



# Legal Compliance Baseline

Deliverable 6.1, version 1.0

|                               |                                                             |                                                                                                                          |
|-------------------------------|-------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| <b>Project acronym</b>        | LLMs4EU                                                     |                                                                                                                          |
| <b>Grant agreement number</b> | 101198470                                                   |                                                                                                                          |
| <b>Project website</b>        | www.llms4eu.eu                                              |                                                                                                                          |
|                               |                                                             |                                                                                                                          |
| <b>Deliverable Lead</b>       | ELDA                                                        | Kossay Talmoudi                                                                                                          |
| <b>Contributors</b>           | ALT-EDIC<br>ALT-EDIC<br>TK<br>TK<br>TK<br>INRIA<br>APTUS.AI | Florine Astruc<br>Trine Platou<br>Lex Keukens<br>Nicky Willemsen<br>Dagmar Klein Hesseling<br>Anne Combe<br>Davide Colla |
| <b>Internal reviewers</b>     | ATHENA<br>PWR                                               | Maria Giagkou<br>Przemysław Kazienko                                                                                     |

## History of Changes

| Version | Date     | Changes       |
|---------|----------|---------------|
| 1.0     | 20251205 | First version |

*This document is a public report. However, the information herein is provided as is and no guarantee is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.*



**Co-funded by  
the European Union**

*This project is co-funded by Digital Europe programme  
under Grant Agreement no. 101198470*

## Table of Contents

|                                                                                                                                             |    |
|---------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1. Introduction .....                                                                                                                       | 5  |
| 1.1. Presentation of LLMs4EU .....                                                                                                          | 5  |
| 1.2. Scope of the deliverable .....                                                                                                         | 6  |
| 2. Personal data protection in data collection and LLM fine-tuning .....                                                                    | 7  |
| 2.1. GDPR applicability to LLM fine-tuning datasets .....                                                                                   | 8  |
| 2.2. Legitimate interests as the legal basis in AI model development .....                                                                  | 10 |
| 2.3. Implementation of the data minimisation principle .....                                                                                | 12 |
| 2.4. Adapting safeguards to web scraped data .....                                                                                          | 12 |
| 2.5. Definition of a data retention period .....                                                                                            | 14 |
| 2.6. Reuse of certain personal data .....                                                                                                   | 15 |
| 2.7. Information of the data subjects .....                                                                                                 | 16 |
| 2.8. Data subject rights and their implications for LLM fine-tuning .....                                                                   | 18 |
| 3. Intellectual Property Rights in LLMs4EU .....                                                                                            | 22 |
| 3.1. The intersection of Intellectual Property rights and generative AI: the Copyright Digital Single Market Directive and the AI Act ..... | 23 |
| 3.2. Text and data mining exceptions and limitations .....                                                                                  | 28 |
| 3.3. Licensing strategies for training data .....                                                                                           | 34 |
| 4. Trade secrets .....                                                                                                                      | 36 |
| 4.1. International protection of trade secrets .....                                                                                        | 38 |
| 4.2. Trade Secrets Directive (EU) 2016/943 .....                                                                                            | 39 |
| 4.3. Ensuring trade secrets are protected in the fine-tuning of LLMs .....                                                                  | 41 |
| 5. Aspects related to the Artificial Intelligence Act .....                                                                                 | 45 |
| 5.1. General-Purpose AI models .....                                                                                                        | 46 |
| 5.2. High-Risk AI systems .....                                                                                                             | 52 |
| 5.3. Implementation timeline and evolving framework .....                                                                                   | 56 |
| 6. Conclusion .....                                                                                                                         | 58 |
| 7. Bibliography .....                                                                                                                       | 60 |
| 7.1. Academic references .....                                                                                                              | 60 |
| 7.2. Institutional references .....                                                                                                         | 61 |

## Table of abbreviations

|               |                                                                                           |
|---------------|-------------------------------------------------------------------------------------------|
| AI            | Artificial Intelligence                                                                   |
| ALT-EDIC      | Alliance for Language Technologies - EDIC                                                 |
| CJEU          | Court of Justice of the European Union                                                    |
| CNIL          | Commission nationale de l'informatique et des libertés (French Data Protection Authority) |
| DSM Directive | Directive on Copyright in the Digital Single Market                                       |
| EC            | European Commission                                                                       |
| EDIC          | European Digital Infrastructure Consortium                                                |
| EDPB          | European Data Protection Board                                                            |
| EU            | European Union                                                                            |
| EUIPO         | European Union Intellectual Property Office                                               |
| FLOP          | Floating-point operations                                                                 |
| GDPR          | General Data Protection Regulation                                                        |
| GPAI          | General Purpose Artificial Intelligence                                                   |
| InfoSoc       | Information Society                                                                       |
| IP            | Intellectual Property                                                                     |
| IPR           | Intellectual Property Rights                                                              |
| LLM           | Large Language Model                                                                      |
| OECD          | Organisation for Economic Co-operation and Development                                    |
| TDM           | Text and data mining                                                                      |
| TRIPs         | (Agreement on) Trade-Related Aspects of Intellectual Property Rights                      |
| WTO           | World Trade Organisation                                                                  |

### **Executive summary**

This deliverable provides a legal analysis of data usage for model fine-tuning within the LLMs4EU project, focusing on compliance with applicable European data protection, artificial intelligence and copyright frameworks. It identifies the relevant legal bases for processing personal data under the General Data Protection Regulation, evaluates the permissibility of using different data sources, and examines the limitations imposed by IPRs, including the scope of exceptions and limitations under the EU acquis. The report further highlights practical considerations for lawful data acquisition, risk mitigation, and documentation obligations, ensuring that the fine-tuning process within LLMs4EU adheres to EU legal standards.

# 1. Introduction

## 1.1. Presentation of LLMs4EU

Large Language Models for the European Union (LLMs4EU) is a project funded under the Digital Europe Programme that aims to develop fine-tuned LLMs for various applications while preserving linguistic diversity, regulatory compliance, and European technological sovereignty. The project is one of a series designed to consolidate an EU-centred ecosystem for developing state-of-the-art language technologies and is under the coordination of the Alliance for Language Technologies (ALT-EDIC), established through European Commission Implementing Decision (EU) 2024/458.<sup>1</sup> ALT-EDIC operates as a European Digital Infrastructure Consortium under the Digital Decade Policy Programme 2030<sup>2</sup>, enabling Member States to pool resources efficiently for important digital projects. The LLMs4EU project thus has the underlying ambition of consolidating the ecosystem of language technologies and creating synergies between economic and academic actors.

The LLMs4EU project's approach is to focus on five strategically selected use case domains that address critical economic sectors and public service functions in order to attain fine-tuned effective LLMs that will serve the community. The use cases are:

- Scientific Research
- Public Services
- Tourism
- Telecommunications
- Energy Sector

LLMs4EU intervenes at a moment where the emergence of generative artificial intelligence has fundamentally altered the technological landscape<sup>3</sup>, creating an ecosystem dominated by English-language training data and predominantly non-European technological actors. This dominance poses significant challenges to European linguistic sovereignty and threatens to marginalise Union's official languages and regional and minority languages. This is exacerbated by the fact that European entities face disadvantages in accessing appropriate language resources for AI model development.

---

<sup>1</sup> Commission Implementing Decision (EU) 2024/458 of 1 February 2024 on setting up the European Digital Infrastructure Consortium for the Alliance for Language Technologies (ALT-EDIC) [https://eur-lex.europa.eu/eli/dec\\_impl/2024/458/oj/eng](https://eur-lex.europa.eu/eli/dec_impl/2024/458/oj/eng)

<sup>2</sup> Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030

<sup>3</sup> Please note that the term “generative artificial intelligence” refers to broad set of technologies capable of generating content. This includes image and video generation, speech generation, text generation and many other modalities. However, in this project, we focus solely on LLMs and text generation.

