Public AI and EU Copyright Law

ECS 10th Annual Conference (Versailles, 2026)

João Pedro Quintais
Associate Professor
University of Amsterdam
Institute for Information Law (IViR)

Ole-Andreas Rognstad
Professor
University of Oslo
Department of Private Law

Outline

- What is “Public AI”?
- EU law and Public AI, especially copyright
- Survey of national Public AI initiatives
- Case studies: Netherlands and Norway
- Discussion and Implications

Outline

- **What is “Public AI”?**
- EU law and Public AI, especially copyright
- Survey of national Public AI initiatives
- Case studies: Netherlands and Norway
- Discussion and Implications

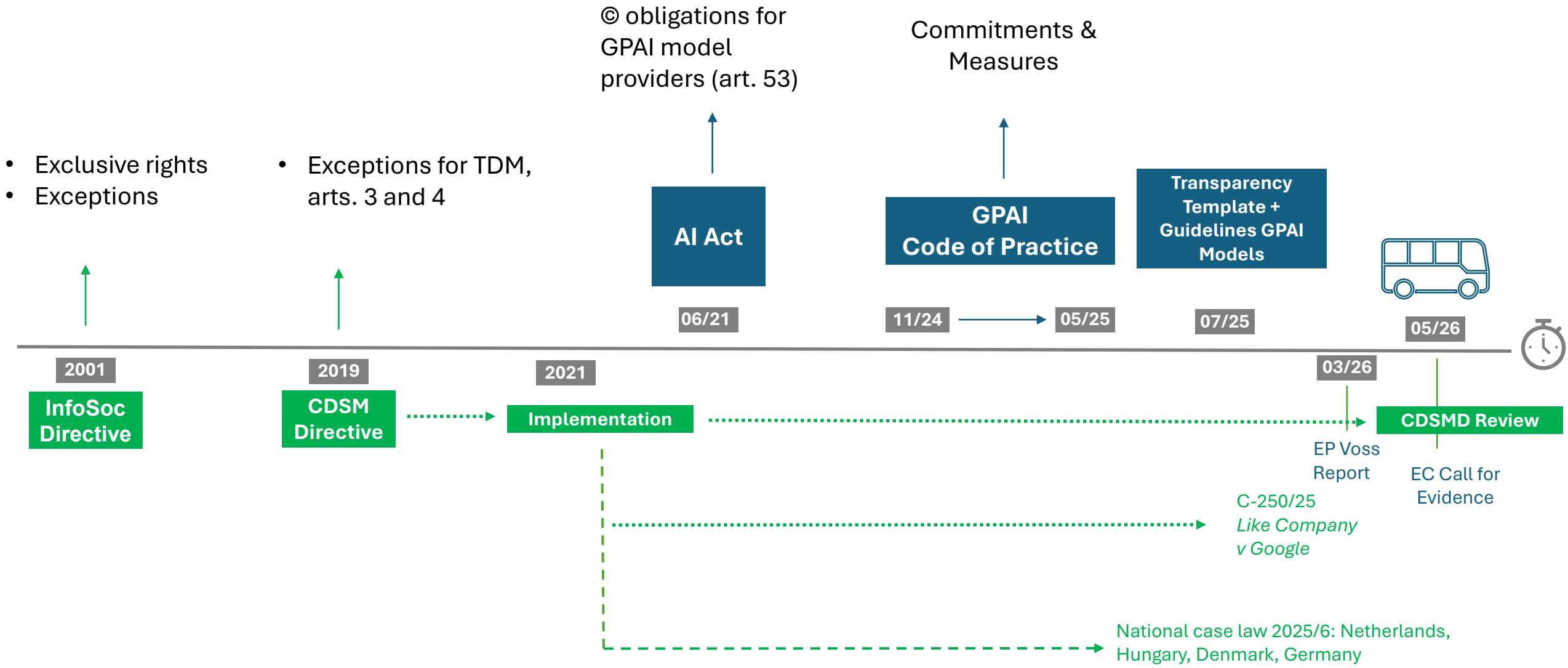
Public AI in European Policy and Literature

- Public AI as alternative to concentrated proprietary AI ecosystems.
- European debates link Public AI to digital sovereignty and public digital infrastructure.
- Core features: openness, public-interest goals and public accountability.
- Policy themes: Open-source models, public compute and shared datasets.
- Public AI remains conceptually broad, ranging from open ecosystems to fully public AI stacks.

