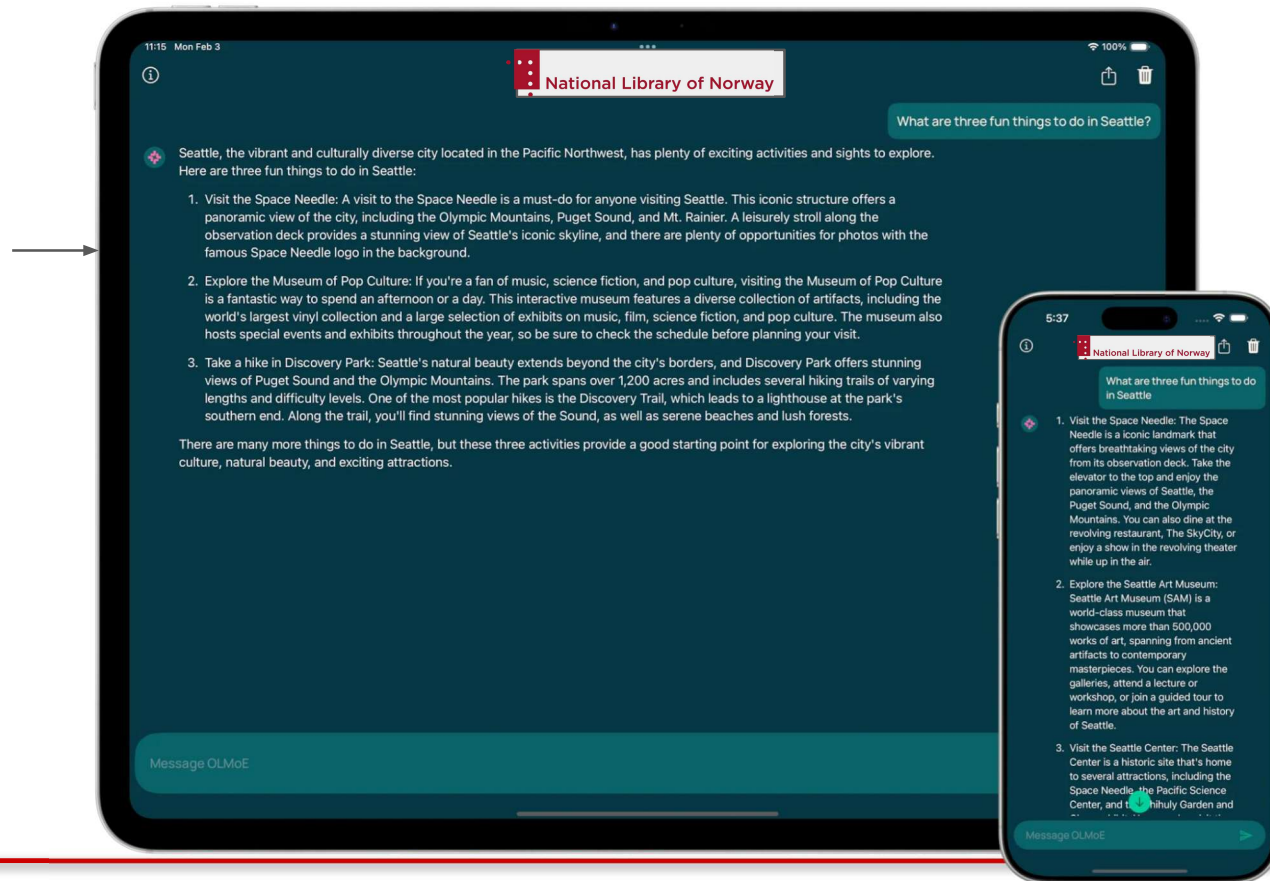
Angelina, Freddy, Ines, Javier, Katri, Lucas, Markus,  
Marthe, Pawel, Per Egil, Phoebe, Rolv-Arild, Victoria

+ Arne, Magnus, Marie, Marie, Thea, Tita (Språkbanken)

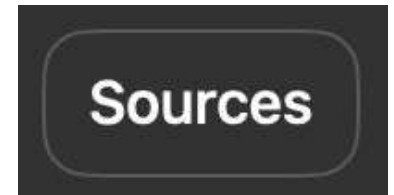
~1.5M GPU hours/year

# Language Models Ambition

Multimodal  
Aligned LLM



Voice  
input/output



RAG

Sovereign Norwegian AI and the European Legal Landscape  
Språkteknologi, rettsstat og demokrati – Oslo, Feb 2026

# Language Models Plan

High priority:

1. A series of GPT-based textual models ← Aligned LLM

Medium priority:

2. Improving automatic speech recognition and text-to-speech ← Voice in/out

Lower priority:

3. Modernized encoder-only models }  
4. Embedding models } ← RAG  
5. Image encoder ← Multimodality

# Project Lines

- NB-GPT
  - Started Feb 1st, 2025
- NB-TTS
  - Started Sept 5th, 2025
- NB-ASR
  - Started Oct 1st, 2025
- NB-EMBED
  - Started Feb 4th, 2026
- NB-VLM
  - Q4, 2026

# “Open” data

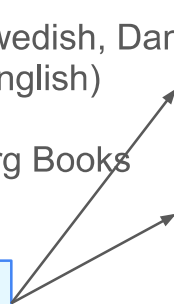
- aurora-open v1 (open license)
  - ✓ Books (public domain)
  - ✓ Newspapers (public domain)
  - ✓ Målfrid
  - ✓ Parliament
  - ✓ Public reports
  - ✓ Norsk Lovtidend
  - ✓ Wikipedia
  - ✓ HPLTv3
  - ✓ FineWeb2
  - ✓ FineWeb-HQ (Swedish, Danish)
  - ✓ FineWeb-Edu (English)
  - ✓ Digimanus
  - ✓ Project Gutenberg Books
  - ✓ MegaMath
  - ✓ StarCoder
  - ✓ GiellaT Sámi
  - ✓ Sámi Web
  - ✓ FinePDFs