The lack of European language resources for training large language models creates dependencies on external technological solutions that may not align with European regulatory requirements.

It is to be highlighted that the LLMs4EU project operates within a complex and rapidly evolving regulatory environment where the intersection of artificial intelligence development, data protection compliance, and linguistic diversity preservation requires careful legal analysis. WP6 will provide guidelines and orientations to project partners, with the aim of aligning with all relevant regulations, directives and soft law, such as the ethics guidelines for trustworthy AI.<sup>4</sup>

## 1.2. Scope of the deliverable

This deliverable D6.1 “Legal Compliance Baseline” is a building block in the compliance work carried out under the LLMs4EU project. It constitutes the initial output of WP6, which is designed to ensure conformity with the applicable legal frameworks governing data and artificial intelligence within the European Union. This deliverable falls under Task T6.1, whose purpose is to establish a baseline framework summarising and interpreting the relevant legal obligations in relation to data collection, curation, use, and reuse for the fine-tuning and alignment of large language models<sup>5</sup>. The fine-tuning activities carried out in LLMs4EU aim to extend LLMs to specific domains, terminology, and requirements without the cost of building models from scratch.

The primary mission of Task T6.1 is to provide guidance on intellectual property rights clearance, data protection requirements, and confidentiality considerations. In this context, the legal compliance baseline is instrumental in guiding the participants to ensure that datasets collected are appropriately documented with respect to their terms of use and IPR status, adequately cleared (rendering the content usable from an IPR point of view) if/when needed, respecting the interests and rights of third parties, including copyright holders and individual data subjects. It is to be recalled that the scope of the task and relevant deliverables concerns the data collected and processed for the purposes of the project.

Moreover, this deliverable sets out a comprehensive framework for integrating compliance requirements, particularly those derived from the General Data Protection Regulation, but also following the evolving regulatory landscape, including the AI Act. The findings and principles consolidated in this document form the foundation for

<sup>4</sup> European Commission, 2019, Ethics guidelines for trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

<sup>5</sup> Note that the term “fine-tuning” refers in this document to all activities aimed at permanently training the LLM, in particular continuous pre-training, fine-tuning based on instructions (e.g. SFT) and various types of alignment, e.g. using human preferences. They are performed to provide to LLMs new knowledge, communication skills in various European languages, security-related constraints, etc.

subsequent legal evaluations and assessments during the project's life cycle. It also serves as a practical reference that enables the technical teams to meaningfully address legal compliance in their fine-tuning of LLMs.

Accordingly, the deliverable functions as a theoretical reference and starting point for all LLMs4EU partners seeking legal direction from WP6 partners.

The structure of the deliverable is thematic and is meant to allow the navigation between the relevant frameworks. The main topics addressed are data protection, intellectual property rights, confidentiality and trade secrets, and compliance with the AI Act. The scope is strictly centred on "AI models" understood here as large language models, thereby excluding a broader analysis of all generative AI systems.

## **2. Personal data protection in data collection and LLM fine-tuning**

Personal data protection requirements apply throughout the project life cycle, namely during the data collection phase as well as to all data processing activities, including the fine-tuning of large language models for the defined use cases.

It is to be noted that the data protection measures described in this deliverable should be applied by the project partners, as data controllers, together with their own internal data protection measures. However, the legal governance structure established under WP6 is responsible for providing compliance tools and legal support, which ensures the compliance of data processing activities, especially the respect of the rights of data subjects.

Data protection compliance in the context of AI development is central. However, authoritative guidance from Data Protection Authorities and relevant case law remains limited. The Opinion 28/2024 of the European Data Protection Board (EDPB)<sup>6</sup> offers targeted analysis concerning the processing of personal data in AI models and serves as a main interpretative reference for the project's compliance strategy. Moreover, relevant documentation has been produced by the French Data Protection Authority (CNIL). As the project advances, the future WP6 deliverables will reference further evolving guidance from authoritative sources.

---

<sup>6</sup> Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models

## 2.1. GDPR applicability to LLM fine-tuning datasets

### 2.1.1. The applicability of the material and territorial scope of the GDPR

The applicability of the General Data Protection Regulation (GDPR)<sup>7</sup> to LLM fine-tuning datasets in LLMs4EU is determined by the material and territorial scope of the Regulation. Regarding material scope, the GDPR applies to any operation involving the processing of personal data, including collection and use for LLM fine-tuning.

Article 4(1) of the GDPR defines personal data as *“any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;”* Given this broad scope, datasets assembled or used for the fine-tuning of large language models are likely to include personal data. Language data, even when not explicitly collected for the purpose of identifying individuals, frequently contains information that can be linked, directly or indirectly, to natural persons. This may arise through the inclusion of names, email addresses, user IDs, or contextual information that, when combined with other data, enables identification. Furthermore, the aggregation and transformation of language data from diverse sources, including open data and proprietary datasets, magnify the risk that personal data will be present, whether in structured fields or unstructured content. Thus, the collection, aggregation, transformation, and use of datasets for the purpose of LLM fine-tuning fall within the material scope of the GDPR.

The act of collecting data and processing it for the purpose of creating a fine-tuned model constitutes processing of personal data wholly or partly by automated means, as defined in Article 2(1) GDPR.

In regards to the resulting AI model, the European Data Protection Board (EDPB) has answered to the question regarding the “anonymity” of a model and clarified that the mere fact that the AI model is the result of a training process does not render the model anonymous. The assessment of whether data processed during training or retained in the model remains personal data must consider the likelihood of direct or indirect

<sup>7</sup> Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>

identification of individuals, taking into account all means reasonably likely to be used. If the model retains personal data or enables its extraction, the GDPR remains applicable.

Concerning territorial scope, Article 3(1) GDPR provides that the Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or processor, in the sense of Article 4(7) and 4(8) of the GDPR, in the Union, regardless of whether the processing itself takes place within the Union. Where project participants are established in the EU, the establishment criterion is satisfied and the GDPR applies to all related processing activities, including those conducted outside the Union.

### **2.1.2. Responsibilities of the beneficiaries carrying out data processing activities**

When project participants jointly determine the purposes and means of processing personal data for LLM fine-tuning, they are considered to be joint controllers under Article 26 GDPR.<sup>8</sup> In LLMs4EU, these correspond to all project participants that process the personal data either by collecting, cleaning, modifying, using and evaluating the data. Joint controllers must transparently define and allocate their respective responsibilities for compliance, including the facilitation of data subject rights. This arrangement requires a clear agreement between the parties and must be communicated to data subjects, ensuring accountability and effective cooperation in fulfilling GDPR obligations.<sup>9</sup> In LLMs4EU such cooperation on personal data issues is referenced in clause 4.4 of the LLMs4EU Consortium Agreement that allows for further agreements between the partners in regard to personal data processing.

It is also important to note that the usage of personal data regarding the deployment of the model itself may pose different configurations regarding controllership.<sup>10</sup> Controllership in this context needs to be examined while taking into account the determination of means and purposes of such deployment in the project.

