COMMUNIA

STUDY

Copyright challenges in open-source AI development in the European Union

JUNE 2026[1]

[1] This study was prepared by Maria Drabczyk (Centrum Cyfrowe) and Alek Tarkowski (Open Future), with contribution from Maja Bogataj Jančič (Intellectual Property Institute).

Table of contents

Executive abstract

The European Union faces a misalignment between its pursuit of digital sovereignty based on open source AI development, and its existing legal frameworks. European research consortia, supported by significant public funding, are building open-source, public-interest Large Language Models. Yet a web of copyright restrictions and legal uncertainties constrains the development of an open source AI ecosystem in Europe.

The Copyright in the Digital Single Market (CDSMD) establishes two mandatory exceptions for text and data mining: Article 3, which allows research organisations and heritage institutions to carry out TDM for scientific research, and Article 4, which allows anyone to carry out general purpose TDM but includes an opt-out mechanism. This distinction introduces profound friction, as AI training projects conducted by research institutions routinely shift away from training under Article 3 and opt for the more restrictive Article 4.

Based on eight in-depth interviews with technical leads, principal investigators, and legal or data experts from European initiatives—OpenEuroLLM, Pleias, GAMS, PLLUM, SOOFI, GPT-NL, and an unnamed repository of climate change publications—this study empirically maps the landscape of legal challenges surrounding open-source AI development.

Unclear rules for training under the TDM exceptions are the biggest legal challenge for training open-source LLMs. Related to this, compliance with opt-out requirements is a challenge, made more complex by a lack of standard, machine readable information on rights reservations. Data sharing is also not harmonised across the EU, leading to a waste of resources, as researchers have to each work with crawl data on their own.

To secure the future of European open-source AI, the authors recommend that EU law explicitly clarify that the training and development of AI systems constitute legitimate TDM activities protected under Articles 3 and 4; that a statutory right to share data for scientific research purposes be introduced; that researchers and public-interest intermediaries acting in good faith be protected from statutory copyright claims; and that Europe build a public training corpus that facilitates training AI under the CDSMD exceptions.

1. Introduction

As artificial intelligence becomes part of the foundational infrastructure of the digital economy, the European Union (EU) faces a misalignment between its pursuit of "digital sovereignty" and its existing legal frameworks. European research consortia, supported by significant public funding, are building open-source, public-interest, and sovereign Large Language Models (LLMs) designed to be transparent, linguistically representative, and accessible to local businesses and the public sector. However, a web of copyright restrictions and legal uncertainties constrain the development of this open ecosystem.

1.1. Goals

The ambition of this study is to empirically map the landscape of legal challenges surrounding open-source AI development by uncovering the breadth of structural obstacles and identifying key, repeating problems across the sector. The focus of this study is on non-commercial research organisations (including those engaged in public-private partnerships) and on the Text and Data Mining (TDM) exceptions provided in Articles 3 and 4 of the Copyright in the Digital Single Market Directive (CDSMD). We seek to evaluate why these public-interest teams frequently decide not to rely on scientific research exceptions. We examine potential fears of copyright litigation, and risk-mitigation strategies that are being deployed in the face of uncertainty.

Ultimately, this research aims to translate specific organisational choices and regulatory expectations into a cohesive, evidence-based narrative that articulates exactly why public-interest exceptions matter in practice. In doing so, it provides policymakers with the necessary data and arguments to defend the scope of Article 3 CDSMD and structurally strengthen the copyright framework for scientific research, in order to protect the future of European open science.

1.2. Context

The CDSMD is the primary EU copyright regulatory framework that is relevant for the development of the open-source AI ecosystem. It establishes two mandatory exceptions for text and data mining (TDM). The first one, under Article 3, allows research organisations and heritage institutions to carry out TDM activities for the purposes of scientific research, using works and other subject matter to which they have lawful access. The exception also allows for retaining works for the purposes of scientific research, including verification of research results. The second exception, under Article 4, allows anyone (most prominently, commercial companies) to carry out general purpose TDM activities, but includes an opt-out mechanism that allows rights holders to reserve the right to use the work.

This distinction between the two exceptions is introducing profound friction for those building open-source, public-interest AI models and systems.

Publicly funded projects are mandated to publish open-source AI models. While there is significant uncertainty regarding what that entails precisely, there is agreement that at least model weights need to be shared openly. This means that the models can be freely downloaded and used, also for commercial purposes. For this reason, developers are routinely preferring to rely on Article 4. Under this regime, public-interest teams exhaust limited resources (both financial and compute), as they deal with an unstandardised pool of machine-readable opt-outs—or face the risk of significant legal liabilities.