# “Open” data

## - aurora-open v1 (open license)

- ✓ Books (public domain)
- ✓ Newspapers (public domain)
- ✓ Målfrid
- ✓ Parliament
- ✓ Public reports
- ✓ Norsk Lovtidend
- ✓ Wikipedia
- ✓ HPLTv3
- ✓ FineWeb2
- ✓ FineWeb-HQ (Swedish, Danish)
- ✓ FineWeb-Edu (English)
- ✓ Digimanus
- ✓ Project Gutenberg Books
- ✓ MegaMath
- ✓ StarCoder
- ✓ GiellaT Sámi
- ✓ Sámi Web
- ✓ FinePDFs

| Type               | Language   | Compiler           | Dataset           | Words              | License          | Lang Identification |
|--------------------|------------|--------------------|-------------------|--------------------|------------------|---------------------|
| Speech             | North Sámi | Divvun/Giellatekno | salmon-whisper    | 24,639             | Restricted       | Manual              |
| Speech             | Lule Sámi  | Divvun/Giellatekno | salmon-whisper    | 18,857             | Restricted       | Manual              |
| Book               | North Sámi | NB (up to 2021)    | sámi-bibliography | 6,033,738          | In-copyright     | NB                  |
| Book               | South Sámi | NB (up to 2021)    | sámi-bibliography | 194,407            | In-copyright     | NB                  |
| Book               | Lule Sámi  | NB (up to 2021)    | sámi-bibliography | 231,053            | In-copyright     | NB                  |
| Newspaper          | North Sámi | NB (2018-2022)     | avvir-pdf         | 6,437,918          | In-copyright     | NB                  |
| Web                | North Sámi | Divvun/Giellatekno | sikor-freecorpus  | 13,283,080         | Uncertain/Public | N/A                 |
| Web                | South Sámi | Divvun/Giellatekno | sikor-freecorpus  | 646,273            | Uncertain/Public | N/A                 |
| Web                | Lule Sámi  | Divvun/Giellatekno | sikor-freecorpus  | 535,367            | Uncertain/Public | N/A                 |
| API/Reports/Web    | North Sámi | NB (2020-2025)     | målfrid           | 14,602,310         | Public           | GlottID             |
| API/Reports/Web    | Lule Sámi  | NB (2020-2025)     | målfrid           | 3,306,972          | Public           | GlottID             |
| API/Reports/Web    | North Sámi | NB (2020-2025)     | målfrid           | 1,660,958          | Public           | GlottID             |
| Web                | North Sámi | UiO LTG            | saami-web         | 11,896,087         | Uncertain/Public | GlottID             |
| Book/Newspaper/Web | North Sámi | NB (up to 2024)    | mimir-extended    | 99,965,100         | In-copyright     | GlottID             |
| Book/Newspaper/Web | South Sámi | NB (up to 2024)    | mimir-extended    | 48,003             | In-copyright     | GlottID             |
| Book/Newspaper/Web | Lule Sámi  | NB (up to 2024)    | mimir-extended    | 22,519             | In-copyright     | GlottID             |
| <b>TOTAL</b>       |            |                    |                   | <b>158,907,281</b> |                  |                     |



# Sámi textual data

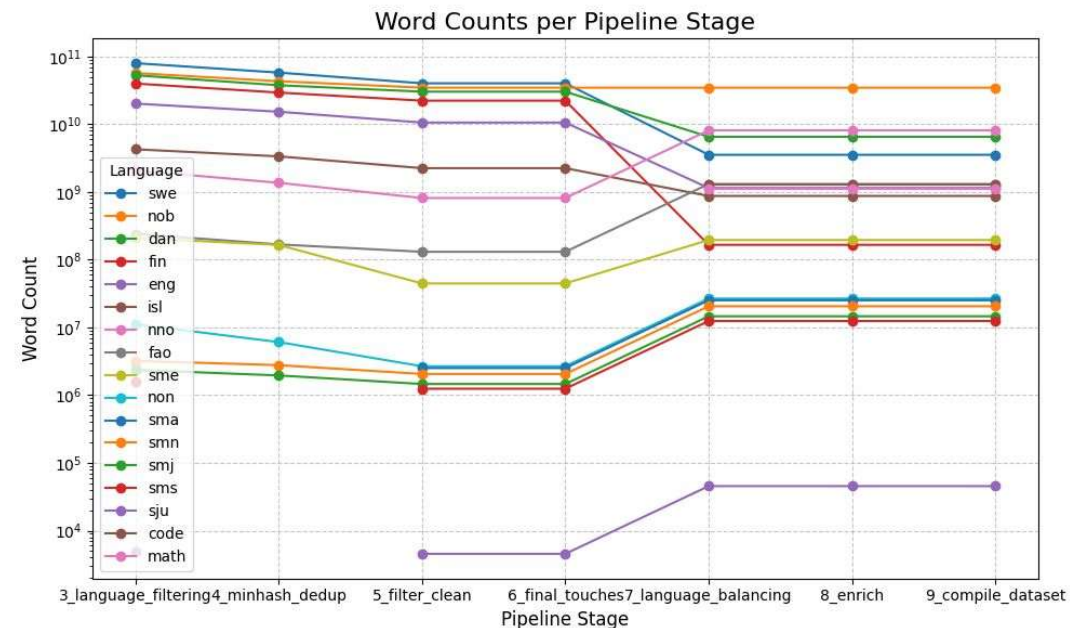
| Type            | Language   | Compiler           | Dataset          | Words      | License          | Lang Identification |
|-----------------|------------|--------------------|------------------|------------|------------------|---------------------|
| Speech          | North Sámi | Divvun/Giellatekno | salmon-whisper   | 24,639     | Restricted       | Manual              |
| Speech          | Lule Sámi  | Divvun/Giellatekno | salmon-whisper   | 18,857     | Restricted       | Manual              |
| Web             | North Sámi | Divvun/Giellatekno | sikor-freecorpus | 13,283,080 | Uncertain/Public | N/A                 |
| Web             | South Sámi | Divvun/Giellatekno | sikor-freecorpus | 646,273    | Uncertain/Public | N/A                 |
| Web             | Lule Sámi  | Divvun/Giellatekno | sikor-freecorpus | 535,367    | Uncertain/Public | N/A                 |
| API/Reports/Web | North Sámi | NB (2020-2025)     | målfrid          | 14,602,310 | Public           | GlottLID            |
| API/Reports/Web | Lule Sámi  | NB (2020-2025)     | målfrid          | 3,306,972  | Public           | GlottLID            |
| API/Reports/Web | North Sámi | NB (2020-2025)     | målfrid          | 1,660,958  | Public           | GlottLID            |
| Web             | North Sámi | UiO LTG            | saami-web        | 11,896,087 | Uncertain/Public | GlottLID            |
| TOTAL           |            |                    |                  | 45,974,543 |                  |                     |