---

<sup>8</sup> Article 26 GDPR

<sup>9</sup> See similar analysis in Van Veen EB, Boeckhout M, Schlünder I et al. Joint controllers in large research consortia: a funnel model to distinguish controllers in the sense of the GDPR from other partners in the consortium [version 2; peer review: 3 approved, 1 approved with reservations]. Open Res Europe 2024, 2:80 (<https://doi.org/10.12688/openreseurope.14825.2>)

<sup>10</sup> “The relationship between users and providers of generative AI presents therefore a special case that cannot be seamlessly subsumed under the categories of the GDPR.” Ruschemeier H. Generative AI and data protection. Cambridge Forum on AI: Law and Governance. 2025;1:e

## 2.2. Legitimate interests as the legal basis in AI model development

Work undertaken for the purpose of fine-tuning large language models in LLMs4EU necessarily entails the ingestion of large corpora that will likely contain personal data. All processing must respect the principle of lawfulness provided by Article 5(1)(a) GDPR, which requires that one of the six lawful bases in Article 6(1) be fulfilled, namely: consent; contract; legal obligation; vital interests; public task; or legitimate interests. The European Data Protection Board has expressly acknowledged that “legitimate interests” can constitute a legal basis for AI training provided that controllers satisfy a three-step assessment test consisting of: (i) identifying a legitimate interest, (ii) demonstrating objective necessity, and (iii) conducting a careful balancing of that interest against the rights and freedoms of the data subject.<sup>11</sup>

Article 6(1)(f) GDPR<sup>12</sup> allows processing where “*processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party*,”. Recital 39 clarifies that personal data should be processed only where the purpose cannot reasonably be accomplished through less intrusive means. In this sense, the principle of data minimisation is a criterion in the assessment of the possibility of using legitimate interests in this context. In Opinion 28/2024 the EDPB explicitly recognised the feasibility of relying on Article 6(1)(f) for the training of AI models, applicable here. The EDPB however, highlighted that such reliance demands accountability documentation and ongoing risk management. The opinion also refers to Guidelines 1/2024, adopted on 8 October 2024<sup>13</sup>, that provide the interpretative framework and detail criteria of the three-step test.

In this sense, to determine whether a processing of personal data may be based on Article 6(1)(f) GDPR<sup>14</sup>, three cumulative conditions need to be satisfied: (i) the pursuit of a legitimate interest by the controller or by a third party (identification); (ii) the processing is necessary to pursue the legitimate interest (necessity test); and (iii) the legitimate interest is not overridden by the interests or fundamental rights and freedoms of the data subjects (balancing test).

The first element requires clearly defining the interest relied on.<sup>15</sup> This interest must be lawful, clearly and precisely articulated, real and present. The latter meaning that the interest cannot be speculative. In the case of LLMs4EU, the aim is to develop models in

<sup>11</sup> Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models

<sup>12</sup> Article 6 GDPR

<sup>13</sup> EDPB, Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR

<sup>14</sup> Article 6 GDPR

<sup>15</sup> Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 21

domains that offer tangible societal benefits. However, the legitimacy of each specific use case requires a documented, case-by-case analysis.

The second element consists of assessing necessity.<sup>16</sup> Article 5(1)(c) and Recital 39 GDPR clarify that personal data should only be processed if the intended purpose cannot reasonably be achieved by other means. In the practical context of LLM development, the data to be processed are inherently large and, despite minimisation efforts, cannot entirely exclude personal data. Each LLMs4EU partner must assess the volume of personal data required and actively consider whether less intrusive alternatives could fulfil the project's purposes as effectively. Measures such as pseudonymisation are relevant for meeting the necessity requirement. It is to be noted that the Court of Justice of the European Union has repeatedly insisted on a strict necessity standard. In its judgement of 4 July 2023 in Case C-252/21<sup>17</sup> the Court observed that the controller must demonstrate that processing is "objectively indispensable" for the stated purpose. Although the decision arose in the competition-law context, the reasoning is directly transposable to AI training that may use personal data such as the case of LLMs4EU.

The third element is the balancing test.<sup>18</sup> Controllers must demonstrate that the interest relied on is not overridden by the interests or fundamental rights and freedoms of data subjects. Large-scale data collection by AI systems can create particular risks, such as a perception of surveillance, especially when data is scraped from public sources. Partners are required to consider potential risks to data subjects' privacy, autonomy, and financial or social interests, as well as the prospects for positive societal impacts, such as improved access to information or services. The assessment should include the nature of the data processed, the context in which it is processed, and the possible consequences for data subjects. Special attention must be paid to whether the data include special categories under Article 9 GDPR (such as data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health, data concerning a natural person's sex life or sexual orientation), and whether data subjects are children or otherwise vulnerable persons. Another relevant criterion here is the reasonable expectation of data subjects. Recital 47 GDPR requires assessing whether data subjects can reasonably expect that their data may be processed for the stated purposes at the time and in the context of collection. In the context of LLMs4EU, reasonable expectations can for example be determined. The organisation and clarity of information provided to data subjects are central, including whether the personal data was publicly available, the type of relationship with the controller, the context of collection, the possible further uses,

<sup>16</sup> Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 22

<sup>17</sup> CJEU - C-252/21 - Meta Platforms and Others (General terms of use of a social network)

<sup>18</sup> Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 23

and the level of awareness of data subjects regarding the online availability of their personal data.

## 2.3. Implementation of the data minimisation principle

According to Article 5(1)(c) GDPR, only data necessary for the purpose of processing may be collected or processed.

In this regard, data protection authorities such as the French Data Protection Authority (CNIL) published recommendations for the development of AI systems<sup>19</sup>. In practice, it is necessary to prioritise techniques that achieve the desired result using as little personal data as possible. The choice of the learning protocol used may, for example, allow access to data to be restricted to authorised persons only, or allow access only to encrypted data. This principle does not prevent the use of large training databases. The data used must, in principle, be selected and cleaned in order to optimise the training of the algorithm while avoiding the use of unnecessary personal data. Several steps are recommended to ensure that the data collected is relevant to the objectives pursued:

1. Defining which categories of data are useful before starting collection;
2. Excluding from the collection certain categories of data, by filtering or excluding certain types of sites, particularly sensitive data if its collection is not relevant to the processing carried out.
3. Deleting data accidentally collected that is not relevant.
4. Not collecting data from sites that oppose the web scraping of their content by means of technical protections (such as CAPTCHAs or robots.txt files).

## 2.4. Adapting safeguards to web scraped data

Data scraping has become commonplace in LLM development and is mentioned in LLMs4EU as a means of collecting fine-tuning data. There is currently no generally accepted standard definition, and it can encompass a range of techniques and activities. Data scraping involves the technique of systematically searching sources using software (e.g. scrapers or crawlers) and copying relevant information that is used to train a language model underlying an LLM. Data scraping is commonly referred to as "web scraping" because it mainly involves extracting data from websites. However, it may also include other methods of data access, such as screen scraping. "Web crawling" is also not clearly defined but generally refers to the automatic indexing or other collection of information from websites by systematically searching web pages, often for the purpose of indexing by search engines. Data Protection Authorities such as the CNIL issued

<sup>19</sup> CNIL Synthèse des recommandations pour développement des systèmes d'IA : <https://www.cnil.fr/fr/developpement-des-systemes-dia-les-recommandations-de-la-cnil-pour-respecter-le-rgpd>

guidelines in regard to the protection of personal data contained in scraped data<sup>20</sup>. This is achieved by:

1. Cleaning data in order to build a high-quality learning base and thus strengthen the integrity and relevance of the data by reducing inconsistencies, as well as the cost of learning.
2. Identifying relevant data to ensure that certain categories that are unnecessary for the task at hand are not represented and to identify data that is not relevant for learning (which must then be removed from the database).
3. Implementing measures from the design stage in order to transform the data (such as generalisation and/or randomisation measures, etc.) to limit the impact on individuals.
4. Monitoring and updating data because the data collected may lose its accuracy, relevance, adequacy, and limited nature due to possible data drift, updates, or technological developments. Regular analyses must be conducted to monitor the database that has been created.
5. Documenting the data used for the development of an AI model in order to demonstrate the traceability of the datasets used, which can be difficult due to their large size.

Regarding the reasonable expectations of the data subjects mentioned previously, the following elements must be taken into account: the publicly accessible nature of the data, the nature of the source websites (social networks, online forums, etc.), the type of publication (e.g. an article published on a freely accessible blog or a restricted publication), etc.

