

NORWEGIAN OPEN LLMS AT UIO

David Samuel, Vladislav Mikhailov,
Andrey Kutuzov, Stephan Oepen, Yves Scherrer, Erik Velldal, Lilja Øvrelid
11.12.2025

Language Technology Group (LTG),
Department of Informatics,
Section for Machine Learning,
University of Oslo



1. The importance of openness in the era of generative AI

LLM-as-a-service

Open software and open LLMs

Aspects of openness

LTG and the open LLM ecosystem

2. Web-derived LLM training data for Norwegian

3. Training data for Norwegian LLM research

4. Developing NorMistral-11B

5. NorEval: Native benchmarking for Norwegian LLMs

6. The language quality of the Norwegian output in language models

7. Outlook

For most end users, 'generative AI' means 'chatbot with a web interface or an app'.



For most end users, 'generative AI' means 'chatbot with a web interface or an app'.

This blurs an important distinction

1) large language model (LLM) itself, doing all the work



For most end users, 'generative AI' means 'chatbot with a web interface or an app'.

This blurs an important distinction

- 1) large language model (LLM) itself, doing all the work
- 2) point of access to the model(s): interface



For most end users, 'generative AI' means 'chatbot with a web interface or an app'.

This blurs an important distinction

1) large language model (LLM) itself, doing all the work

2) point of access to the model(s): interface



- What we are dealing with nowadays is usually a **combination of the actual model and a user interface.**
- This allows to 'serve' an LLM or LLMs **as a service**
 - good for business!

For most end users, 'generative AI' means 'chatbot with a web interface or an app'.

This blurs an important distinction

1) large language model (LLM) itself, doing all the work

2) point of access to the model(s): interface



- What we are dealing with nowadays is usually a **combination of the actual model and a user interface**.
- This allows to 'serve' an LLM or LLMs **as a service**
 - good for business!
- **Not so good for public understanding.**
- **The model** is becoming something very far from real life, almost mythical: unlike the inference interface.

For most end users, 'generative AI' means 'chatbot with a web interface or an app'.

This blurs an important distinction

- 1) large language model (LLM) itself, doing all the work
- 2) point of access to the model(s): interface



- What we are dealing with nowadays is usually a **combination of the actual model and a user interface**.
- This allows to 'serve' an LLM or LLMs **as a service**
 - good for business!
- **Not so good for public understanding.**
- **The model** is becoming something very far from real life, almost mythical: unlike the inference interface.

But in fact, **models** are powering 'AI apps', and it is extremely important who owns them.

Open software and open LLMs



Let's recall **open source / free software**.

Open software and open LLMs

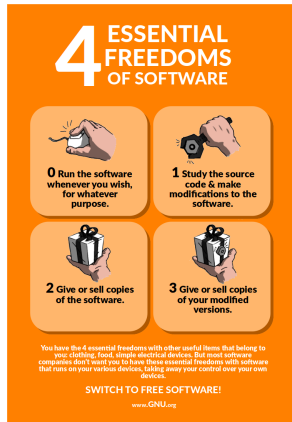


Let's recall **open source / free software**.

Four essential freedoms, according to Free Software Foundation

1. The freedom **to run the program** as you wish, for any purpose.
2. The freedom **to study how the program works, and change it to make it do what you wish**.
3. The freedom **to redistribute copies** so you can help others.
4. The freedom **to distribute copies of your modified versions** to others.

Most of modern computing infrastructure (including Internet) is built on open software. Openly licensed content powers foundational projects like Wikipedia.



Open software and open LLMs



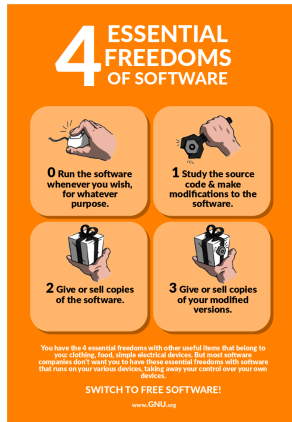
Let's recall **open source / free software**.

Four essential freedoms, according to Free Software Foundation

1. The freedom **to run the program** as you wish, for any purpose.
2. The freedom **to study how the program works, and change it to make it do what you wish**.
3. The freedom **to redistribute copies** so you can help others.
4. The freedom **to distribute copies of your modified versions** to others.