# Sámi textual data

| Type               | Language   | Compiler           | Dataset           | Words              | License          | Lang Identification |
|--------------------|------------|--------------------|-------------------|--------------------|------------------|---------------------|
| Speech             | North Sámi | Divvun/Giellatekno | salmon-whisper    | 24,639             | Restricted       | Manual              |
| Speech             | Lule Sámi  | Divvun/Giellatekno | salmon-whisper    | 18,857             | Restricted       | Manual              |
| Book               | North Sámi | NB (up to 2021)    | sámi-bibliography | 6,033,738          | In-copyright     | NB                  |
| Book               | South Sámi | NB (up to 2021)    | sámi-bibliography | 194,407            | In-copyright     | NB                  |
| Book               | Lule Sámi  | NB (up to 2021)    | sámi-bibliography | 231,053            | In-copyright     | NB                  |
| Newspaper          | North Sámi | NB (2018-2022)     | avvir-pdf         | 6,437,918          | In-copyright     | NB                  |
| Web                | North Sámi | Divvun/Giellatekno | sikor-freecorpus  | 13,283,080         | Uncertain/Public | N/A                 |
| Web                | South Sámi | Divvun/Giellatekno | sikor-freecorpus  | 646,273            | Uncertain/Public | N/A                 |
| Web                | Lule Sámi  | Divvun/Giellatekno | sikor-freecorpus  | 535,367            | Uncertain/Public | N/A                 |
| API/Reports/Web    | North Sámi | NB (2020-2025)     | målfrid           | 14,602,310         | Public           | GlottLID            |
| API/Reports/Web    | Lule Sámi  | NB (2020-2025)     | målfrid           | 3,306,972          | Public           | GlottLID            |
| API/Reports/Web    | North Sámi | NB (2020-2025)     | målfrid           | 1,660,958          | Public           | GlottLID            |
| Web                | North Sámi | UiO LTG            | saami-web         | 11,896,087         | Uncertain/Public | GlottLID            |
| Book/Newspaper/Web | North Sámi | NB (up to 2024)    | mimir-extended    | 99,965,100         | In-copyright     | GlottLID            |
| Book/Newspaper/Web | South Sámi | NB (up to 2024)    | mimir-extended    | 48,003             | In-copyright     | GlottLID            |
| Book/Newspaper/Web | Lule Sámi  | NB (up to 2024)    | mimir-extended    | 22,519             | In-copyright     | GlottLID            |
| <b>TOTAL</b>       |            |                    |                   | <b>158,907,281</b> |                  |                     |

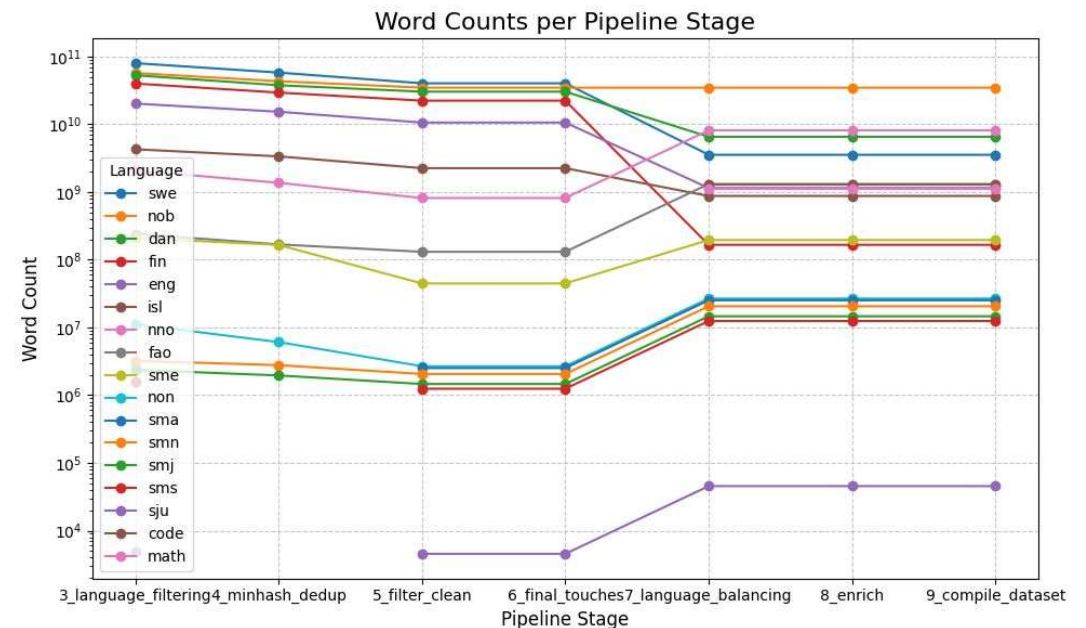
# Data Pipelines

- Datatrove
- GlotLID LangID
- Dedup
- Heuristic filtering/cleaning
- Language balancing
- Copyright index
- Domain filter
- Shuffling
- Quality estimation
- Wiki category classifier
- HQ and Long subsets