This study aims to move past theoretical debates on the challenges to the development of AI models, and in particular model pre-training, under EU's copyright regimes and provide empirical evidence of these barriers based on the direct, practical experiences of public and open-source European AI developers. In the research process we tried to systematically trace the AI development lifecycle, covering training data acquisition, opt-out compliance, licensing constraints, and the strategic effects of legal uncertainty on technical architecture and more broadly, innovation.

Ultimately, this report aims to articulate exactly why public-interest exceptions matter in practice. By delivering concrete evidence and arguments it supports defending—and potentially strengthening—the scope of scientific research exceptions to protect the future of European open innovation.

1.3. Methodology

To examine the practical intersection of copyright law and public-interest artificial intelligence development, we employed a qualitative, empirical research design. We utilised an in-depth semi-structured interview methodology to capture nuanced operational insights. This approach provided us with a framework ensuring consistency across core themes while simultaneously allowing the flexibility needed to explore unique institutional barriers, national copyright variations, and technical workarounds discovered by different development teams across the European Union.

Participants were selected using a purposeful sampling strategy to target expert stakeholders directly engaged in building open-source AI systems (or their components) within the European AI ecosystem. Most of the projects represented in the study are committed to open-sourcing the various components of AI systems that they are building, to the broadest extent possible.

The target sample consisted of representatives from public AI initiatives, academic consortia, and foundational open-source European AI development projects. A total of eight in-depth interviews were conducted, mapping across a diverse geographical and institutional matrix to capture variations in national implementations of EU copyright exceptions that benefit "research organisations" including these under the EU

Copyright in the Digital Single Market Directive (CDSMD). This report presents the findings from interviews with technical leads, principal investigators, and legal or data experts from European initiatives developing open-source AI systems or their components. The final pool of expert interviewees included:

OpenEuroLLM is a European consortium developing a series of strong, multilingual AI models for official European languages. Three choices made by the consortium have significant impact on legal aspects of the project. First, a commitment to publish everything openly, including models, parameters, guidelines, software and data. Secondly, the need to make the produced model available for commercial use – a precondition of Digital Europe funding. And, third, a decision to train the model solely on publicly available web crawl data.

Pleias is a French company that adopted an Open Data approach in their work with Common Corpus—a training dataset for language models, and has built open-source models on its basis. They define their models as fully open, with open training data, training loop and weights. Their goal is to demonstrate that a fully open model is viable, even if not as capable as proprietary ones. They see the need of creating a precedent of fully open models, even if they are not as omnipotent as the models based on proprietary data.

GAMS is a Slovenian initiative building a family of small language models, and exploring state of the art open-source models for less-resourced languages. The project is committed to publishing all model components, including parameters, code and pipelines, under an open-source license. Opening up models is meant to help communities get to grips with high-performance, state-of-the-art technologies. Open sourcing technologies funded with public money is required, based on the Slovenian Scientific Research and Innovation Activities Act. It is also considered a moral obligation by the research team.

PLLUM is a family of Polish open-source LLMs developed by a consortium of Polish public research institutions and commissioned by the Ministry of Digital Affairs, which owns the intellectual property rights to the model. The main aim was to create models that could be used not only by researchers, but by society in the broadest sense: civil servants, institutional representatives, businesses of all sizes.

SOOFI is a German consortium, led by the KI Bundesverband (German AI Association) aimed at building open-source foundation models, with a focus on industry uses. The project is publicly funded by the Federal Ministry of Economic Affairs. The models are intended to be released under open-source licenses - most probably Apache 2.0 or MIT. The primary focus is on the German language, with some multilingual capacity.

GPT-NL is a Dutch language model built by TNO. The goal of the project is to build a state of the art sovereign model in terms of ensuring openness and transparency, and compliance with copyright law, GDPR, and the AI Act. The approach can be described as building a model that is “slow, sustainable and as good as possible”—targeting

public sector users that expect higher ethical standards and compliance from AI technologies.

The GPT-NL project stands out as the one with the weakest commitment to open sourcing AI technologies. It is committed to openly releasing code and data, but the model itself is not open-sourced. Instead, the model is dual licensed under a free research license, or a paid commercial license. This is due to the need to secure sustainability of further model development, which is meant to be covered by commercial licensing. Furthermore, the project assumes that under Dutch antitrust law, publicly funded AI technologies could not be put in the market for free, as that would hinder market competition.

An **unnamed repository of publications on climate change** is a non-profit project to aggregate grey literature on public interest topics and the environment. It provided complementary evidence to our sample of AI development projects. This data intermediary is aimed at addressing integrity of climate information in AI systems and search engines. The goal of this project is to make civil society climate content more visible, and to allow computational research on climate discourse.