Most of modern computing infrastructure (including Internet) is built on open software. Openly licensed content powers foundational projects like Wikipedia.

What does 'open' and 'closed' mean for **large language models**?



Obviously closed models



- ChatGPT, Gemini, Claude, Grok ≥ 3 etc are entirely closed.
 - No direct access to the models: only via API or web interfaces.
 - Training datasets unknown, no docs or peer-reviewed papers on training

Obviously closed models



- ChatGPT, Gemini, Claude, Grok ≥ 3 etc are entirely closed.
 - No direct access to the models: only via API or web interfaces.
 - Training datasets unknown, no docs or peer-reviewed papers on training
 - Meaningful comparison and studying impossible.

Obviously closed models



- ChatGPT, Gemini, Claude, Grok ≥ 3 etc are entirely closed.
 - No direct access to the models: only via API or web interfaces.
 - Training datasets unknown, no docs or peer-reviewed papers on training
 - Meaningful comparison and studying impossible.
- Roughly equivalent to closed proprietary software like MS Windows...

Obviously closed models



- ChatGPT, Gemini, Claude, Grok ≥ 3 etc are entirely closed.
 - No direct access to the models: only via API or web interfaces.
 - Training datasets unknown, no docs or peer-reviewed papers on training
 - Meaningful comparison and studying impossible.
- Roughly equivalent to closed proprietary software like MS Windows...
 - ...or, rather, Office 365.

Obviously closed models



- ChatGPT, Gemini, Claude, Grok ≥ 3 etc are **entirely closed**.
 - No direct access to the models: only via API or web interfaces.
 - Training datasets unknown, no docs or peer-reviewed papers on training
 - **Meaningful comparison and studying impossible.**
- Roughly equivalent to **closed proprietary software** like MS Windows...
 - ...or, rather, Office 365.

More open alternatives

Some actors advertise their models as 'open source':

- **Llama** by Meta Touvron et al. (2023), **Deepseek** by HighFlyer DeepSeek-AI (2025)...

Obviously closed models



- ChatGPT, Gemini, Claude, Grok ≥ 3 etc are **entirely closed**.
 - No direct access to the models: only via API or web interfaces.
 - Training datasets unknown, no docs or peer-reviewed papers on training
 - **Meaningful comparison and studying impossible.**
- Roughly equivalent to **closed proprietary software** like MS Windows...
 - ...or, rather, Office 365.

More open alternatives

Some actors advertise their models as 'open source':

- **Llama** by Meta Touvron et al. (2023), **Deepseek** by HighFlyer DeepSeek-AI (2025)...
- You can indeed download these models and self-host them. But...

Obviously closed models



- ChatGPT, Gemini, Claude, Grok ≥ 3 etc are **entirely closed**.
 - No direct access to the models: only via API or web interfaces.
 - Training datasets unknown, no docs or peer-reviewed papers on training
 - **Meaningful comparison and studying impossible.**
- Roughly equivalent to **closed proprietary software** like MS Windows...
 - ...or, rather, Office 365.

More open alternatives

Some actors advertise their models as 'open source':

- **Llama** by Meta Touvron et al. (2023), **Deepseek** by HighFlyer DeepSeek-AI (2025)...
- You can indeed download these models and self-host them. But...

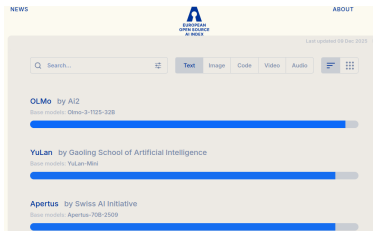
More correct to use the term '**open-weights models**'. For LLMs, **openness** is more complex.

'Openness' of an LLM is a multi-faceted property Liesenfeld et al. (2023)

- are **model weights** available? **intermediate checkpoints**?
- are **training data** available or reproducible?
- is the architecture and **training methodology** transparent?
- are **inference endpoints** easily available?
- what are **licences / terms of use**?