# Data Pipelines

- Datatrove
- GlotLID LangID
- Dedup
- Heuristic filtering/cleaning
- Language balancing
- Copyright index
- Domain filter
- Shuffling
- Quality estimation
- Wiki category classifier
- HQ and Long subsets



# Data Pipelines

## Copyright index

- Check for long form copyright matches
- ~1000 matches from the Internet

## Domain filter

- Checks for domains from a list of copyrighted newspapers
- ~20M matches from the Internet!

# Data Pipelines

From a final corpus of 90M docs, almost 20M scraped from the Internet belonged to in-copyright newspapers under the agreement



# Språkmodeller på norsk leaderboard

Enkelt leaderboard for modellene vi har evaluert sålangt

### Select Evaluation Suite:

NorEval

✕ ▼

### Filter models by size (B)



Selected Model Size Range: 0B - 4.5B

### Selected models ranked

Filter model names (substring)

Filter column names (substring)

☒  Instruction-tuned

☒  Pre-trained

Settings

| model_name                                                                                                            | all_categories_all_languages_rank | all_categories_nno_rank | all_categories_nob_rank | all_tasks_all_languages_rank | all_tasks_nno_rank |
|-----------------------------------------------------------------------------------------------------------------------|-----------------------------------|-------------------------|-------------------------|------------------------------|--------------------|
|  google/gemma-2-2b                     | 1                                 | 1                       | 3.5                     | 1                            | 1                  |
|  google/gemma-2-2b-it                  | 2                                 | 3                       | 2                       | 2                            | 2                  |
|  tiiuae/Falcon-H1-3B-Instruct          | 3                                 | 4                       | 3.5                     | 3                            | 4                  |
|  HuggingFaceTB/SmolLM3-3B              | 4.5                               | 2                       | 6                       | 4.5                          | 3                  |
|  Qwen/Qwen3-1.7B                     | 4.5                               | 7                       | 1                       | 4.5                          | 7                  |
|  BSC-LT/salamandra-2b-instruct       | 6                                 | 8                       | 7                       | 10                           | 9                  |
|  google/gemma-3-1b-pt                | 7                                 | 5                       | 11.5                    | 6                            | 5                  |
|  utter-project/EuroLLM-1.7B          | 8                                 | 9.5                     | 9                       | 8.5                          | 12                 |
|  utter-project/EuroLLM-1.7B-Instruct | 9.5                               | 15                      | 5                       | 7                            | 13                 |
|  Qwen/Qwen2.5-1.5B-Instruct          | 9.5                               | 6                       | 14                      | 8.5                          | 6                  |
|  Qwen/Qwen2.5-1.5B                   | 11                                | 13                      | 10                      | 14                           | 16                 |
|  google/gemma-3-1b-it                | 12                                | 12                      | 11.5                    | 11                           | 8                  |

# Språkmodeller på norsk leaderboard

Enkelt leaderboard for modellene vi har evaluert så langt

## Select Evaluation Suite:

NorEval

## Filter models by size (B)



Selected Model Size Range: 0B - 4.5B

## Selected models ranked

Filter model names (substring)

Filter column names (substring)

☒ Instruction-tuned ☒ Pre-trained

Settings

| model_name                          | all_categories_rank | all_tasks_rank | all_tasks_all_languages_rank | all_tasks_nno_rank |
|-------------------------------------|---------------------|----------------|------------------------------|--------------------|
| google/gemma-2-2b                   | 1                   | 1              | 1                            | 1                  |
| google/gemma-2-2b-it                | 2                   | 2              | 2                            | 2                  |
| tiiuae/Falcon-H1-3B-Instruct        | 3                   | 3              | 4                            | 4                  |
| HuggingFaceTB/SmolLM3-3B            | 4.5                 | 4.5            | 3                            | 3                  |
| Qwen/Qwen3-1.7B                     | 4.5                 | 4.5            | 7                            | 7                  |
| BSC-LT/salamandra-2b-instruct       | 6                   | 7              | 10                           | 9                  |
| google/gemma-3-1b-pt                | 7                   | 11.5           | 6                            | 5                  |
| utter-project/EuroLLM-1.7B          | 8                   | 9              | 8.5                          | 12                 |
| utter-project/EuroLLM-1.7B-Instruct | 9.5                 | 15             | 7                            | 13                 |
| Qwen/Qwen2.5-1.5B-Instruct          | 9.5                 | 6              | 8.5                          | 6                  |
| Qwen/Qwen2.5-1.5B                   | 11                  | 13             | 14                           | 16                 |
| google/gemma-3-1b-it                | 12                  | 12             | 11                           | 8                  |

Sámi evaluation  
is missing!