The project, while international, aims to comply with EU copyright law. The project runs its own scraper, with AI-powered tools that discover legal metadata: robots.txt files, terms of use, and document-level copyright claims. Based on this, the system creates a metadata pipeline that flags risky content, and creates a record of legal decisions.

The content will be published as a dataset for RAG uses, and depending on the automated analysis of legal risk, it will be republished either as abstracts or full text. In addition, a dataset for a benchmark on integrity of climate information provided by LLMs will be created.

2. Data strategy and navigating the copyright framework

The different open-source AI development teams adopt various strategies with regard to their data strategy: the types of data that they source and use to train the models. All of the projects in this study are focused on the development of language models. Their data strategies are therefore focused on obtaining large quantities of high quality textual content. Data used for training the models falls into four categories:

- Web crawled data. Typically, existing datasets based on web crawling are used, including Common Crawl dumps and datasets built on their basis, such as Nemotron, FineWeb or HPLT. These datasets largely consist of proprietary,

in-copyright materials, mixed with a small portion of content that is either openly licensed or out-of-copyright.
- Openly licensed data and datasets. These include various sources of Open Data and other openly licensed content, including Public Domain book collections, Wikipedia content, Open Access research publications, or resources shared openly by public institutions. These are licensed under various open licenses or are in the Public Domain.
- In-copyright data obtained through licensing deals and other types of agreement. To a small extent, open-source AI developers in Europe sign licensing deals with various entities – mainly commercial ones – that hold rights to high quality data. This typically includes book publishers or media. These are both non-remunerated and remunerated licensing deals. In Slovenia, LLM builders have reached an agreement with the National Library to use its collection, based both on copyright and legal deposit laws.
- Synthetic data. This includes various types of AI generated data, typically created on the basis of either openly licensed or in-copyright sources. This can entail translation of content or generation of new content, based on existing texts.

The projects can be divided into those that train on in-copyright materials by using datasets with web crawled content, and those that do not. For the majority of projects that rely on web crawled data, the TDM exceptions in Articles 3 and 4 of the CDSMD provide a key legal framework for training AI models. Below, we describe in detail data strategies of OpenEuroLLM, GAMS, PLLUM, SOOFI and the climate change publications repository, all of which rely on CDSM exceptions for model training.

This is followed by two examples of alternative approaches, adopted by Pleias and GPT-NL. Pleias has been training models solely on openly licensed texts, and GPT-NL has been training models on both openly licensed text and those obtained through bespoke deals. Both initiatives refrain from using web crawled data, and therefore do not rely on copyright exceptions.

OpenEuroLLM

OpenEuroLLM's data pipeline is based solely on existing web crawl data. This is largely due to financial constraints: the project budget does not include funds either for first-party crawling, or licensing of data. The crawl data comes from Common Crawl, the Internet Archive and smaller crawling projects. A mix of various datasets based on crawl data is used, including DCLM, FineWeb, Nemotron CC and HPLT. The last dataset was previously created by some of the OpenEuroLLM partners.

In addition, OpenEuroLLM depends on synthetic data to build a multilingual LLM. For each of the languages other than English, French, Italian, Spanish and German, an additional 100 billion words were needed. This data is obtained by automated translation of the Nemotron CC dataset, a derivative of Common Crawl.

The consortium assumes that model training falls under the TDM exception for scientific research uses (Article 3). Legal ambiguity is most acute around sharing and retention, not around training itself. And despite a lack of opt-out provisions in Article 3, the consortium aims to not train on data that has been opted-out under Article 4 of the Directive. This is made easier by the fact that Common Crawl follows opt-out information, based on robots.txt files.

The consortium assumes that the datasets they work with have been properly sanitised. For some of the data in the HPLT dataset that predates the introduction of the exception, the consortium went back through crawl packages and removed domains with robots.txt denials, to comply with opt-out rules.

GAMS

The project is based on continued pre-training, an approach where a base model trained predominantly on English continues to be trained on linguistic, in this case Slovene, content. The project uses a variety of sources, including both web crawl data and a corpus of Public Domain academic works and public documents. For web crawl data and other in-copyright sources, the project relies on the scientific research TDM exception (Article 3).

The project also collaborates with the Slovenian National Library, which has initially provided them with a corpus of non-copyrighted works from their collections. The project has also been negotiating access to in-copyright materials and, after two years, an agreement was reached in June 2026, making the library collection that is available in digitised form available for LLM development.

The team has also negotiated with a range of publishers, and some have agreed to share data for the purpose of training GAMS. Participation in the project, through licensing deals, is seen by some Slovenian stakeholders as beneficial to their public image. The Slovenian Press Agency has also provided access to its materials, in exchange for technological support.

PLLUM

The PLLUM project has relied on a range of sources for model training, including web crawl data, openly licensed and Public Domain resources, as well as licensed materials.