Outline

- What is “Public AI”?
- **EU law and Public AI, especially copyright**
- Survey of national Public AI initiatives
- Case studies: Netherlands and Norway
- Discussion and Implications

Legal landscape



Assumptions (not all consensual)

- **AI value chain**: copyright issues re: training, model and outputs

- Most of the *explicit* regulatory focus of harmonized EU law is currently on the **training stage** *lato sensu* (incl. web scraping, data collection and processing, pre-training, fine-tuning, etc.)…
  - But… emerging issues of model memorisation & outputs as infringement (per se or as proxy for memorisation)

- AI Act
  - GenAI = type of GPAI models → Copyright issues (Rs 99, 104)
  - GPAI model providers as primary targets of regulation (Chapter V)

- CDSMD + AI Act logic
  - AI model training involves reproductions & extractions of protected materials
  - Concept of TDM (2(2)) encompasses all, most or significant part of training activities
  - AI model training requires license or exception, predominantly for TDM (arts 3 + 4) [but see InfoSoc (arts 5(1) + 5(3)a)) (R.105)

The EU TDM exceptions

	Art 3 TDM for scientific research	Art 4 E&L for TDM
Nature	**Mandatory**	
Acts	TDM	Reproductions and extractions
Rights	**Mostly reproduction** DB: 5(a), 7(1) / InfoSoc: 2 / CDSMD: 15(1)	**Mostly reproduction** DB: 5(a), 7(1) / InfoSoc: 2 / SW: 4(1)(a)-(b) // CDSMD: 15(1)
Purpose	Scientific research	TDM → **including commercial**
Beneficiaries	Research organisations; Cultural heritage institutions	All (unrestricted)
Lawful Access	YES (no access, no mining)	YES (no access, no mining)
Other Conditions for beneficiary	Works must be stored with "appropriate level of security"	N/A
Contractual Derogation	NO	YES
Other Reservations or exclusions	NO	YES: via "machine-readable means in the case of content made publicly available online"
Other conditions	Three-step test; Partial TPM protection under 6(4) InfoSoc Dir.	Three-step test; Partial TPM protection under 6(4) InfoSoc Dir.
Interface *acquis* E&Ls	InfoSoc: 5(3)(a)	InfoSoc: Art. 5(1)

2 key obligations for <u>all GPAI providers (including FOSS)</u>

Policies to respect EU © Law

- Art. 53(1)(c)
- Put in place a **policy to respect EU ©copyright law in particular to ID + respect**, including through **state-of-the-art technologies**, the reservations of rights [“**opt-out**”] expressed pursuant to art. 4(3) CDSM Directive

Transparency Training Data

- Art. 53(1)(d)
- Draw up and make publicly available a **sufficiently detailed summary** about the **content used for training of the GPAI model**, according to a template provided by the AI Office.

Code of Practice

Transparency Template

Public AI, Research Organisations and FOSS under the AI Act

- AI Act includes special rules for research organisations and FOSS AI projects.
- Articles 2(6) and 2(8) create research exemptions, but:
  - scope remains legally uncertain for Public AI projects with collaborative research, open model releases and public-sector deployment beyond purely academic experimentation.
- “Research organisations” benefit from Article 3 CDSMD TDM exception, but…
  - uncertainty for other participants + non-scientific research parts of projects.
- FOSS GPAI models have exemptions but **not from AI Act copyright obligations**
- Public AI projects face compliance burdens:
  - dataset documentation, rights reservation compliance (if outside Art. 3 CDSMD), transparency reporting & implementation of internal copyright governance systems.