Additional measures depending on the intended use of the AI model and on the risks involved in the processing must be implemented.

1. Establishing a default exclusion list of sites containing particularly sensitive data (health forums, etc.).
2. Excluding sites that oppose the web scraping of their content through technical or legal measures (e.g. terms and conditions of use).
3. Limiting collection to freely accessible data (without the need to create an account) for which individuals are aware of its public nature.
4. Informing data subjects as widely as possible (online articles, social media accounts, etc.).
5. Providing for a discretionary and prior right to object, prior to collection, with a reasonable time frame before training a model

---

<sup>20</sup> CNIL, Fiche pratique de l'IA sur les mesures en cas de collecte de données par moissonnage, June 2025, <https://www.cnil.fr/fr/focus-interet-legitime-collecte-par-moissonnage>

6. Anonymising or pseudonymising data immediately after collection and prevent any cross-referencing of data based on individuals' identifiers. Such anonymisation needs to be envisaged while keeping in mind the limitations of the available techniques.<sup>21</sup>

## 2.5. Definition of a data retention period

The storage limitation principle according to Article 5(1)(e) GDPR makes it mandatory to determine a data retention period after which the data must be deleted, anonymized, or in some cases archived. This retention period must be determined by the data controller on the purpose for which the data was collected.

In the context of data collection for AI development, a retention period must be set as well.

**For the fine-tuning phase**, LLMs4EU partners must:

1. Set a data retention period for the use made for the finetuning of the model. During this phase, the provider uses the data for the creation of a database limited to strictly necessary data, cleaned, pre-processed, and ready to be used for finetuning purposes. During this phase, the data must be shared securely and be accessible to authorized persons. Depending on the case, this phase can last from a few weeks to several months, or, conversely, be carried out iteratively in the case of continuous learning. This duration should be defined in advance and justified (taking into account the data controller's past experience, their knowledge of the duration of IT developments, the human and material resources they can make available to carry them out, etc.).
2. Plan in advance the data retention and monitor it over time. This period must correspond to the time from data collection and model finetuning to the testing phase of the final product. The data subjects must be informed of the data retention period (e.g. in information notices).

**Beyond the fine-tuning phase**, when the data is no longer needed for finetuning the AI model, it should in principle be deleted. However, it may be retained for product maintenance or improvement if safeguards are put in place (segregated storage, access restricted to authorised persons only, etc.).

The retention of data can enable audits to be carried out and facilitate the measurement of certain biases. In these cases, and when a similar result could not be achieved by retaining general information about the data (such as documentation produced on the

---

<sup>21</sup> “[M]odern day technologies like Big Data analytics and generative AI render “full” anonymization impossible, while on the other hand, legal frameworks for personal data are being interpreted to have a broad scope of applicability” Ramya Chandrasekhar. Legal frictions for data openness: Reflections from a case-study on re-use of the open web for AI training. Centre Internet et Société - CNRS; Inno3; Open Knowledge Foundation. 2025, pp.75.

model), prolonged data retention may be justified. It must be limited to the necessary data and accompanied by enhanced security measures.

Once the data has been sorted, it can be stored on a partitioned medium, i.e. physically or logically separated from the data constituting the database. This partitioning enhances data security and restricts access to authorized persons only. The duration of the maintenance phase can vary from a few months to several years when the retention of this data poses little risk to individuals and appropriate measures have been taken. In the case of data from open sources, the retention period specified by the data source must be taken into account when determining the duration of the maintenance phase. However, this period must be limited and justified by a real need.

## 2.6. Reuse of certain personal data

Reuse of personal data in the context of AI development is common. According to the CNIL's recommendations<sup>22</sup>, **if the AI developer reuses the data it originally collected for another purpose**, in principle, it is necessary to determine whether this reuse is compatible with the initial processing through "compatibility test". This test takes into account.

1. The existence of a link between the initial objective and that of creating a database for training an AI model,
2. The context in which the personal data was collected,
3. The type and nature of the data depending on their sensitivity,
4. the potential consequences for the individuals concerned
5. the existence of appropriate safeguards (e.g. data encryption or pseudonymisation of data).

This test is not necessary when reuse is carried out for scientific research purposes. In this case, the processing is presumed to be compatible if it complies with the GDPR and is not used to make decisions about individuals.

**If the developer uses publicly accessible data**, it must ensure that the creation of this dataset is not manifestly unlawful (e.g. from a data leak). A case-by-case analysis must be performed. Reusers must ensure that:

1. The description of the database mentions its source. Example: a database whose description explains that it was compiled from publications on a specifically named professional social network.

---

<sup>22</sup> CNIL, Ensuring the lawfulness of the data processing - In case of re-use of data, carrying out the necessary additional tests and verifications, 2024 : <https://www.cnil.fr/en/ensuring-lawfulness-data-processing-re-use> and [https://www.cnil.fr/sites/cnil/files/2024-04/jeu\\_de\\_donnees\\_ia\\_modele\\_de\\_documentation.odt](https://www.cnil.fr/sites/cnil/files/2024-04/jeu_de_donnees_ia_modele_de_documentation.odt)

2. The creation or dissemination of the database is not clearly the result of a crime or offence or has been the subject of a conviction or public sanction by a competent authority involving the deletion or prohibition of further use of the data.
3. There should be no obvious doubts as to the lawfulness of the database, ensuring in particular that the conditions for data collection are sufficiently documented.
4. The database does not contain sensitive data (e.g. health data or data revealing political opinions) or criminal data (within the meaning of Articles 9 and 10 of the GDPR), or, if it does, that additional checks are carried out to ensure that such processing was lawful.

**If the developer reuses a dataset acquired from a third party**, the data controller must ensure that they are reusing a database which was created and shared in a lawful manner.

It is therefore recommended that an agreement be concluded between the reuser and this party in order to enable the latter to ensure the lawfulness of its own processing operations. In this regard, the CNIL recommends providing a number of details in the contract, such as

1. the source, the context of the data collection,
2. the legal basis for the processing, and the data protection impact assessment if necessary, in order to rule out the risk of having an unlawful database,
3. the information provided to individuals (in particular regarding the purpose and recipients),
4. any guarantees regarding the lawfulness of this data sharing by the original data holder (e.g. on the compatibility of the purpose, the lawfulness of the sharing, etc.).

## 2.7. Information of the data subjects

According to Article 12(1) GDPR, the data controller shall take appropriate measures to provide any information referred to in Articles 13 and 14.

The data subjects must be informed of the processing of their data (why, how, in what way) and how they can exercise their rights (e.g. access, rectification, etc.).<sup>23</sup> The obligation to provide information applies to data collected directly from the data subjects or indirectly (in particular through web scraping).

In practice, the recommendations are

1. Ensuring accessibility of information: the data controller must ensure that information is easily accessible. Individual information may be included on the form used to collect data, via a pre-recorded voice message, etc. General information may take the form of information published on the website of the project. It is recommended that a reasonable period of time is allowed between

---

<sup>23</sup> See section 2.8.

informing individuals about data collection and training, due to the difficulty individuals may have in exercising their rights on the model once it has been trained.