The licensed materials accounted for a small share of training data, but constituted high-quality content that is particularly valuable. These resources were obtained directly from publishers and other rightsholders, through negotiations. This required a dedicated team whose job was to establish relationships, and negotiate the licence agreements.

> “This approach is relatively rare in model development globally—the idea of actually going to a publisher and signing a contract”.

For web crawl content, PLLUM could have relied on the scientific research TDM exception (Article 3), as all consortium partners are research institutions. Yet a decision was made to rely instead on the general purpose TDM exception (Article 4). The interviewee indicated that the reason for this was that the open-sourced model was meant to also be used commercially.

SOOFI

SOOFI data strategy is a mixed approach, using multiple data sources. These include web crawled data, openly licensed datasets and some licensed content. The project is closely collaborating with NVIDIA and using the Nemotron datasets. No web scraping activities are conducted, and raw Common Crawl dumps are not used - the project relies on datasets that use cleaned and pre-processed web crawl data.

The project initially relied on Article 3, but over time shifted to Article 4 as the legal basis. This is due to uncertainty whether open-source AI model development can rely on the scientific research TDM exception—if the intention is to publish the model as open-source with commercial use allowed. Our interviewee has signalled significant legal uncertainty around whether and how the TDM exceptions apply to AI training. Due to this, the project has been investing extensively in copyright policies and extensive checks, even of individual data sources.

Repository of climate change publications

The project partners are research organisations that can rely on the Article 3 scientific research TDM exception to conduct their own web crawls, focused on obtaining publications related to climate and environmental policies. However, for reputational management, the organisations decided to proactively respect opt-outs, even though they understand that this is not strictly required of them.

> “We don't want someone to decide we don't qualify for the exception and force us to reevaluate everything”. (IDI)

The project considered licensing deals, but decided against them as an approach that does not comply with the nonprofit status of the host organisation, and that would necessitate significant additional funding.

An assumption is made that organisations whose content will be used for the project—mainly environmental non-profits—have the intent to broadly distribute content to further public interest goals. Thus, there is a low risk of copyright liability.

Pleias

Pleias builds models based on the belief that open, traceable data is a robust training dataset that is legally compliant in any IP regime. They see this approach as a form of regulatory hedging, in a regulatory landscape that continues to be in shift.

> “The regulatory landscape might change. It's not crystal clear which data you can use for training and which you can't. I'm pretty sure the laws will change, one way or another." (IDI)

Provenance is another key feature of open training data, which ensures greater trust in AI systems. This is very different from models built with classical pre-training datasets, largely built through scraping—where provenance is unclear.

An Open Data approach means that a project does not need to invest resources in copyright vetting. They see Open Data as offering clear cut training conditions, and therefore quality assurance can be limited to deduplication, filtering etc.

GPT-NL

The GPT-NL model was trained solely on data that was either in the Public Domain or licensed. The first category consists of a public corpus created mainly from government data. This entailed aggregating data from across the public sector, and in some cases even digitising content. The public corpus also consisted of established, openly licensed collections such as Project Gutenberg or OpenAlex, and Public Domain materials in Dutch and English.

A small amount of synthetic data was used to enlarge the training dataset—this was limited to narrow generation of new sentence forms from existing knowledge graphs, and to translation of texts from other languages.

The private corpus, in turn, is licensed content provided by the Content Board: a group of data providers in the Netherlands with whom GPT-NL collectively agreed on the terms and conditions of data sharing. A revenue sharing mechanism was introduced, with revenue proportional to how much data a party contributed and how much was actually used. This approach attracted various data providers, and therefore provided more high-quality, curated data for the project.

Web crawl data was used to a minimal extent: Common Crawl datasets were used to extract content that explicitly carries a CC0 or CC BY license declaration. In addition, false positives were manually checked for: mainly cases where openly licensed images were included in pages where the text itself was not openly licensed.

3. Key challenges: the cost of uncertainty

According to our interviewees, unclear rules for training under the TDM exceptions are the biggest legal challenge for training open-source LLMs. More generally, fragmented legal regimes, lack of clear rules and complexity of obtaining data means that not enough data can be obtained to build models, especially for less-resourced languages. And the complexity of Member State level copyright regimes creates a challenge for European consortia aiming to build LLMs, with partners located in various Member States, and falling under various legal regimes.

3.1. Uncertainty about training under Article 3 and Article 4 TDM exceptions

> "The fact that we are a scientific institution didn't give us the right to operate solely under Article 3, because purpose matters". (IDI)

All of the interviewed projects are run by research institutions that fall under the Article 3 scientific research TDM exception. Yet, save for one of the interviewed projects, all of those who rely on the TDM exceptions to train on in-copyright data have decided, for various reasons, to rely instead on the narrower, general purpose TDM exception in Article 4. This means, in particular, complying with the opt-out requirements. It is not clear to what extent, otherwise, the projects decided to rely on specific provisions of Article 3 or Article 4, for example those on data retention.