Outline

- What is “Public AI”?
- EU law and Public AI, especially copyright
- **Survey of national Public AI initiatives**
- Case studies: Netherlands and Norway
- Discussion and Implications

National Public AI Initiatives

Country	Project and model name	Governance structure	Public AI characteristics	Copyright operationalisation
Spain	ALIA, model: ALIA	Publicly backed initiative coordinated through the Spanish government and Barcelona Supercomputing Center	Public compute, multilingual public-interest model development, partial open-source orientation	Reliance on publicly curated datasets, transparency documentation and public compute governance
Spain	AINA, model: AINA	Catalan-language initiative coordinated by the Government of Catalonia and Barcelona Supercomputing Center	Linguistic diversity, regional public infrastructure and open-language resources	Use of curated public-language corpora and research-oriented data governance
Netherlands	GPT-NL, model: GPT-NL Model	Publicly funded research project coordinated by TNO with SURF and the Netherlands Forensic Institute	Dutch-language public-interest model and research infrastructure serving the national public interest	Licensed or otherwise lawful content, Content Contributor governance, revenue-sharing mechanism and documented training content protocol
Germany	Aleph Alpha, model: Luminous	Privately organised company with substantial public support	European sovereignty focus, explainability and regulatory compliance orientation	Documentation practices, controlled dataset governance and AI Act compliance positioning
Germany	Teuken, model: Teuken	Publicly supported multilingual research model initiative	European multilingual open model development and research collaboration	Reliance on research datasets, transparency documentation and open research practices
France	Mistral AI, model family: Mistral	Private company benefiting from French and European policy support	Open-weight release strategies and European ecosystem positioning	Open-weight licensing combined with selective dataset transparency practices
Poland	PLLuM, model: PLLuM	Publicly backed Polish large language model initiative	Public-sector deployment orientation and national language preservation	Research-oriented data governance and publicly curated Polish-language datasets
Poland	Bielik, model: Bielik	Open-weight Polish language model initiative supported by public and community actors	Open-source ecosystem development and national language capability building	Open licensing and community-based dataset curation practices
Sweden	GPT-SW3, model: GPT-SW3	Public research collaboration led by AI Sweden	Nordic language preservation and public-sector deployment focus	Research-oriented dataset governance and reliance on public research infrastructures
Finland	Poro, model: Poro	Public-private Nordic collaboration involving research institutions	Multilingual Nordic model development and open-source orientation	Use of research datasets and open-source licensing strategies
Nordic collaboration	Viking, model: Viking	Nordic research collaboration involving universities and public institutions	Regional multilingual infrastructure and open-source ecosystem building	Publicly curated datasets and research-oriented governance practices
European Union	OpenEuroLLM, model family: OpenEuroLLM	European Commission funded consortium	Multilingual open-source European foundation model development	Commitment to open-source licensing, documentation and AI Act compliance
Switzerland	Apertus, model: Apertus	Swiss AI Initiative involving ETH Zurich, EPFL and the Swiss National Supercomputing Centre	Fully open multilingual sovereign AI model developed as public digital infrastructure	Training on openly available data, respect for robots.txt opt-outs, full release of weights, datasets and documentation in line with EU AI Act transparency and copyright compliance principles

Key points

- European governments increasingly support sovereign and public-interest AI models.
- Public AI projects combine public funding, multilingualism and OSS strategies.
- Universities, national libraries and public research institutions play central roles (NB beneficiaries of art. 3 CDSMD).
- Many projects rely on curated, openly licensed or research-based datasets.
- Copyright compliance is operationalised through transparency, documentation and controlled training environments.

Outline

- What is “Public AI”?
- EU law and Public AI, especially copyright
- Survey of national Public AI initiatives
- **Case studies: Netherlands and Norway**
- Discussion and Implications

Case Study 1: The Netherlands GPT-NL

GPT-NL

Contact

Over GPT-NL | Commitments | Samenwerken | Planning | Veelgestelde vragen | Nieuws

