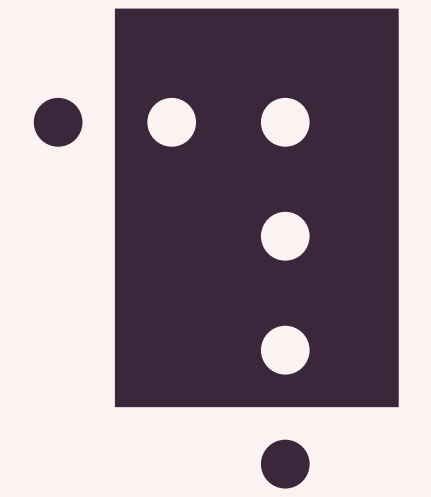


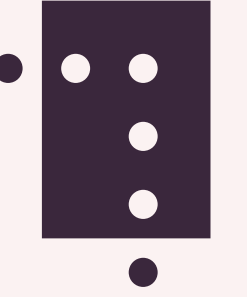
SEMINAR AT UIO, 11.12.25

Training data for LLMs

Yngvil Beyer, head of section, DH-lab and language bank



Nasjonalbiblioteket



We develop and provide training data for

GPT models – large scale language understanding and generation

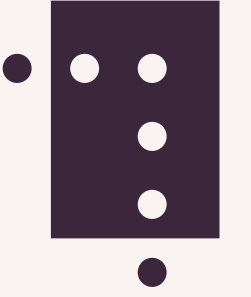
- Large, diverse text corpora
- Learns language patterns, not exact documents
- Goal: generate text

ASR models – automatic speech recognition models

- Audio + transcripts
- Learns speech acoustics and language structure
- Goal: turn speech into text

TTS models – text-to-speech models

- Text aligned with high-quality speech
- Learns prosody, pronunciation, voice characteristics
- Goal: generate natural speech



NB-GPT: open training data

Digitized text from NB collection:

- Public domain

Text sources from Språkbanken:

- Text from different public sector domains and data with various licences

Internet text corpora:

- Filtered to exclude robots.txt-protected sites and known copyright-restricted domains such as Norwegian newspapers

Books (public, OCR)

Digimanus

FinePDFs

FineWeb2

FineWeb2-HQ

FineWeb-Edu

FineWeb-HQ

GiellaLT

HPLT 3.0

Kudos

Lovdata

Newspapers (public, OCR)

Maalfrid (crawl)

Maalfrid (government)

MegaMath

ParlaMint

Project Gutenberg

Saami Web

Swallow Code

FineWiki

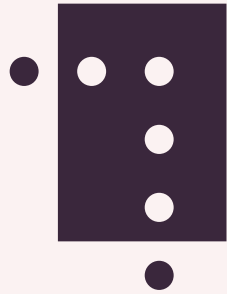
Datasets include

Pre-training data: general language ability

Post-training data: instructions and safety

Evaluation data: measure model quality

Adaptation: domain-specific data





Nasjonalbiblioteket AI Lab

Team

Non-Profit

Verified

https://ai.nb.no/

NbAiLab

Activity Feed

New

Organization settings

Following 250

AI & ML interests

None defined yet.

Recent Activity



Rolv-Arild updated a dataset about 4 hours ago

NbAiLab/aurora-open-2511-long



Yngvil authored a paper 19 days ago

Comparative analysis of optical character recognition method...



Rolv-Arild updated a dataset 20 days ago

NbAiLab/aurora-open-2511

View all activity

Team members 25



NbAiLab's collections 21

Norwegian Language Understanding Datasets

Dictionaries, word datasets, and other semantic and linguistic annotations th...

NbAiLab/orddev-combined

private

Viewer • Updated 16 days ago • 641k • 2

NbAiLab/ordbank-combined

private

Viewer • Updated 15 days ago • 1.96M • 2

NbAiLab/ordbokene-articles-raw

private

Viewer • Updated 13 days ago • 227k • 4

NDLA_NPK

Collection of datasets for a merged NDLA and NPK dataset. devided intoeach ...

NbAiLab/merged_ndla_npk_initial_pairs

private

Viewer • Updated Oct 3 • 1.34M • 1

NbAiLab/merged_ndla_npk_dedup

private

Viewer • Updated Oct 3 • 1.19M

NbAiLab/merged_ndla_npk_dedup_structural

private

Viewer • Updated Oct 6 • 1.15M

NbAiLab/merged_ndla_npk_dedup_semantic

private

Posttrain Datasets

A collections of datasets for the nb-gpt-posttrain group.

NbAiLab/ndla_npk_conversational

private

Viewer • Updated Sep 19 • 2M • 2

NbAiLab/ndla_parallel_paragraphs

Viewer • Updated Oct 16 • 192k • 61

NbAiLab/nynorsk_norm_200eval

Viewer • Updated Oct 14 • 200 • 54 • 1

NbAiLab/ndla_npk_conversational_nb_to_nn

Red-teaming datasets

A collection of red-teaming datasets

NbAiLab/attaq_translated

private

Viewer • Updated about 1 month ago • 1.4k • 3

NbAiLab/aart_translated

private

Viewer • Updated about 1 month ago • 3.27k

NbAiLab/do-not-answer_translated

private

Viewer • Updated about 1 month ago • 939 • 1

NbAiLab/gandalf_ignore_instructions_translated

private

NB-GPT-Eval-lighteval-v0.1

NbAiLab/nb-gpt-eval-lighteval-v0-norcommonsens...

private

Viewer • Updated Jul 4 • 1.09k

NbAiLab/nb-gpt-eval-lighteval-v0-norcommonsens...

private

Viewer • Updated Jul 4 • 272 • 234

NbAiLab/nb-gpt-eval-lighteval-v0-belebele-val

private

Viewer • Updated Jul 4 • 225 • 268

NbAiLab/nb-gpt-eval-lighteval-v0-belebele-test

private

NB-GPT-Eval-lighteval-v0

New splits for the NorEval tasks

NbAiLab/nb-gpt-eval-lighteval-v0-ask-gec-val

private

Viewer • Updated Jul 4 • 40.9k • 26

NbAiLab/nb-gpt-eval-lighteval-v0-belebele-test

private

Viewer • Updated Jul 4 • 900

NbAiLab/nb-gpt-eval-lighteval-v0-ncb-test

private

Viewer • Updated Jul 4 • 840

NbAiLab/nb-gpt-eval-lighteval-v0-ncb-val

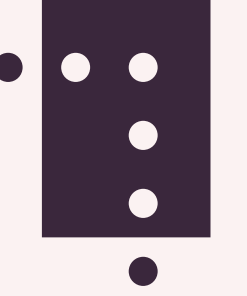
private

Licenced data: Newspapers

- Agreement from January 1st 2026
- 45 million NOK
- 1 year delay
- Opt-out in place
- Can be used for whole life-cycle
- Models to be shared on an open licence

Regjeringen vil styrke
norsk kunstig intelligens
med 45 millioner





Research Access and Model Publication

- Data with open licence will be shared openly
- All models will be shared with open licences
- Data in copyright: need to find ways to share data for training and research (e.g., LUMI and Olivia)