The interviewees provide various reasons for this. Typical arguments include reputational issues – complying with opt-outs is considered a good practice in a situation where the right to train under TDM exceptions is considered to be legally untested, and contested in the public debate about copyright and AI training. Another reason is related to the purpose of releasing open-source AI models. According to some of our interviewees, the fact that open-source models can in principle be used also for commercial uses means that the projects need to comply with the narrower Article 4 exception.

3.2. The burden of complying with opt-outs

Compliance with opt-out provisions is a major challenge for the projects that we researched that decided to rely on Article 4. This is mainly due to a lack of standard methods for expressing copyright reservations and the lack of machine readable information on rights reservations. The robots.txt protocol is commonly used to signal use restrictions, but was not designed for this purpose. In other cases, this information

is included only in terms of use of a site. Also, copyright terms included in robots.txt file for a web domain, and those defined for individual documents published within the domain can also be in conflict. Finally, there is a lack of legal clarity whether textual opt-outs need to be taken into account.

> “There's no standard formula – unlike, say, a pharmaceutical disclaimer with a fixed legal form that everyone recognises. Restriction notices appear in all kinds of forms. This raises a key question: which form is sufficient to consider a reservation valid, and which is not?” (IDI)

This means that in order to establish the copyright status of a work, often various information sources need to be consulted. In order to gain legal certainty, research consortia adopt various measures. Some have introduced manual verification of rights reservation statements, even though Article 4 speaks of machine-readable checks. Moreover, copyright information is seen as often conflicting and incomplete. Interviewees cite cases where some publicly available corpora and data collections were not used, due to poor metadata that did not provide sufficient clarity as to the provenance and copyright status of the data.

The compliance burden mainly affects research organisations and smaller AI development labs. And the complexity of running copyright information verification pipelines can be a factor that limits research, especially for more risk averse organisations. Verification of copyright information does not scale easily, without significant resources. One interviewee noted that filtering out works based on textual opt-out information is so compute-intensive, that it would consume a significant part of the compute budget allocated for the whole project, meant to be used for model development.

These compliance burdens are compounded by problems with the underlying datasets themselves. Web crawl sources carry provenance issues that make opt-out compliance retroactively difficult or impossible. One interviewee considered web-crawled datasets, such as Common Crawl, as controversial sources of training data, due to challenges with respecting opt-outs. He cited issues such as impossibility of retroactive filtering in the case of content crawled before the opt-out rules were introduced; loss of opt-out information for data that exists in multiple places on the web, and availability of crawled content that is currently being contested in court. Problems with the quality of copyright information stored by Common Crawl means, according to the interviewee, that the datasets are inherently non-transparent, as provenance data is missing.

> “What you could do in 10 seconds is just use the internet – but we do not want to do it in an unlawful or unethical way, and right now that’s very hard. That's the bind a lot of AI developers are in”. (IDI)

This leads to a difficult choice for model developers. On one hand, web crawl is the only source of data at a scale sufficient to train sufficiently capable models—most of the projects that we researched rely on this data source. On the other hand, using this data can be a source of legal liability. Other options, such as licensing deals, do not scale easily due to financial and time constraints.

3.3. Mindset and lack of overall legal certainty regarding AI development

Challenges are also related to the unwillingness of various institutions to share data for AI training. These are related to the high level of uncertainty about the legal status of such activities. As these uses are still novel, institutions expect additional assurances. For example, the climate change publications repository project has been negotiating with an institution running an EU-funded project, with an obligation to share research outcomes. Still, the project refused to make the data available for model training.

There are also positive examples of how agreements can be reached, to secure access to novel types of content. The Slovenian case should be considered a best practice with regard to supporting LLM development. The National Library agreed to provide access to in-copyright content from their collection based in large part but not only on mandatory legal deposit rules. The National Library conditioned cooperation on the agreement that data sets based on the library collections will be reshared back to the National Library, so that it can further make them available for new research projects.

3.4. Data sharing challenge

Data sharing among research projects and various institutions is a major challenge identified by some of the interviewees. Problems arise also within consortia that consist of multiple partners, located in various Member States, especially when some of them are commercial entities. Data sharing is not harmonised across the EU. Furthermore under the general purpose TDM exception (Article 4), data cannot be kept once the TDM activities end, and there is no verification clause.

One example of this issue is the HPLT project, which used a budget of 6 million Euro to clean up, deduplicate and refine Common Crawl data. This resulted in 50 terabytes of clean, high quality text for model training in 200 languages. Yet there is legal uncertainty whether this can be shared, leading to a waste of resources, as researchers have to each work with crawl data on their own.