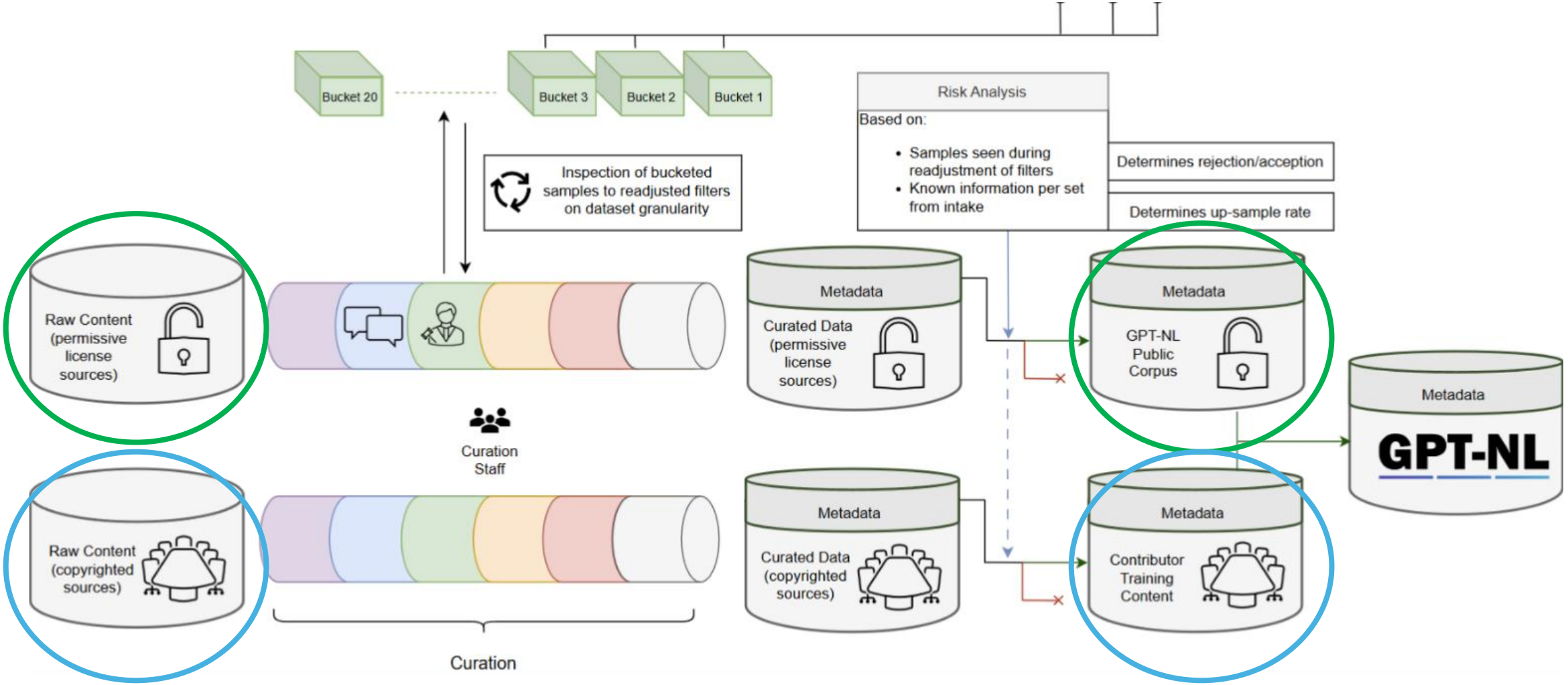
Home

Een verantwoord alternatief op bestaande LLMs

GPT-NL biedt een verantwoord alternatief op bestaande taalmodellen. Wij ontwikkelen een taalmodel met kwalitatieve, Nederlandse data en, gebruiken alleen data als we die rechtmatig hebben verkregen. Daarnaast zijn we transparant over welke trainingsdata we gebruiken, en laten een deel van de opbrengsten terugvloeien naar de auteursrechthebbenden. Op die manier helpen we de digitale positie van Europa versterken, en geven we auteursrechthebbenden een eerlijke plek in de ontwikkeling van technologie.