2. Ensuring the intelligibility of information: the data controller must ensure that the information is concise and clear. The complexity of AI systems must not prevent the information from being properly understood by those concerned. It is recommended to detail, for example by means of diagrams, how the data is used during training, how the AI system developed works, and the distinction that must be made between the training database, the AI model, and the model outputs. The information may be included in AI dataset, model, and system cards, but must be clearly highlighted in these documents.
3. Choosing between individual and general information: the information must be individual, but the GDPR provides for exceptions:
  - All information on the processing has already been provided to the data subject (e.g. by a third party).
  - The information would require disproportionate effort, taking into account the effort required (lack of contact details, age of the data, etc.) and the level of privacy infringement. In particular, individual information is often disproportionate when collecting pseudonymized data through web scraping, since finding the means to contact individuals may involve collecting additional or more identifying data. In this case, a comprehensive general information notice can simply be published on the project's website. Article 53 of the AI Act requires providers of general-purpose AI models to complete a public summary of the content used to train an AI system, based on a template provided by the AI Office<sup>24</sup>. This summary may serve as general information on data sources.
4. Specific information to be provided: the data controller must generally provide all the information specified in Articles 13 and 14 of the GDPR. For the development of artificial intelligence systems, a few specific points must be taken into account.
  - If the number of sources which are processed is limited: in principle, the precise identity of the sources must be provided. This applies in particular to harvesting from a limited number of websites.
  - If many sources are used, the data controller can just indicate the categories of sources, in particular the names of a few main or typical sources<sup>25</sup>. This applies to harvesting from a large number of sources. If a dataset or AI model subject to the GDPR is reused, the CNIL recommends providing the

<sup>24</sup> see section 5.1.3

<sup>25</sup> "[...] the feasibility of individually informing those whose data form part of the training set is often impractical due to the extensive effort required, potentially exempting it under Article 14(5)(b) of the GDPR" Novelli, Claudio, et al. "Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity." *Computer Law & Security Review* 55 (2024)

means to contact the data controller from whom it was obtained, especially if the dataset or model poses a significant risk to the individuals concerned.

- Informing individuals that they cannot be identified: individuals must be informed that they cannot be identified, including for the purpose of responding to their requests to exercise their rights (Article 11 of the GDPR). The CNIL recommends, where possible, providing individuals who wish to exercise their rights with additional information that could help identify them.
- The specific case of AI models subject to the GDPR: when an AI model or system is subject to the GDPR, the CNIL recommends specifying the nature of the risk associated with the extraction of personal data from the training database through the sole manipulation of the AI model, the measures taken to limit these risks, and the mechanisms for recourse if these risks materialize.

## **2.8. Data subject rights and their implications for LLM fine-tuning**

It is important that LLMs4EU allows the respect of data subject rights throughout the project's life cycle. It is to be highlighted that the GDPR grants a number of data subject rights aimed at ensuring individual control over personal data processing. These rights are:

1. The right of access under Article 15, enabling individuals to obtain information about data processing including processing purposes, data categories, and recipient information;
2. The right of rectification under Article 16, allowing correction of inaccurate or incomplete personal data;
3. The right to erasure under Article 17, permitting deletion under specific circumstances;
4. The right to restriction of processing under Article 18, enabling temporary limitation of processing operations;
5. The right to data portability under Article 20, facilitating data transfer between controllers when processing is based on consent or contract;
6. The right to object under Article 21, applicable to processing based on legitimate interest or public interest; and
7. The right to withdraw consent under Article 7(3), allowing revocation of consent-based processing at any time.

These rights are not absolute and may be subject to limitations. Controllers may for example refuse exercise of rights when requests are manifestly unfounded or excessive under Article 12(5). There can also be situations in the LLMs4EU project where processing

operations serve scientific research purposes, and the exercise of rights would render impossible or seriously impede achievement of research objectives, or where the entity receiving a request from a data subject is not the data controller for the processing in question. In all instances of refusal, controllers must provide justification and respond to the data subject.

When personal data is used in the context of fine-tuning LLMs, which can be the case at hand, data subjects must be able to exercise their rights on both training datasets and the model itself.

### **2.8.1. Data subject rights regarding the datasets**

With respect to the personal data that could potentially be included in the finetuning datasets, LLMs4EU partners need to check the necessity of the inclusion of personal data in the training/fine-tuning process, before the actual model training/fine-tuning starts. Indeed, under Article 11, controllers who do not need to identify data subjects and can demonstrate inability to do so are not obligated to store or collect additional information solely for rights exercise purposes. LLM developers typically do not require identification of individuals whose data appears in training datasets and therefore need not maintain identifiers in pseudonymised datasets specifically for rights exercise. However, transparency obligations remain in effect and following the ECJ ruling in Case C-154/21 of January 12, 2023, controllers must provide actual recipient identities rather than merely recipient categories when technically feasible.

The data subject rights can evidently be applicable in many situations beyond the applicability of Article 11. Article 15 grants individuals the right to obtain, free of charge, copies of all processed data concerning them. This extends to training dataset extracts when essential for effective exercise of other rights, as established by the ECJ in Case C-487/21 of May 4, 2023. Such provision must occur without prejudice to third-party rights and must include all relevant information to enable data subjects to assess processing lawfulness. Also, data subjects may supplement or correct data concerning them, for example by addressing incorrect or incomplete annotations within training datasets. The right encompasses both correction of inaccurate information and completion of incomplete data through supplementary statements. The framework also needs to allow a right to deletion which applies in multiple scenarios. Additionally, where processing relies on legitimate interest or public interest performance, data subjects may object based on their particular situation.

### **2.8.2. Data subject rights regarding the fine-tuned model**

The implementation of data subject rights concerning model parameters themselves presents significantly greater technical and practical challenges due to the diffuse nature

of information representation within trained models. Current research demonstrates that individual data contributions become distributed across millions or billions of parameters, making direct identification and modification extremely difficult. Thus, extracting or modifying specific personal information from model parameters faces significant technical barriers due to the non-linear transformations applied during training<sup>26</sup>.

In LLMs4EU, training data remains accessible throughout the project's life cycle. Model retraining is thus possible and represents the most reliable method for implementing rights exercise following dataset modifications. Retraining should be considered when not disproportionately burdensome to controller rights while accounting for data sensitivity levels. However, complete retraining involves substantial computational costs and time requirements, making it impractical for large-scale models in many scenarios.

Where retraining is not feasible, controllers may implement rights exercise through alternative technical measures demonstrated to be sufficiently effective. These approaches include approximate unlearning methods that modify model parameters to reduce the influence of specific data points without complete retraining. However, current unlearning techniques often lack strong guarantees regarding effectiveness.

As an additional step, LLMs4EU partners may implement safeguards at the model output level rather than modifying parameters directly. These measures include direct filtering systems that prevent generation of specific personal information, which can be achieved via named entity recognition techniques that identify and redact personal data before output delivery.

Again, it is important to note that the implementation of model-level data subject rights must consider proportionality between the technical burden on controllers and the effectiveness of rights protection. The assessment of technical feasibility should account for the current state of research and available implementation methods that are currently evolving.

#### **Data protection aspects in a nutshell**

**Assessment of personal data presence:** Technical teams must conduct an initial assessment of the datasets to determine whether they contain personal data as defined by Article 4(1) GDPR. Given the aggregation of diverse sources, datasets would likely contain personal data unless anonymisation can be demonstrated. In cases of uncertainty, consultation with WP6 is mandatory prior to data processing.

<sup>26</sup> "Future research should focus on developing scalable machine unlearning techniques to efficiently remove specific data points without necessitating complete retraining" Feretzakis, G.; Vagena, E.; Kalodanis, K.; Peristera, P.; Kalles, D.; Anastasiou, A. GDPR and Large Language Models: Technical and Legal Obstacles. *Future Internet* 2025, 17, 151.

**Determine the legal basis:** All data processing operations, especially in WP2 and WP4, should determine its legal basis. It is likely that legitimate interests are invoked as the legal basis (Article 6(1)(f) GDPR). In such cases, technical teams shall document: (i) the legitimate interest pursued by fine-tuning models for specific EU language and domain applications, (ii) demonstrate necessity by proving that data minimisation techniques have been applied and no less intrusive alternatives exist, and (iii) conduct balancing tests considering data subject rights against project benefits. The legitimate interest assessment must be completed before any data collection activities start.