Our interviewee notes that sharing is hard to justify both under the original formulation in the Directive, and under most country-level implementations of the TDM exception. It is also unclear whether the verification rules under Article 3 allow content to be shared for the purpose of reviewing or reproducing research—something that in the opinion of our interviewee should be explicitly

allowed. For now, the solution is to work within the bounds of country-level implementations of the scientific research TDM exception that also include sharing rights.

3.5. Data memorisation

Memorisation of data by the model is a potential challenge considered by some of the consortia. This means that while LLMs are transformative, they can also reproduce content in some modes of operation. One interviewee told us that their consortium assumes that there is currently no definitive answer, based on technical knowledge, whether models memorise training data. There is also no method to fine tune AI training in a way that makes them memorise less. The consortium is aware of a tool built by Allen Institute of AI that tests for memorisation, by checking generated content against training datasets. For this to be effective though, a database of opted-out content would be required—and such a database does not exist at the moment.

3.6. Challenges related to openly licensed content

The GPT-NL representative pointed to a specific challenge of using open content licensed with a CC BY SA license, due to the lack of clarity on whether models or their outputs count as derivatives of the training dataset. The uncertainty means that models themselves might potentially need to be released under a Share Alike license. This resulted in a decision not to use Wikipedia content (which is licensed under the CC BY SA licence), even though there is a general assumption that nobody would object to Wikipedia being used. Such a strategic approach meant a significant downgrade to the training dataset.

4. Towards desired solutions: empirical insights on current policy

The empirical testimonies from the study reveal a widening chasm between the political ambition of creating sovereign European AI and the practical understanding of the EU legal framework by the AI developers. EU funding from programs like Digital Europe or Horizon Europe supports the creation of open-source, commercially viable models, aimed to support European SMEs and public infrastructure. Yet the rigid mechanics of the CDSMD, unclear to the developers, seem to be structurally handicapping these exact projects. In particular, both Article 3 and Article 4 TDM exceptions are often misunderstood and therefore fail to provide legal certainty and to enable developers to build European AI models.

The interviewees signaled the following areas that need policy intervention. Currently misunderstood or identified as challenging, these hold a negative impact on unlocking public-interest AI innovation in Europe:

4.1. Reframing the distinction between Article 3 and Article 4 CDSMD around public-interest open-source deployment

The current EU dual-use copyright framework outlined in Articles 3 and 4 of the CDSMD is described as unclear and does not fit for the purpose of development of open-source AI solutions delivered by research institutions. Public-interest consortia—composed at least in part of universities and public research institutes—would still in many cases rather rely on the more restrictive Article 4, instead on Article 3.

Under Article 4, public-interest developers may need to allocate additional resources to address varied and unstandardised opt-out requests. Additionally, they face the legal obligation to delete or remove the commercial utility of their training datasets once the initial research project concludes. This requirement presents challenges for the iterative, cyclical processes necessary for LLM validation and fine-tuning. To solve this issue, according to the respondents, the definition of scientific research within Article 3 should be broader and explicitly encompass any public-interest, open-source, or sovereign AI development funded by public grants, provided that model is open-sourced: model outputs (weights and code) are distributed under open licenses (the definition of open-source AI, provided in the AI Act, could be adopted for this purpose).

> “The project falls under Article 4 (commercial TDM exception) rather than Article 3 (research), because the model is intended to be commercially usable by third parties—a condition built into the project’s design from the outset. Article 4 imposes stricter constraints: no data retention after project ends, no sharing provisions, no verification clause.” (IDI)
>
> “The project falls under Article 4 (commercial use) rather than Article 3 (research) because Digital Europe requires outputs be commercially usable. Academic institutions in consortium technically fall under Article 3 but the project follows Article 4 requirements.” (IDI)

According to the interviewees, the boundary for the exception should shift from who is doing the training (academic versus commercial entity) to how the output is distributed (proprietary/closed versus open-source/public-interest).

4.2. Introducing a statutory right to share

A cornerstone of scientific integrity is peer verification and reproducibility. However, the current TDM exceptions are described as silent or, at the very least, highly ambiguous regarding whether a research consortium can legally host, share, or republish the copyrighted text corpora they used to train a model so that other researchers can audit it for bias, safety, or accuracy. This ambiguity may create institutional risk aversion. University legal departments, fearing copyright infringement suits, tend to routinely advise research labs against publishing full replication datasets. This approach may stall open science, degrade model transparency, and force researchers to take institutional risks when engaging in standard algorithmic evaluation.