‘Openness’ of an LLM is a multi-faceted property Liesenfeld et al. (2023)

- are **model weights** available? **intermediate checkpoints**?
- are **training data** available or reproducible?
- is the architecture and **training methodology** transparent?
- are **inference endpoints** easily available?
- what are **licences / terms of use**?

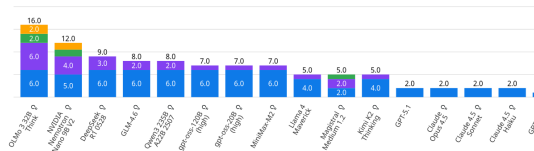


<https://osai-index.eu/the-index?type=text>

Artificial Analysis Openness Index: Components

Openness Index underlying score contribution by components, up to a maximum of 18 (higher is more open)

Model Availability Transparency - Methodology Transparency - Post-training Data Transparency - Pre-training Data



<https://artificialanalysis.ai/?openness=openness-index-bar-chart>

Why we should prefer **open LLMs**?

- They stimulate **progress and common good**:
 - reproducible and transparent research
 - *'everyone can download, use, study and improve'* paradigm leads to a better world

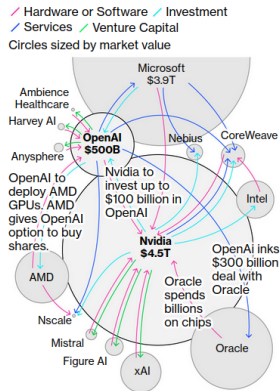
Why we should prefer **open LLMs**?

- They stimulate **progress and common good**:
 - reproducible and transparent research
 - 'everyone can download, use, study and improve' paradigm leads to a better world
- Imagine *a world without Linux, Apache/Nginx and Wikipedia*
 - every web site buys a proprietary OS
 - subscription fees for access to knowledge

Why we should prefer **open LLMs**?

- They stimulate **progress and common good**:
 - reproducible and transparent research
 - 'everyone can download, use, study and improve' paradigm leads to a better world
- Imagine *a world without Linux, Apache/Nginx and Wikipedia*
 - every web site buys a proprietary OS
 - subscription fees for access to knowledge
- **Closed models** owned by BigTech monopolies increase **centralization** of the LLM ecosystem
- they incentivize the 'AI hype': in the end detrimental to the real advances in the field.

How Nvidia and OpenAI Fuel the AI Money Machine



Source: Bloomberg News reporting

Practical risks and harms of closed models

- Concentration of power (mostly USA and China players)
 - OpenAI is a subsidiary of **Microsoft**; Mistral AI linked to **MS Azure Cloud**
 - Anthropic AI is a (kind of) subsidiary of **Google** and **Amazon**
 - Llama models are proprietary products of **Meta**
 - Deepseek models are owned by **High-Flyer**, a large Chinese hedge fund
 - Silo AI is acquired by **AMD** in 2024



Practical risks and harms of closed models

- Concentration of power (mostly USA and China players)
 - OpenAI is a subsidiary of **Microsoft**; Mistral AI linked to **MS Azure Cloud**
 - Anthropic AI is a (kind of) subsidiary of **Google** and **Amazon**
 - Llama models are proprietary products of **Meta**
 - Deepseek models are owned by **High-Flyer**, a large Chinese hedge fund
 - Silo AI is acquired by **AMD** in 2024



Practical risks and harms of closed models

- **Concentration of power (mostly USA and China players)**
 - OpenAI is a subsidiary of **Microsoft**; Mistral AI linked to **MS Azure Cloud**
 - Anthropic AI is a (kind of) subsidiary of **Google** and **Amazon**
 - Llama models are proprietary products of **Meta**
 - Deepseek models are owned by **High-Flyer**, a large Chinese hedge fund
 - Silo AI is acquired by **AMD** in 2024
- **vendor lock-in**
 - you buy subscription from OpenAI and your pipelines are now 'addicted' to it



Practical risks and harms of closed models

- **Concentration of power (mostly USA and China players)**
 - OpenAI is a subsidiary of **Microsoft**; Mistral AI linked to **MS Azure Cloud**
 - Anthropic AI is a (kind of) subsidiary of **Google** and **Amazon**
 - Llama models are proprietary products of **Meta**
 - Deepseek models are owned by **High-Flyer**, a large Chinese hedge fund
 - Silo AI is acquired by **AMD** in 2024
- **vendor lock-in**
 - you buy subscription from OpenAI and your pipelines are now 'addicted' to it
- **private information leaks**: *'a cloud is just someone else's computer'*