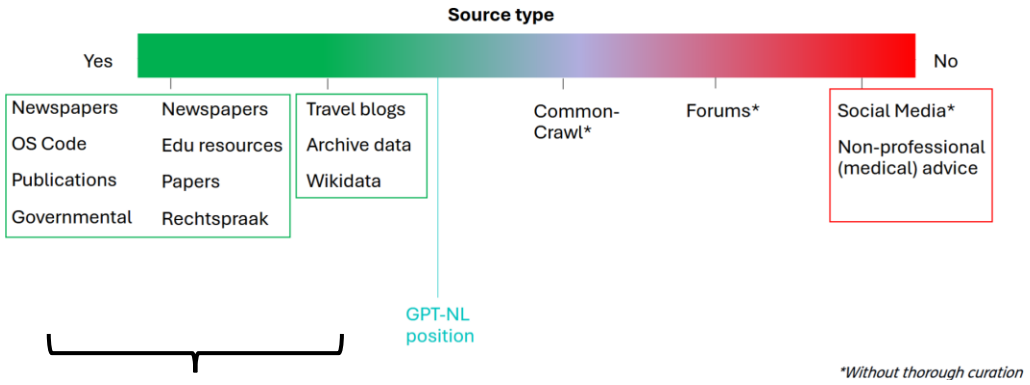
GPT-NL: Copyright-Oriented Public AI Training Model

- Publicly funded Dutch sovereign AI project coordinated by TNO, SURF and the Netherlands Forensic Institute.
- Combines copyright compliance with public-interest and transparency objectives.
- Trains models only on licensed or otherwise lawful content.
- Rejects large-scale indiscriminate web scraping practices used by many commercial AI model providers.
- **Training data** provided through selected publicly available lawful sources + negotiated agreements with “Content Contributors”.
- Content Contributors participate directly in governance, licensing and enforcement decisions.

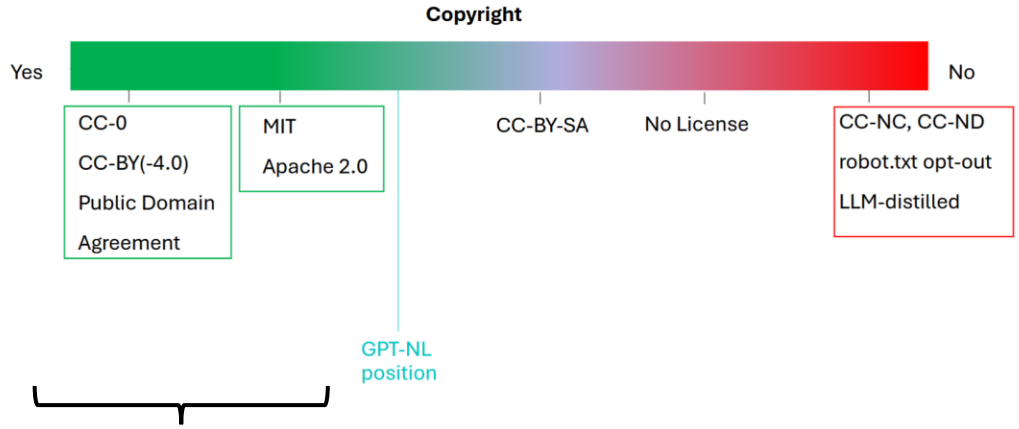


Recipe for Training GPT-NL, available at https://gpt-nl.nl/nieuws/presentatie-dnb-data-science-event-2025/

Data requirements



Data requirements



"Permissive license sources"

Adapted from: *Recipe for Training GPT-NL*, available at https://gpt-nl.nl/nieuws/presentatie-dnb-data-science-event-2025/

GPT-NL: Remuneration, Licensing and Compliance

- GPT-NL is initially funded through a Dutch government research grant.
- Further model development requires additional funding via licences or new public grants.
- TNO must license GPT-NL under market-based conditions to comply with EU and Dutch state aid rules.
- Content Contributors may choose:
  - **(1)** no financial compensation,
  - **(2)** a proportional share of 50% of GPT-NL net revenues plus a discounted professional licence, or
  - **(3)** a one-time upfront payment plus a discounted professional licence.
- Compensation calculated through “data value” methodology assessing the quantity and quality of contributed datasets.
- Licence terms prohibit use of unlawfully obtained or opt-out-protected content.
- GPT-NL operationalises AI Act and copyright compliance through governance, documentation and downstream use restrictions.