# Pre-training

- Baseline (minimal)
- Same setup
- Models up to 1B
- Repeat on aurora-open
  - Compare with baseline
    - Stage 1: train on full dataset (warmup + standard)
    - Stage 2: + aurora-open-high (high quality pre-training)
    - Stage 3: + aurora-open-long (long-context expansion)
- Transfer to law models

# Post-training

## General capabilities (SFT)

-  Tülu (machine translated, +700k, openly licensed)
-  NPK & NDLA (parallel bokmål-nynorsk dataset)
-  mimir-instructions

## Instruction evolution

- <https://github.com/arcee-ai/EvolKit> (easy)
- <https://github.com/QwenLM/AutoIF> (complex)

# Norwegian LLM Leaderboard DEV SUITE

Simple leaderboard ranking LMs on Norwegian tasks

### Filter models by size (B)



### Selected models ranked ?

Filter model names (substring)

Filter column names (substring)

☒ Instruction-tuned ☐ Pre-trained

Settings

Download table

| model name                                                            | Category: all Language: nor (nob and nno) (rank) | Category: all Language: nob (rank) | Category: all Language: nno (rank) | Category: all Language: eng (rank) | Category: all Language: all (rank) | Task: all Language: nor (nob and nno) (rank) | Task: all Language: nob (rank) | Task: all Language: nno (rank) |
|-----------------------------------------------------------------------|--------------------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|----------------------------------------------|--------------------------------|--------------------------------|
| NbAiLab/nb-gpt-gemma3-27b-instruct-aurora-sft-2512-filtered-posttrain | 1                                                | 1                                  | 2                                  | 6                                  | 1                                  | 2                                            | 2                              | 1                              |
| NbAiLab/nb-gpt-qwen3-14b-base-aurora-sft-2512-filtered-posttrain      | 2                                                | 2                                  | 5                                  | 2.5                                | 2                                  | 1                                            | 1                              | 7                              |
| NbAiLab/nb-gpt-gemma3-12b-instruct-aurora-sft-2512-filtered-posttrain | 3                                                | 3                                  | 6                                  | 11                                 | 5                                  | 5                                            | 5                              | 4                              |
| swiss-ai/Apertus-8B-Instruct-2509                                     | 4                                                | 6                                  | 1                                  | 25                                 | 10                                 | 9                                            | 11                             | 2                              |
| mistralai/Mistral-Nemo-Instruct-2407                                  | 5                                                | 9                                  | 3                                  | 14                                 | 6                                  | 10                                           | 14                             | 5                              |
| tiiuae/Falcon-H1-34B-Instruct                                         | 6                                                | 4                                  | 9                                  | 1                                  | 3                                  | 4                                            | 3                              | 13                             |
| mistralai/Mistral-Small-3.1-24B-Instruct-2503                         | 7                                                | 5                                  | 7                                  | 2.5                                | 4                                  | 3                                            | 4                              | 8                              |
| google/gemma-3-12b-it                                                 | 8                                                | 10                                 | 4                                  | 13                                 | 7.5                                | 7                                            | 10                             | 3                              |
| NbAiLab/nb-gpt-gemma3-4b-instruct-                                    | 9                                                | 8                                  | 12                                 | 24                                 | 14.5                               | 17                                           | 17                             | 15                             |

# Aurora Borealis

---

- Aurora corpus (open)
- Borealis 4B (preview)
- Released on Dec 2025
- Borealis 270M, 2B, 12B, 27B released on Jan 2026

Photo: Freddy Wetjen



## Borealis Preview

updated 2 days ago

Preview release of the Borealis family of instruction tuned models by the National Library of Norway.

 **NbAiLab/borealis-27b-instruct-preview**

 Image-Text-to-Text •  27B • Updated 3 days ago •  38 •  3

 **NbAiLab/borealis-12b-instruct-preview**

 Image-Text-to-Text •  12B • Updated 3 days ago •  25

 **NbAiLab/borealis-4b-instruct-preview**

 Image-Text-to-Text •  4B • Updated Dec 23, 2025 •  492 •  9

 **NbAiLab/borealis-1b-instruct-preview**

 Image-Text-to-Text •  1.0B • Updated 3 days ago •  32

 **NbAiLab/borealis-270m-instruct-preview**

 Text Generation •  0.3B • Updated 3 days ago •  31

**Note**  Native precision models (f32, bf16).



Photo: Freddy Wetjen



Språkmodeller  
National Library of Norway

# Alignment and safety

- Declines harmful, illegal, discriminatory, or adult content. SAFE
- Handles sensitive topics respectfully and factually. SAFE FACTUAL
- States it is a language model, not a person, with no gender or emotions. STYLE
- Uses clear, natural Norwegian (Bokmål or Nynorsk depending on input). STYLE
- Responds directly, stays on topic, avoids unnecessary length or repetition. HELPFUL
- Adapts level of detail to user request (short or detailed as needed). STYLE
- Matches user's tone, including humor or formality. STYLE

The background features a dark blue gradient with horizontal streaks of purple and blue. Scattered throughout are small white and blue dots, resembling a starry night sky. Four stylized, hand-drawn floral or star-like shapes are positioned around the central text: a red one in the top left, a green one in the top center, a light blue one in the top right, and a yellow one in the bottom right. A red one is also partially visible in the bottom left.

# MEANWHILE

*Acquiring better data...*

# Acquiring better data



State Budget 2026

# Training on newspapers yields better language models

Ministry of Culture and Equality

Press release | No: 51/25 | Date: 15.10.2025

The government proposes to allocate 45 million kroner for the National Library to enter into an agreement to use newspapers in training Norwegian and Sami language models.