**Data minimisation implementation:** Technical teams shall implement data cleaning protocols that remove inconsistencies and identify irrelevant data. Data transformation measures including generalisation, randomisation, and anonymisation techniques must be applied at the design stage. For web-scraped data, technical teams must exclude sites with technical protections (CAPTCHAs, robots.txt), establish default exclusion lists for sensitive content, and limit collection to freely accessible data. Cross-referencing based on individual identifiers must be prevented through immediate pseudonymisation post-collection.

**Joint controller arrangements:** Partners processing personal data, especially in WP2 and WP4, operate as joint controllers under Article 26 GDPR. Technical teams must implement clear responsibility allocation mechanisms and maintain documentation enabling data subject rights facilitation. Data processing activities require coordination between partners as specified in the LLMs4EU Consortium Agreement clause 4.4.

**Data retention measures:** Technical teams must establish retention periods for development and maintenance phases. During development, data must be stored securely with restricted access to authorised persons. Upon completion of model training and testing, data shall be deleted unless retained for maintenance with enhanced security measures and segregated storage. Retention periods must be documented and communicated to data subjects.

**Rights implementation framework:** Technical teams must maintain training dataset throughout the project life cycle to enable data subject rights exercise. Model retraining capabilities must be adopted as the primary method for implementing rights exercise following dataset modifications. Where retraining is disproportionately burdensome, technical teams may implement approximate unlearning methods or output-level safeguards including direct filtering systems with named entity recognition.

**Information obligations:** Technical teams must ensure accessibility and intelligibility of information provided to data subjects. Individual information requirements may be

replaced by general information notices published on the LLMs4EU website when individual contact would require disproportionate effort, particularly for pseudonymised web-scraped data. Information must specify data source categories and, where applicable, the inability to identify individuals for rights exercise purposes.

**Compliance monitoring:** Technical teams will need to implement continuous monitoring mechanisms coordinated with WP6 legal team. Regular compliance checks must be integrated into data processing workflows. Legal consultations with WP6 partners must be conducted for unclear legal situations before proceeding with data processing activities.

### 3. Intellectual Property Rights in LLMs4EU

The emergence of generative AI gives rise to significant legal challenges. AI models are trained with large sets of existing data and the methods of collecting and using this data results in concerns regarding intellectual property rights of original creators of this data.

In general, there are two categories of sources of data used by developers of generative AI models for training purposes.

The first is targeted sources data, meaning data of a small size, targeted by well-defined parameters. One key benefit of using targeted sources is that their origins are easily traceable, which enables developers to assess the quality and reliability of the targeted data. From an IP point of view, this way of collecting data assists developers in ensuring adherence to IP rights and legal obligations.<sup>27</sup> The second source of collected data is large volumes of scraped data.<sup>28</sup> This source of data contains data from websites, social media pages, databases, and has proved to be important in the development of generative AI models by supplying large and diverse datasets required for effective training.<sup>29</sup>

Further reference to the term “data scraping” in this chapter refers to the automated process of extracting information from (online) sources in the broadest sense of the word, unless defined otherwise.

In Chapter 3.1. the intersection of IP rights and generative AI is elaborated, with a specific focus on copyright and database rights. Chapter 3.2 discusses the text and data mining exceptions (TDM), as a possible solution to the challenges that currently exist with regard

<sup>27</sup> OECD, ‘Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data’ (OECD Artificial Intelligence Papers, February 2025) 33, 14; Baack S., ‘A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl’ (5 June 2024) *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

<sup>28</sup> Baack S., ‘A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl’ (5 June 2024) *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

<sup>29</sup> OECD, ‘Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data’ (OECD Artificial Intelligence Papers, February 2025) 33, 14; Soldaini, L. et al., ‘Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research’ (31 January 2024) arXiv:2402.00159.

to respecting intellectual property rights when using datasets for training and developing LLMs.

### **3.1. The intersection of Intellectual Property rights and generative AI: the Copyright Digital Single Market Directive and the AI Act**

The emergence of LLMs has given rise to a number of concerns, including those relating to copyright and database protection. The methods of collecting data by data scraping can be at odds with the rights of holders of IP. As a result, it is of great importance to understand and comply with the existing legal framework under the Copyright Digital Single Market Directive (DSM Directive)<sup>30</sup> in order to ensure lawful and ethical deployment and/or production of generative AI, particularly LLMs.

The DSM Directive establishes obligations regarding copyright (protection) and provides protection to creators of works. It aims to harmonise copyright legislation of EU member states in the digital environment and includes exceptions to copyright protection in the context of text and data mining.<sup>31</sup>

In the following chapters, copyright protection is discussed (3.1.1), as well as database protection (3.1.2). Moreover, an overview is provided of the key issues regarding the protection of copyrights and database rights under the AI Act and the DSM Directive (3.1.3).

#### **3.1.1. Copyright**

In general terms, copyright grants the author of an original work exclusive economic and moral rights over that work. These works can for example take the form of videos and images, but also texts or music. The threshold for copyright protection is in general relatively low, and no registration in order to gain copyright protection is necessary. In order for a work of authorship to meet the requirements for copyright protection, the work must satisfy two basic criteria: originality, and creativity. The work must be the author's own intellectual creation and not copied from another existing work. The work must possess a "spark" or "modicum" of creativity. This does not require a work to be novel, unique, or inventive. When an original work is created, this is upon creation automatically protected by copyright, and nobody is allowed to make this work public or reproduce this work without permission of the proprietor.

<sup>30</sup> Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market (DSM Directive).

<sup>31</sup> Chapter 3.2 and further.

The collection of data through data scraping can affect creators' copyrights, as their data can be used as input for generative AI models without their knowledge or consent.<sup>32</sup> The question whether training generative AI forms an infringement upon the copyright of authors is the subject of ongoing debate. For example, in the United States, a judge of the U.S. District Court for the Northern District of California recently ruled that training generative AI with datasets cannot be qualified as 'fair use', which forms a defence to claims of copyright infringement, and that such training therefore forms a copyright infringement.<sup>33</sup> Though this ruling does not directly affect the EU, it is one of the examples that illustrates the complexity of the rise of generative AI, and the legal implications flowing from this.

The EU exception for the use of copyright-protected material as input for training AI models is the 'text and data mining' exception, as anchored in Articles 3 and 4 in the DSM Directive. These exceptions will be further elaborated in Chapter 3.2.

### 3.1.2. Database protection and extraction rights

Database rights are protected in the European Union by Directive 96/9/EC (Database Directive)<sup>34</sup>. The Database Directive defines 'database' as "*a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means*".<sup>35</sup>

A database can be protected by two types of rights. Firstly, the contents of a database can be protected by copyright. In this case the copyright criteria that have been set out in Section 3.1.1 apply. Secondly, the database itself can be protected by a database right. With a database right, the maker of the database can prevent others from extraction and/or re-utilisation of the whole or a substantial part of the contents of that database. Also, with a database right, the rightsholder has *inter alia* the exclusive right to authorise temporary or permanent reproduction by any means and in any form, in whole or in part.<sup>36</sup>

In order to obtain such a database right, the database must be the result of qualitative and/or quantitative investments in either the obtaining, verification or presentation of the contents of the database.<sup>37</sup>

<sup>32</sup> OECD, 'Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data' (OECD Artificial Intelligence Papers, February 2025) 33, 20.

<sup>33</sup> United States District Court Northern District of California 25 June 2025, Case No. 23-cv-03417-VC (*Kadrey / Meta Platforms*).

<sup>34</sup> Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Database Directive).

<sup>35</sup> Article 1(2) Database Directive.

<sup>36</sup> Article 5(a) Database Directive.

<sup>37</sup> Article 7(1) Database Directive.