Case Study 2: Norway: The Kopinor Agreement with the National Library

The Norwegian Kopinor Agreement(s)

- Agreement between the National Library and the CMO Kopinor on the basis on the General Extended Collective License in the Norwegian Copyright Act (Section 63 second paragraph) of 22 December 2025
- Background: 'The Mimir Project' requested by the Ministry of Culture and carried out by the National Library, University of Oslo and the Norwegian University for Science and Technology, The Mímir Project - Technical Report [Short] (eng)
- September 2025: The Norwegian government granted 45 mill. NOK (approx. 4.190 Euros) for training on newspapers content in the development of LLMs
- The Kopinor Agreement is a result of negotiations on the basis of this grant
- The full Borealis model, based on the Agreement, was launched on 25 May 2026, A new family of Norwegian-centric models – AI-lab

Kulturdepartementet gir 45 millioner til ny norsk språkmodell basert på innhold fra norske aviser

Mediepolitikk, Journalistikk

Publisert 02.09.2025

Fra venstre: direktør i Språkrådet Åse Wetås, nasjonalbibliotekar Aslak Sira Myhre, Geir E. Engen, fagsjef ved Mediebedriftenes landsforening, kultur- og likestillingsminister Lubna Jaffery, digitaliserings- og forvaltningsminister Karianne Oldernes Tung. Foto: Gorm K. Gaare/Nasjonalbiblioteket.

– Utviklingen av generativ kunstig intelligens har til nå vært styrt av globale teknologigiganter, som ofte bruker opphavsrettsbeskyttet innhold uten lov. Derfor er vi veldig glade for at Kulturdepartementet vil sette av penger i neste års statsbudsjett. Det gjør det mulig å utvikle en norsk språkmodell trent på innhold med klarerte rettigheter, sier Randi S. Øgrey adm. dir. i MBL.

Mimir study (July 2024): Main findings

- Models trained with a mix of copyrighted and non-copyrighted content generally exhibited superior performance compared to those trained exclusively on non-copyrighted data
- Particularly valuable: newspapers and non-fictional literature