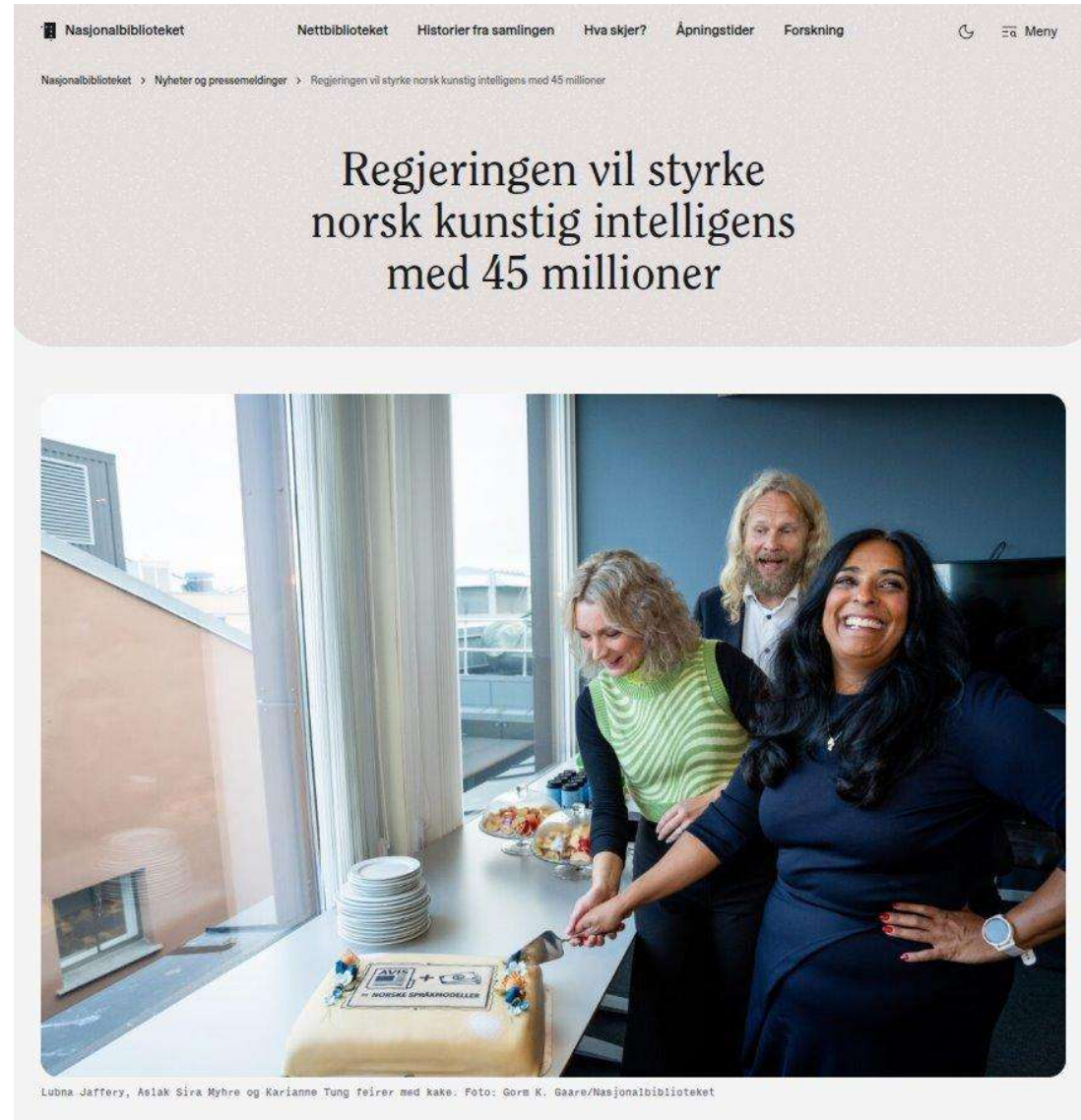
## THEME

[Film and media](#)[Language and literature](#)[State budget](#)

# Agreement

## Newspapers

- 45M NOK/year
- Articles up to 1 year ago
- Compiling ground-truth list every 3 months (opt-outs)
- Non open license: nb-license
- No distribution unless third-party enters into agreement with rights holders
- Support full life-cycle of models
- Signing Dec 19 2025



# Newspaper agreement

- NB representing the Norwegian state
  - Kopinor representing the newspapers
- 
- All newspapers, digitized and born digital, all publishing forms
  - Text only
  - Opt-out (only Morgenbladet has opted out)
  - 1-year delay

# Newspaper agreement

- Two (three) documents
  - Agreement
  - License (template)
  - FAQ
- Research under the agreement
  - Research on models
  - Publishing models trained as part of research
- Use of newspaper content for research
  - Not part of the agreement
  - Covered by Norwegian Law and Regulations

# NB-License

## [4] Use-based restrictions

You can only use the Model in compliance with the following conditions:

- A. You must comply with applicable laws and regulations.
- B. You must not use the model to intentionally recreate material from training data, whether protected by intellectual property rights or as personal data
- C. Neither the name of the National library nor the names of creators or publishers of training data may be used to endorse or promote products derived from this model without specific prior written permission.
- D. You must not use the Model or its outputs, to provide end-user services whose primary purpose is to summarize, restate, copy, or otherwise replace services for access to the press publications licensed as training data [reference to documentation], unless you have a separate agreement with the copyright holders. This includes, but is not limited to, services that provide users with functional substitutes for access to the original press content.

# Meanwhile... in Europe

# Meanwhile in Europe...



# EU AI Act



**1 August 2024**

The Act entered into force.



**2 February 2025**

Prohibited AI practices & AI literacy obligations



**2 August 2025**

Governance & GPAI rules applicable



**2 August 2026**

Most rules, incl. high-risk AI systems



**2 August 2027**

High-risk AI in regulated products

# EU AI Act

- ★ **Bloc-wide law (Regulation (EU) 2024/1689)** with rules for AI placed on the EU market or used in the EU.
- ★ **Support safe, trustworthy AI**, protect **fundamental rights**, and provide legal certainty while enabling innovation.
- ★ It uses a **risk-based approach** (more risk → more obligations).