In this context, extraction means the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form. Re-utilisation means any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission. The first sale of a copy of a database within the EU by the right holder or with his consent shall exhaust the right to control resale of that copy within the EU.<sup>38</sup>

The term of protection of a database right starts at the date of completion of making of the database. It expires after 15 years from the first January of the year following the completion.<sup>39</sup> If any substantial changes are made to the contents of the database, which would result in the database being considered to be a substantial new investment, the database resulting of this new investment will have its own term of protection.<sup>40</sup>

### 3.1.3. Copyright, database rights and AI: An overview of the key issues

Copyright and database rights concerns arise at various stages of the AI life cycle or value chain. Questions can concern the early stages, such as the input (or training data) stage. In addition, questions related to intellectual property rights can also concern the AI model itself or the output that is generated by means of an AI model or system.<sup>41</sup> At the input stage, training and developing AI models involves practices such as web scraping, data collection and processing, pre-training, fine-tuning and training.<sup>42</sup> These activities often involve reproductions of protected content and are generally considered to fall under the broad legal concept of text and data mining, as defined in Article 2(2) of the DSM Directive.<sup>43</sup>

With respect to the model itself, key questions arise on the legal status of models' weights as protected databases, whether a model memorises copyrighted content and, if so, to what extent this would constitute an unauthorised reproduction under existing copyright laws.<sup>44</sup>

<sup>38</sup> Article 7(1)(b) Database Directive.

<sup>39</sup> Article 10 Database Directive.

<sup>40</sup> Article 10 Database Directive.

<sup>41</sup> Guadamuz A, 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs' (2024) 73 *GRUR International* pp.111–127; Quintais J.P, 'Generative AI, Copyright and the AI Act' (30 January 2025) 56 *Computer Law & Security Review* p. 2.

<sup>42</sup> Often, these stages are jointly referred to as "training." See e.g. European Union Intellectual Property Office (EUIPO), *Development of Generative Artificial Intelligence from a Copyright Perspective* (EUIPO, 2025).

<sup>43</sup> 'Text and data mining' means any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations.

<sup>44</sup> Cooper A.F. et al, 'Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models' (18 May 2025) arXiv:2505.12546; Cooper A.F. and Grimmelmann J., 'The Files Are in the Computer: On Copyright, Memorization, and Generative AI' (18 July 2024) arXiv:2404.12590; Dornis T.W., 'The Training of Generative AI Is

In terms of output, key issues include whether AI-generated outputs are eligible for copyright protection, constitute derivative works or infringe upon third-party works used in the training of a model.<sup>45</sup> Within the European Union, a notable and evolving issue is the extent to which copyright exceptions apply to AI-generated outputs, including those based on freedom of expression, such as quotation, criticism, review, caricature, parody and pastiche.<sup>46</sup>

Additional complexities are introduced by the intersection of copyright law with private contractual agreements, such as AI provider's terms and conditions, which can affect authorship, ownership and enforcement.<sup>47</sup>

Finally, an increasingly urgent policy concern is how creators should be remunerated for the use of their works in generative AI models.<sup>48</sup>

### 3.1.4. The DSM Directive and the AI Act

The recently adopted AI Act<sup>49</sup> has partially attempted to address some of the key copyright and database right concerns surrounding AI. An important stipulation in the AI Act in terms of copyright is Article 53 (1) (c), stating that *"Providers of general-purpose AI models [among which generative AI and specifically LLMs] put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3)*

---

Not Text and Data Mining' (2025) 2 *European Intellectual Property Review* 1; Emanuilov I. and Margoni T., 'Forget Me Not: Memorisation in Generative Sequence Models Trained on Open Source Licensed Code' (February 2024) *Zenodo*; European Union Intellectual Property Office, *Development of Generative Artificial Intelligence from a Copyright Perspective* (Alicante: European Union Intellectual Property Office); N. Sousa e Silva, 'Are AI Models' Weights Protected Databases?' (18 January 2024) *Kluwer Copyright Blog*, available at: <<https://legalblogs.wolterskluwer.com/copyright-blog/are-ai-models-weights-protected-databases/>>.

<sup>45</sup> The latter issue, often linked to memorisation, should instead be viewed from the perspective of regurgitation, extraction, and reconstruction during output generation, as highlighted by Cooper A.F. and Grimmelmann J., 'The Files Are in the Computer: On Copyright, Memorization, and Generative AI' (18 July 2024) arXiv:2404.12590. See also Sag M, 'Fairness and Fair Use in Generative AI' (2024) 92 *Fordham Law Review*, p. 1887.

<sup>46</sup> NB that the upcoming Court of Justice of the European Union (CJEU) judgement in *Pelham II* may further define "pastiche" as an autonomous EU legal concept, potentially impacting the treatment of AI outputs.

<sup>47</sup> See e.g. Bulayenko O. et al, 'AI Music Outputs: Challenges to the Copyright Legal Framework' (28 February 2022) *Social Science Research Network*; Cifrodelli G. and Edwards L., 'Copyright and Generative AI: What Can We Learn from Model Terms and Conditions?' (24 July 2024) *Kluwer Copyright Blog*, available at: <<https://legalblogs.wolterskluwer.com/copyright-blog/copyright-and-generative-ai-what-can-we-learn-from-model-terms-and-conditions/>>; Edwards L. et al, 'Private Ordering and Generative AI: What Can We Learn From Model Terms and Conditions?' (17 December 2024) *CREATE*.

<sup>48</sup> Geiger C. and Iaia V, 'The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI' 52 *Computer Law & Security Review* 1; Quintais J.P., 'Generative AI, Copyright and the AI Act' 56 *Computer Law & Security Review* 106107; Senftleben M, 'Generative AI and Author Remuneration' 54 *IIC – International Review of Intellectual Property and Competition Law* 1535.

<sup>49</sup> Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance) (**AI Act**).

of Directive (EU) 2019/790 [DSM Directive].” This stipulation therefore directly refers to one of the text and data mining exceptions as anchored in Articles 3 and 4 DSM Directive.<sup>50</sup>

Recital 105 of the AI Act provides additional explanation regarding the use of text and data mining techniques for training general-purpose AI. It refers to the text and data mining exceptions in the DSM Directive, and adds that “*rightsholders may choose to reserve their rights over their works or other subject matter to prevent text and data mining, unless this is done for the purposes of scientific research.*”<sup>51</sup> Text and data mining for the scientific research refers to the text and data mining exception of Article 3 DSM Directive.<sup>52</sup>

Recital 107 of the AI Act states that providers of general-purpose AI models must create a publicly available detailed summary of the content used for the training of AI models, in order to increase transparency on the data that is used for the training of the AI model. By creating such a summary, the need to protect trade secrets and other confidential information must be taken into account.<sup>53</sup>

Hence, the AI Act refers to (the protection of) copyright. However, the AI Act has left much room for interpretation which leads to legal uncertainty. Although the AI Act clearly requires the development of AI models to adhere to copyright law, such as the DSM Directive, it has also created uncertainties regarding the interaction between the two legal instruments.

One of the most prominent issues in the intersection of AI legislation and copyright legislation is the extraterritorial reach of the AI Act’s copyright provisions. Copyright law is inherently territorial and primarily regulates activities, such as the training of AI models, according to the laws of the territory where the activities take place. In contrast, the AI Act regulates any AI model made available to users in the European Union, while also attempting to extend its reach to training that takes place outside the European Union. This raises a pressing question: if a model is trained outside the European Union in full compliance with the local copyright laws, but those laws differ from European legislation, is it permissible to release the model in the European Union?

The resulting uncertainty poses concrete challenges for the development and adoption of AI in the EU. Furthermore, restrictive copyright rules alone are unlikely to address the broader concerns around fair remuneration for creators.