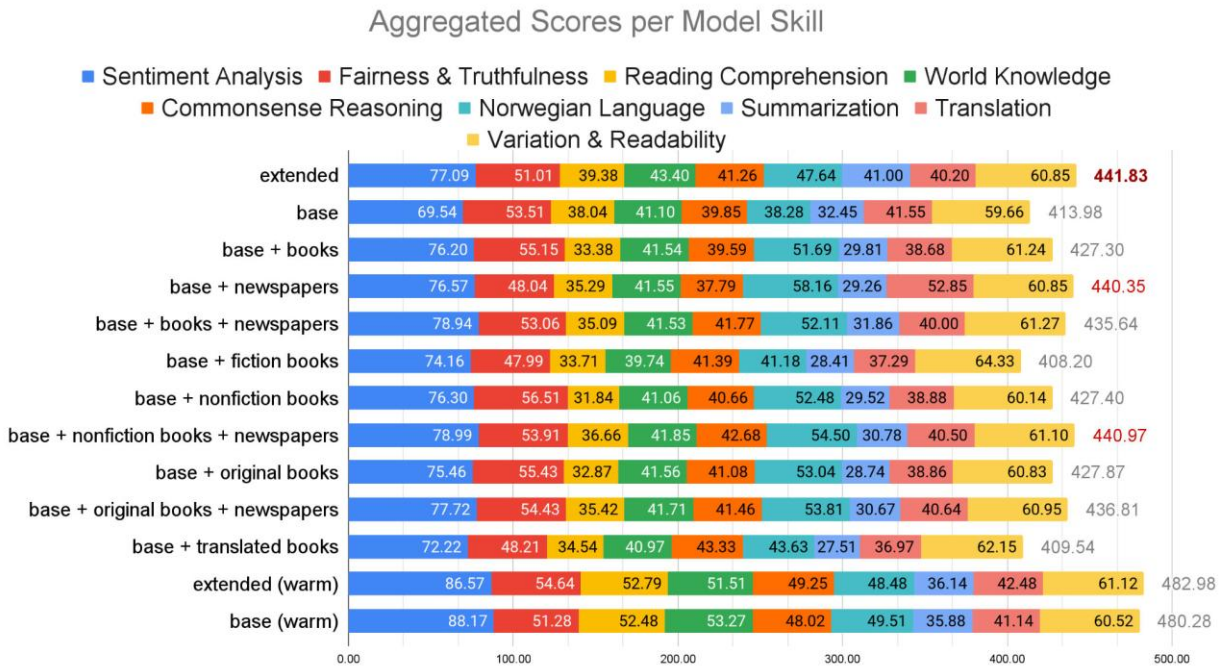


Figure 1. Summary of the total scores (sum) across all skills averaged by task for each model.
Best scores among not warm-started models in red, best overall not warm-started in bold red.

IFRRO member Kopinor signs historic agreement on newspaper content for AI in Norway

IFRRO member Kopinor and the National Library of Norway have signed a historic agreement enabling the use of content from Norwegian newspapers in the training of Norwegian artificial intelligence models. Signed on 19 December 2025, the agreement is funded by the Norwegian state, which will pay NOK 45 million annually. Norway is the first country in the world to establish such an arrangement.

Under the agreement, the National Library of Norway is granted access to content older than one year to develop Norwegian and Sami language models. The trained language models will be made freely available for use by both public and private actors as a basis for AI services. Kopinor concluded the agreement on behalf of the Norwegian Media Businesses' Association and the country's newspapers. The first language models trained under the agreement are expected to be launched in early 2026.

Read the original announcement from Kopinor here.

Foto: Gorm K. Gaare/Nasjonalbiblioteket

Monday, 22nd December 2025

Purpose of the agreement

- “to permit the establishment of a set of training data consisting of copyright-protected content and for this to be used by the National Library to train, develop, maintain and share Norwegian language models with the public”

- “ensure an openly accessible foundation for public and private service development based on language models that preserve the Norwegian language and culture”

- “ensure that the rights holders’ moral and economic interests are safeguarded and not compromised, and that the rights holders are guaranteed compensation for the use of their material for training purposes”

Scope of the agreement

- “Reproduction and use of text for the National Library’s training of language models from copyright-protected press publications published in Norway up to one year from the date of publication”

- Excludes
  - “Opt-outs” of the agreement (in the ECL sense)
  - Illustrations, photographs, audio, moving images and other visual material
  - Text from trade press publications, magazines and weekly newspapers

Rights under the agreement

- The National Library is granted a **non-exclusive license to** use the material covered by the agreement to **train one or more training corpora** (ie. datasets produced by the National Library consisting of rights holders' press publications for use in training language models)

- The National Library **may share the trained language models under royalty-free, non-exclusive terms of use** (licence) **for third-party use** subject to specific license conditions (an adjusted Apache 2.0 license with use restrictions)

Remuneration

- “An **annual total remuneration of NOK 45,000,000** (in 2026) [over EUR 4 Million] shall be paid for the use of protected material under the agreement. The remuneration is subject to indexation, taking effect for the first time in respect of the 2027 remuneration, and thereafter annually on 31 December in accordance with changes in the consumer price index or any other consumer price index that may replace it.”

- (Dependent on government grants)

A new family of Norwegian-centric models

May 26, 2026.

Today we are happy to announce Borealis, a new family of Norwegian-centric instruction-tuned language models from the National Library of Norway.

Borealis is released in five sizes, from **270M** to **27B** parameters, with both full-release and open-release variants. The models are based on the Gemma 3 family and are tuned for Norwegian, Bokmal, Nynorsk, and English assistant use cases, including writing, summarization, question answering, and language-quality assessment.

You can try the demo here.

Why Borealis?

The goal of Borealis is simple: make useful Norwegian language models available in several practical sizes, with transparent documentation, reproducible model artifacts, and formats that people can actually run.

This release also marks an important step for lawful Norwegian language-model development. The full Borealis models are the first Borealis release to incorporate a small amount of data made available through the agreement between rights-holder organizations in Norway and the Norwegian government. So far, we use only a limited supervised fine-tuning subset from this material: around **10,000 tasks** for title and ingress generation.