The respondents call to strengthen Article 3 by adding a clear, un-waivable public good exception that explicitly permits the cross-border sharing and secure hosting of training datasets exclusively for scientific verification, peer review, and bias benchmarking.

> “I think the main problem with the way the AI act assumes scrapping should work, is that there is not a clear carveout for research organisations to republish their dataset for reproducibility, and I can imagine how a lawyer not familiar with internet law, would warn a lab from trying to reproduce full datasets... Everyone who is doing a project like this right now is asking ‘should we take a risk and expose ourselves to copyright claims or not from actors.’” (IDI 2)

> “Sharing copyrighted training data is problematic under most country implementations of TDM exception... My hope was that this clause [verification] could justify making data available for replication... But lawyers have said it's too weak.” (IDI 1)

Verification of data is also mentioned as an issue that should be clarified by law. The verification clause in Article 3 seems to be ambiguous and has not been tested. It is unclear to the respondents whether it allows content to be shared for the purpose of reviewing or reproducing research, something that should be allowed, in the opinion of our interviewees. Also, an opt-out database, if created, would allow for better training data auditing, including on memorisation of training data by models.

Rather than steering open science toward pre-emptive procedures, the respondents see an opportunity for the EU framework to place greater emphasis on trusting established academic and research organisations to manage data ethically under their public-interest mandates and build norms that support model development:

> “There should be more trust in researchers that they are using data ethically, instead

of just regulation." (IDI)

There is a need for a better support redistribution of content for public interest uses, related to AI development. For example, there is a public interest need to monitor media and news sources for information integrity. Doing this in a transparent way would require republishing content in a public dataset. Currently, doing such work entails managing a high level of copyright risk. Therefore, a broad carveout for public interest and research-based redistribution of content published without commercial intent would be welcomed.

"The ideal state: if content is published with the widest possible distribution in mind, and it's non-commercial, you should have the right to aggregate it into usable datasets. There's a meaningful slice of the internet without advertising, meant for distribution, without access barriers. That slice should just be available." (IDI)

A landscape review conducted by one of the organisations showed that, due to copyright infringement risk, attempts to build chatbots that provide climate data with a high level of information integrity focus just on Public Domain sources: content from Wikipedia and public research institutions. This means that the set of information to train on is very narrow, and lacks local context, news of what's happening "on the ground" or solutions that are being explored.

4.3. Standardising opt-out protocols, increasing the availability of high-quality resources and introducing "good faith" safe harbors

Under Article 4, rights holders can express their TDM reservations through "machine-readable means." There is a need, voiced by our interviewees, to clearly specify what a valid opt-out reservation should look like: formally, technically and substantively. The burden of interpreting this technical chaos falls entirely on open-source developers. Public-interest teams must exhaust their limited grants building manual data curation pipelines, checking incomplete metadata, and second-guessing legal validity. A dedicated protocol would be welcome (instead of relying on previously existing frameworks like robots.txt), as well as guidance documents from EU lawmakers - anything that would leave less room for interpretation. Clear rules on what constitutes research objectives would also be helpful.

Furthermore, as the open web rapidly degrades and "shrinks" due to the scale of opt-outs, access to cultural texts is vanishing, disproportionately damaging developers working on smaller European languages (e.g., Slovenian, Polish, Dutch) that lack deep digital data pools. A need for policies was pointed out that would ensure access to cultural works and heritage materials, as high-quality language resources for model training. This is a matter of not just clear copyright rules, but also

infrastructures and pipelines that ensure good technical quality of digital resources. Availability of such human-created data may be key for model development, especially as web crawls might increasingly include AI-generated content.

> “Copyright and data collection, as we've discussed, is certainly an obstacle. But there's another challenge emerging now that's connected to data collection: an increasing proportion of content on the internet is AI-generated. From a substantive point of view, this is a growing problem. To what extent can we be sure that what we're collecting is actually human-created, and to what extent is it something that came about with only some human involvement? This makes access to cultural works and heritage materials even more valuable—these are high-quality, exemplary language. The problem there is more technical: the source language is excellent, but the digital versions often require significant work to reach good technical quality for model training.” (IDI)

Open-source developers who systematically implement and document their adherence to this uniform technical standard would like to be granted a “good-faith” safe harbor, legally shielding them from statutory copyright claims. One of our interviewees argues that a de facto safe harbor for AI model training is possible within the existing legal framework.