Practical risks and harms of closed models

- **Concentration of power (mostly USA and China players)**
 - OpenAI is a subsidiary of **Microsoft**; Mistral AI linked to **MS Azure Cloud**
 - Anthropic AI is a (kind of) subsidiary of **Google** and **Amazon**
 - Llama models are proprietary products of **Meta**
 - Deepseek models are owned by **High-Flyer**, a large Chinese hedge fund
 - Silo AI is acquired by **AMD** in 2024
- **vendor lock-in**
 - you buy subscription from OpenAI and your pipelines are now 'addicted' to it
- **private information leaks**: *'a cloud is just someone else's computer'*
- **lack of reliability**



Practical risks and harms of closed models

- **Concentration of power (mostly USA and China players)**
 - OpenAI is a subsidiary of **Microsoft**; Mistral AI linked to **MS Azure Cloud**
 - Anthropic AI is a (kind of) subsidiary of **Google** and **Amazon**
 - Llama models are proprietary products of **Meta**
 - Deepseek models are owned by **High-Flyer**, a large Chinese hedge fund
 - Silo AI is acquired by **AMD** in 2024
- **vendor lock-in**
 - you buy subscription from OpenAI and your pipelines are now 'addicted' to it
- **private information leaks**: *'a cloud is just someone else's computer'*
- **lack of reliability**
 - a company can promise to 'always provide access to the model'...
 - .. and then deprecate it 'because reasons' (OpenAI did it in the past).



Times are changing



- In the last 1-2 years, **open models** are starting to seriously compete with closed LLMs
- Even BigTech players open up some of their artifacts
 - <https://hf.co/openai/gpt-oss-120b>

Times are changing

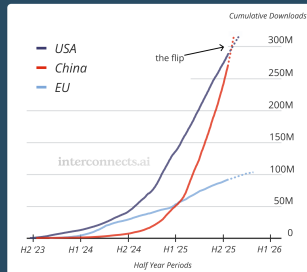


- In the last 1-2 years, **open models** are starting to seriously compete with closed LLMs
- Even BigTech players open up some of their artifacts
 - <https://hf.co/openai/gpt-oss-120b>
- **Chinese projects** are currently taking over the users from the **USA labs**
- **Europe** is (slowly) catching up with projects like HPLT, Tilde, Swiss AI, EuroLLM, OpenEuroLLM and others

Open Source models from China are set to overtake the US, as downloads surge in usage across HuggingFace.

The American ecosystem has stalled in growth, and the “flip” is happening right now.

Interconnects
Nathan Lambert



Times are changing

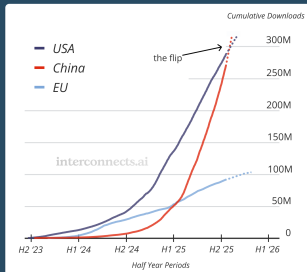


- In the last 1-2 years, **open models** are starting to seriously compete with closed LLMs
- Even BigTech players open up some of their artifacts
 - <https://hf.co/openai/gpt-oss-120b>
- **Chinese projects** are currently taking over the users from the **USA labs**
- **Europe** is (slowly) catching up with projects like HPLT, Tilde, Swiss AI, EuroLLM, OpenEuroLLM and others

Open Source models from China are set to overtake the US, as downloads surge in usage across HuggingFace.

The American ecosystem has stalled in growth, and the “flip” is happening right now.

Interconnects
Nathan Lambert



What is the place of **Norway** in this shifting landscape?

- open datasets for Norwegian NLP (Bokmål and Nynorsk)
- open Norwegian language models
 - 2017–: **word embeddings** for Norwegian and ~70 other languages
 - 2020–: series of transformer-based language models (**NorBERT** and **NorT5**)
 - 2023–: large generative models; **NorBLOOM** and **NorMistral**
- Benchmarking resources and frameworks for evaluation

[illegible][illegible]