The open models do not include material from that agreement. Their SFT dataset is available as NbAiLab/aurora-sft-open. The full models use NbAiLab/aurora-sft, whose only difference from the open dataset is the addition of those 10k newspaper-derived tasks.

Models

All models are available on Hugging Face. GGUF repositories are available for llama.cpp, Ollama, and other local inference tools.

The **full models** are released under NB-license, an adaptation of Apache 2.0 with additional use-based restrictions related to training-data recreation and end-user access to licensed press publications. The **open models** are released under the Gemma license and do not include material from the press-publication agreement.

The release collection is here: NbAiLab/borealis.

New agreements?

- Agreement between National Library and Kopinor on training on books (April 2026)

- Similar conditions as the ‘newspaper agreement’ but more extensive limitations (eg. Excludes books from the last 10 years and post 1980 school textbooks)

- Ongoing negotiations with the Ministry of Culture about the funding

Legal context (1/2)

- Neither the CDSMD nor the AI Act are yet in force for Norway (due to particularities of the EEA Agreement)

- A bill exists (27 March 2026) that proposes to implement CDSMD, incl. Article 3 and 4 more or less as in the Directive

- For now: The following rule in the Copyright Regulation (Art. 4) applies
  - The National Library may produce copies of copyright works in its collections as the basis for linguistic corpora for research purposes, also in formats other than the original copy. This also applies to works covered by the Legal Deposit Act.

Legal context (2/2)

- See the statement of the MoC in the 27 March Bill:

  - “where a cultural heritage institution or research institution wishes to make the result of a TDM process available to the public, or if the process is part of the development of services directed to the general public (eg. an AI tool), this will fall outside the scope of the provision aimed to implement Article 3 DSM so that the institutions will have to rely on the Article 4 provision instead in these situations, including the right holders’ right to make reservations against reproductions”.

- Is this compatible with the Directive?

Outline

- What is “Public AI”?
- EU law and Public AI, especially copyright
- Survey of national Public AI initiatives
- Case studies: Netherlands and Norway
- **Discussion and Implications**

Public AI Initiatives in Europe: Key Observations

- Different approaches to the development of public AI across the EU/EEA, including the legal basis for the use of copyright-protected material in such development.
- The Norwegian licensing model has not yet been adopted in other Nordic countries.
- The survey of national public AI initiatives reveals persistent legal and operational uncertainties.
- But emerging EU framework has structural advantages:
  - Public institutions benefit from stronger governance legitimacy, closer integration with Article 3 research infrastructures and greater willingness to adopt transparency-oriented practices

Uncertainties

- Unclear relationship between open-source release strategies and copyright compliance, including whether (most, all, some) model development activities qualify as TDM under Article 3 CDSM Directive.
- Governance fragmentation: complex determination of which entity qualifies as the “provider” under the AI Act.
- Scalability and compliance costs: AI Act documentation, transparency, and copyright governance obligations may require sophisticated compliance infrastructures.
- Persistent uncertainties regarding the interpretation of Articles 3 and 4 CDSM Directive.
  - Underlying normative question: What do we want out of Article 3? What sharing of TDM results is admissible?

Licensing models vs effectiveness of Article 3

- Caused by legal uncertainties or the notion, in particular among public authorities, that right holders should be entitled to some control over or compensation for use of their works in AI training?

- If uses covered by Article 3 CDSM are subject to licensing…
  - Is that compatible with EU law? Effectiveness of Article 3?
    - For use within scope of exception, authorisation should have no legal effect (see e.g. *Stichting de Thuiskopie* C-496/24, paras 57-58, re: private copying)
  - Possible solution: deduction of license fee?

- Do different licensing approaches in national Public AI projects represent a fragmentation of the DSM?

Public AI and EU Copyright Law

ECS 10th Annual Conference (Versailles, 2026)

João Pedro Quintais
Associate Professor
University of Amsterdam
Institute for Information Law (IViR)

Ole-Andreas Rognstad
Professor
University of Oslo
Department of Private Law