# EU AI Act

- Entered into force on August 2025
- Entered into force 1 Aug 2025; banned practices + AI literacy apply 2 Feb 2025; GPAI obligations apply 2 Aug 2025; most rules fully apply 2 Aug 2026; some high-risk product transitions extend to 2 Aug 2027.
- In effect in Norway from July 2026
- Omnibus directive relaxing some of the requirements
  - High-risk rules originally set for Aug 2026, proposal to delay to Dec 2027

# EU AI Act

- Bans “unacceptable-risk” AI
  - Manipulative/deceptive uses, some social-scoring-type practices, etc.
- Regulates “high-risk” AI systems
  - Defined mainly by use case / sector, e.g., employment, education, critical infrastructure
  - Requirements: risk management, data governance, logging, transparency to users, human oversight, accuracy/robustness/cybersecurity, and (in many cases) conformity assessment before placing on the market.
- Adds transparency duties for “limited-risk” AI
  - E.g., letting people know they’re interacting with AI and marking certain synthetic content/deepfakes.

# EU AI Act

- Introduces rules for General-Purpose AI (GPAI) models and extra duties for GPAI with systemic risk
  - More stringent evaluation, risk mitigation, monitoring, cybersecurity, and incident processes
- Sets up governance and enforcement via the European AI Office and national authorities (KI-Norge), plus advisory bodies.
- Encourages innovation with tools like regulatory sandboxes (KI-Norge) in Member States.

# Meanwhile... in Europe

## EU AI Act

- Bans “unacceptable-risk” AI (certain manipulative/deceptive uses, some social-scoring-type practices, etc.).
  - Regulates “high-risk” AI systems (defined mainly by use case / sector, e.g., employment, education, critical infrastructure), requiring things like risk management, data governance, logging, transparency to users, human oversight, accuracy/robustness/cybersecurity, and (in many cases) conformity assessment before placing on the market. [OBJ]
- Adds transparency duties for “limited-risk” AI, e.g., letting people know they’re interacting with AI and marking certain synthetic content/deepfakes.
- Introduces rules for GPAI models and extra duties for GPAI with systemic risk (more stringent evaluation, risk mitigation, monitoring, cybersecurity, and incident processes).
- Sets up governance and enforcement via the European AI Office and national authorities, plus advisory bodies.
- Encourages innovation with tools like regulatory sandboxes (KI-Norge?) in Member States.

## Tildelingsbrev 2025 ("Letter of allocation")

**The National Library has been tasked with training, updating, and making available Norwegian and Sámi language models, and for this purpose, its allocation has been increased by NOK 20 million. As a general**

*“[An] indicative criterion for a model to be considered a general-purpose AI model is that its training compute is greater than  $10^{23}$  FLOP and it can generate language [...].”*

*“If a general-purpose AI model [...] does not display significant generality or is not capable of competently performing a wide range of distinct tasks, it is not a general-purpose AI model.”*

## NB as Provider of General Purpose AI Models

# EU AI Act

- All GPAI model providers must
  - Provide technical documentation and instructions for use
  - Comply with the Copyright Directive
  - Publish a summary about the content used for training
- Free and open licence GPAI model providers only need to comply with copyright and publish the training data summary, unless they present a systemic risk.
- All providers of GPAI models that present a systemic risk –open or closed– must also conduct model evaluations, adversarial testing, track and report serious incidents and ensure cybersecurity protections.

# EU AI Act Compute: GPAI vs Systemic Risk

| GPU Type           | Models ( $10^{25}$ FLOPs) | Models ( $10^{23}$ FLOPs) |
|--------------------|---------------------------|---------------------------|
| NVIDIA A100 (BF16) | 0.06 (~6% of one)         | ~6.7 models               |
| NVIDIA H100 (BF16) | 0.21 (~21% of one)        | ~21.4 models              |
| NVIDIA H100 (FP8)  | 0.43 (~43% of one)        | ~42.7 models              |
| NVIDIA B200 (BF16) | 0.49 (~49% of one)        | ~48.6 models              |
| NVIDIA B200 (FP8)  | 0.97 (~97% of one)        | ~97.2 models              |

# Tildelingsbrev 2025 ("Letter of allocation")

*The National Library has been tasked with training, updating, and making available Norwegian and Sámi language models, and for this purpose, its allocation has been increased by NOK 20 million. As a general rule, the National Library shall offer **base models** that actors in the public sector or business can fine-tune and further develop to perform more specific tasks.*

*Through the Ministry of Education's budget, an additional NOK 20 million has been allocated to the state high-performance computing company Sigma2, for financing computing power for training the language models.*

# Tildelingsbrev 2025 ("Letter of allocation")

*The National Library has been tasked with training, updating, and making available Norwegian and Sámi language models, and for this purpose, the National Library shall offer base models that actors in the public sector or business can fine-tune and further develop to perform more specific tasks.*

**General Purpose AI Model <  $10^{23}$  FLOPs  
~20 models/year**

*Through the Ministry of Education's budget, an additional NOK 20 million has been allocated to the state-owned computing company Sigma2, for financing computing power for training the language models.*

**~1.5M GPU hours/year**



# Tildelingsbrev 2025 ("Letter of allocation")

*The National Library has been tasked with training, updating, and making available Norwegian and Sámi language models, and for this purpose, its allocation is capped by NOK 20 million. As a general rule, the National Library shall offer base models that actors in the public sector or business can fine-tune and further develop to perform more specific tasks.*

**Systemic Risk >  $10^{25}$  FLOPs**  
**At best 20%**

*Through the Ministry of Education's budget, an additional NOK 20 million has been allocated to the state-owned supercomputing company Sigma2, for financing computing power for training the language models.*