> “The most important thing would be any change that clearly specifies what a valid opt-out reservation should look like: formally, technically, and substantively. The clearer the better. Perhaps a dedicated protocol... I've come to appreciate procedures more than I used to. A procedure is a standard, and following it is itself a demonstration of good faith.” (IDI)

> “Open-source AI for smaller European languages depends on public access to data and computation. Without access to recent, high-quality national-language content, open models will remain weaker than proprietary models developed by large companies... [OpenEuroLLM] is basically doomed for a long time, because it just won't be able to get enough national data... This should be better harmonised across the EU.” (IDI 3)

4.4. Reduction of hidden structural barriers

The existing regulatory environment in Europe for open-data-driven and public-interest AI initiatives is not streamlined and is filled with pre-emptive administrative burdens. The interviewees flag that the current AI development ecosystem imposes on them heavy compliance requirements and "red tape" before their models are even built.

> “Our biggest problems are not about not allowing risks. Our biggest problems are around just the red tape and the bureaucratisation of things before they're even built. It's more an ‘implementation of regulation before the implementation of technology’ approach.” (IDI)

The regulatory framework in the EU is seen as limiting innovation. A shift towards a technology-first approach would be desired. A policy supportive of open-source model training should favor agile implementation and the use of voluntarily shared open resources, and offer clear guidelines on the existing legal framework and its implementation.

These burdens fall hardest on open-source AI developers and public-interest actors, while well-resourced global competitors can absorb legal ambiguity. Pre-emptive compliance operates as a structural barrier that disadvantages European actors.

5. Recommendations from the authors

While the EU aims for digital sovereignty through open-source AI, the fragmented implementation of Text and Data Mining exceptions under Articles 3 and 4 of the Copyright CDSM Directive is actively handicapping European open AI developments. Direct evidence from interviews reveals a risk-averse legal environment that stifles the development of domestic frontier AI.

To secure the future of European open-source AI, this chapter provides actionable policy recommendations, in line with observations and recommendations shared with us by the interviewees. These include strengthening and clarifying legal basis for open-source model training, expanding public-interest data sharing and safeguarding open science research outputs and innovation.

1. To foster AI innovation and legal certainty, EU law should explicitly clarify that the training and development of AI systems constitute legitimate text and data mining (TDM) activities protected under Articles 3 and 4 of the DSM Directive. Furthermore, the EU legislator must clarify that when research organisations and cultural heritage institutions publish model weights, documentation, and other research outputs under open licensing terms, this activity is already fully permitted under Article 3 DSM and does not disqualify the institution from the scope of scientific research and TDM exceptions.
2. EU law should introduce a statutory right to share data for scientific research purposes, i.e. an explicit, un-waivable public good exception ensuring that

scientific research institutions can legally host, share, and republish curated training datasets for peer-review, evaluation, and algorithmic validation.

3. The EU should protect its researchers and public-interest intermediaries acting in good-faith and actively following standard procedural compliance protocols from statutory copyright claims and legal liability hindering their work on open-source AI.
4. To support the development of European LLMs, Europe needs a European public training corpus as digital infrastructure. This would bring into life the data sharing-related provisions of Article 3 and 4 CDSM. The Slovenian case, where the National Library provides access to in-copyright content from their collections, demonstrates the possibility of developing such a public training corpus.

About Centrum Cyfrowe

The Centrum Cyfrowe foundation is a Polish think-and-do tank that cares about the social dimension of technology, especially as it pertains to public life and civil rights in Poland. It acts to make the world more inclusive, more cooperative and more open by changing the way people learn, participate in culture, use the internet and exercise their rights as internet users.

For more information on Centrum Cyfrowe visit our website: https://centrumcyfrowe.pl/en; or contact us at: kontakt@centrumcyfrowe.pl.

About Open Future

Open Future is a European think tank that develops new approaches to an open internet that maximise societal benefits of shared data, knowledge and culture. Our mission is to reimagine and reframe openness by tackling power-related issues and addressing the imbalances it can create. Our primary focus is safeguarding the open internet to maximise the positive impacts of shared data, knowledge, and culture on society. To this end, it develops civil society strategies and policies for cultivating Digital Commons.

For more information on Open Future visit our website: www.openfuture.eu; or contact us at: hello@openfuture.eu.

About COMMUNIA

The COMMUNIA association advocates for policies that expand the Public Domain and increase access to and reuse of culture and knowledge. It acts as a network of like-minded activists, researchers and practitioners based in Europe and the United States who seek to limit the scope of exclusive

copyright to sensible proportions that do not place unnecessary restrictions on access and use.

COMMUNIA is grateful for the financial support of Arcadia, a charitable fund of Lisbet Rausing and Peter Baldwin.

For more information on COMMUNIA visit our website: www.communia-association.org; or contact us at: communia@communia-association.org.

Thanks

We would like to express our gratitude to all the individuals who accepted our invitation to participate in the interviews that supported this publication.

Tools

We used a range of analytical approaches and tools to analyse the research data: from traditional content analysis methodologies, to using LLM-based tools (like HappyScribe, Gemini, and Notion AI).