**~1.5M GPU hours/year**



# EU AI Act and the National Library of Norway AI Models

- If released as open source (not systemic risk)
  - Copyright and Opt-outs
  - Data
  - Weights
  - Architecture
  - Usage
- If released as nb-license (not OSI compliant)
  - Same as open source plus
    - Training and testing procedures
    - Evaluation
  - Keep documentation up-to-date
  - Facilitate deployment
  - If trained from scratch
    - Document pre-training data
    - Document post-training and alignment
  - If trained from pre-existing weights
    - Document changes (fine-tuning data)
    - GPAI only if trained for 1/3 of compute of original model



# EU AI Act and the National Library of Norway AI Models



If released as Open Source  
(not systemic risk)

- © Copyright and Opt-outs
- 📁 Data
- 🔑 Weights
- 🏗️ Architecture
- ⚙️ Usage



If released as **nb-license**  
(not OSI compliant)

- Same as open source plus
  - 📋 Training and testing Evaluation
  - 📝 Keep documentation up-to-date
  - 🚀 Facilitate deployment

*All providers of GPAI models that present a systemic risk –open or closed– must also conduct model evaluations, adversarial testing, track and report serious incidents and ensure cybersecurity protections.*



If trained from scratch

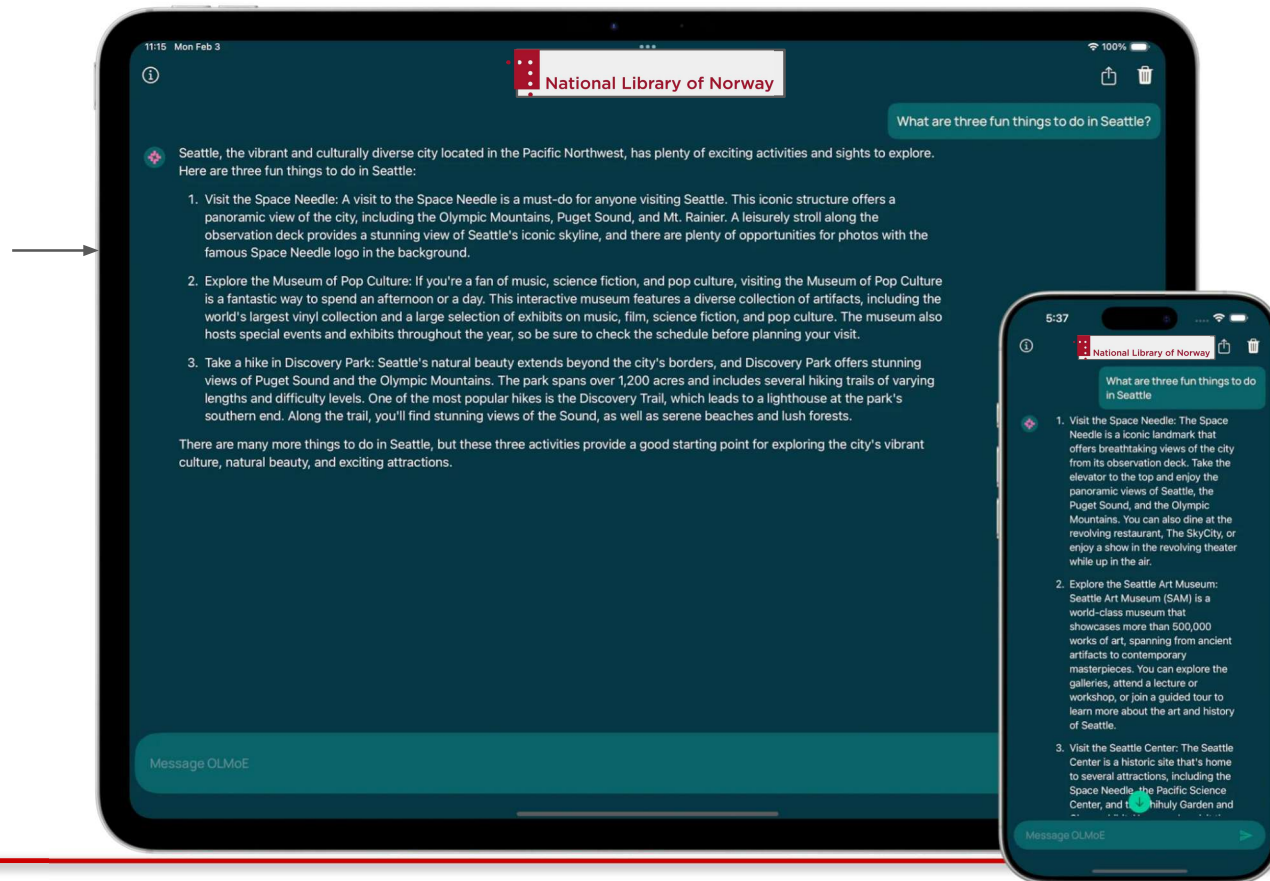
- ✓ Document pre-training data
- ✓ Document post-training and alignment

If trained from pre-existing weights

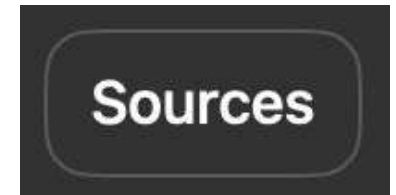
- ✓ Document changes (fine-tuning data)
- 🖨️ GPAI only if trained for  $> 1/3$  of original model compute

# Language Models Ambition

Multimodal  
Aligned LLM



Voice  
input/output



RAG

Sovereign Norwegian AI and the European Legal Landscape  
Språkteknologi, rettsstat og demokrati – Oslo, Feb 2026

# Takk!

## Questions?



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

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