<sup>50</sup> See chapter 3.2 for an elaboration of these exceptions.

<sup>51</sup> Recital 105 AI Act.

<sup>52</sup> See chapter 3.2.1.

<sup>53</sup> Recital 107 AI Act.

## 3.2. Text and data mining exceptions and limitations

According to the DSM Directive, text and data mining means “any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations”.<sup>54</sup> Therefore, in the event that a computer model analyses copyright works or protected database rights as an automated analytical technique, in order to generate output, these acts qualify as text and data mining.<sup>55</sup> Recital 106 AI Act states that “providers of general-purpose AI models should put in place a policy to comply with Union law on copyright and related rights, in particular to identify and comply with the reservation of rights expressed by rightsholders pursuant to Article 4(3) of Directive (EU) 2019/790. Any provider placing a general-purpose AI model on the Union market should comply with this obligation, regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place. This is necessary to ensure a level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage in the Union market by applying lower copyright standards than those provided in the Union.”

The DSM Directive includes two exceptions for TDM, Article 3 (exception for the purpose of scientific research) and Article 4 (general exception). This means that in case one of these two exceptions applies, the act of TDM does not constitute copyright and/or database right infringement.

It is relevant to underline that both TDM exceptions under Articles 3 and 4 of the DSM Directive relate specifically to acts of reproduction (for copyright) and extraction (for *sui generis* database rights), but only where such acts are undertaken for the narrow purpose of analysis. The CJEU has consistently affirmed that exceptions to copyright must be interpreted strictly.

As stated above, the recently adopted AI Act<sup>56</sup> refers to the TDM exceptions of the DSM Directive in Article 53 (1) (c), as well as in recitals 105 and 107.<sup>57</sup>

Recital 105 of the AI Act clarifies that the development and training of AI models may be carried out by TDM techniques. However, such techniques can raise concerns regarding

<sup>54</sup> Article 2(2) DSM Directive.

<sup>55</sup> Knowledge Rights 21, ‘Why the EU TDM Exceptions (Art. 3 & 4) Can Be Used for AI – Including Generative AI’ (3 June 2025) available at: <<https://www.knowledgerights21.org/news-story/eu-tdm-exceptions-can-be-used-for-ai/>>.

<sup>56</sup> Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (AI Act).

<sup>57</sup> See chapter 3.1.1.2.

copyright and related rights, particularly when protected works are used as training data. The AI Act thus acknowledges the importance of the TDM exceptions established by the DSM Directive, which allow certain TDM activities under specific conditions.

Recital 107 of the AI Act implies that for training and deploying AI models in the EU, the developers must maintain, and update detailed technical documentation about the training process of the AI model. This information must be provided to the AI Office upon request of national authorities. Documentation must also be created and shared with AI system providers who integrate these models, to make sure that providers are able to understand the capabilities and limitations of the models while IP rights are protected.<sup>58</sup>

In this section, first the TDM exceptions under Article 3 of the DSM Directive (3.2.1) and Article 4 of the DSM Directive (3.2.2) is discussed. Subsequently, the applicability of TDM exceptions to the training of generative AI is discussed (3.2.3). Lastly, the rights reservation mechanisms and technical implementation methods are discussed (3.2.4).

### 3.2.1. TDM exception under Article 3

Article 3 of the DSM Directive provides the TDM exception regarding reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, TDM of works or other subject matter to which they have lawful access (Scientific TDM Exception).<sup>59</sup> Lawful access focuses on the lawfulness of the access instead of the accessed source, as it is hard to establish whether online data is posted with the authorisation of the rightsholder or not.<sup>60</sup> This means that paywalls may not be circumvented and pirate data from websites that repeatedly infringe copyright may not be scrapped. Recital 14 of the DSM Directive provides additional clarification to lawful access, by stating that *"Lawful access should be understood as covering access to content based on an open access policy or through contractual arrangements between rightsholders and research organisations or cultural heritage institutions, such as subscriptions, or through other lawful means."*<sup>61</sup>

Copies of works or other subject matter made in compliance with the Scientific TDM exception shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results.<sup>62</sup>

Article 3 allows the Scientific TDM Exception only for the purposes of a targeted objective, namely, research by this limited group of beneficiaries in an institutional setting. In

<sup>58</sup> Guadamuz A, 'The EU's Artificial Intelligence Act on Copyright' (10 November 2024) 28 *The Journal of World Intellectual Property* 213.

<sup>59</sup> Article 4(1) DSM Directive.

<sup>60</sup> Margoni T, 'Saving Research: Lawful access to unlawful sources under Art. 3 CDSM Directive?' (22 December 2023) *Kluwer Copyright Blog*, available at: <<https://legalblogs.wolterskluwer.com/copyright-blog/saving-research-lawful-access-to-unlawful-sources-under-art-3-CDSM-directive/>>.

<sup>61</sup> Recital 14 DSM Directive.

<sup>62</sup> Article 4(2) DSM Directive.

contrary to the TDM exception provided for in Article 4 DSM Directive, there is no opt-out mechanism in place for Article 3, which means that rights holders cannot contractually exclude this exception.

Recital 11 clarifies that Article 3 may also be applicable to public-private partnerships, provided that private partners are not given preferential access to the results in relation to public partners.<sup>63</sup>

#### *German court case*

The German case Kneschke/LAION<sup>64</sup> provided insights regarding the interpretation of the Scientific TDM Exception, it especially answered the question what activities should be falling under scientific research. The German Court clarified that scientific research can be defined as the “*methodical and systemic pursuit of new knowledge*”. The German Court ruled that not only the work steps directly involved to obtaining knowledge, but also the preparatory steps should be included in the definition, provided that they are aimed at the (subsequent) acquisition of knowledge.<sup>65</sup> In addition, the Court confirmed that the creation of a data set can fall under the definition of scientific research, since it is a fundamental work step with the aim that the acquired data set will be used for scientific research in a later stage.<sup>66</sup>

The German Court further clarified that data sets that are made available by a scientific institution (either free of charge or commercially) to third parties is allowed, only if it serves a scientific purpose. Subsequently, the German Court ruled that the research activity should be non-commercial, and the research success is not a decisive factor for determining if the TDM Exception is applicable.<sup>67</sup>

Lastly, in this case the German Court decided that the Scientific TDM exception does not apply when the researching organisation cooperates with a private enterprise that exerts a certain degree of influence on the organisation and has preferential access to the research results.<sup>68</sup>

---

<sup>63</sup> Lucchi N. And Hunter S., ‘Generative AI and Copyright - Training, Creation, Regulation’ (9 July 2025) Study by the European Parliament (Think Tank) PE 774.095.

<sup>64</sup> 27 September 2024 Hamburg Regional Court (Germany), Case No. 310 O 227/23 (*Robert Kneschke v. LAION e.V.*).

<sup>65</sup> 27 September 2024 Hamburg Regional Court (Germany), Case No. 310 O 227/23 (*Robert Kneschke v. LAION e.V.*), par 113.

<sup>66</sup> 27 September 2024 Hamburg Regional Court (Germany), Case No. 310 O 227/23 (*Robert Kneschke v. LAION e.V.*), par 114.

<sup>67</sup> 27 September 2024 Hamburg Regional Court (Germany), Case No. 310 O 227/23 (*Robert Kneschke v. LAION e.V.*), par 114.

<sup>68</sup> 27 September 2024 Hamburg Regional Court (Germany), Case No. 310 O 227/23 (*Robert Kneschke v. LAION e.V.*), par 127.

### 3.2.2. TDM exception under Article 4

Article 4 of the DSM Directive provides the TDM exception by means of the limitation of rights provided for in Articles 5(a) and 7(1) of the Database Directive, Article 2 of the Copyright Directive (also known as InfoSoc Directive)<sup>69</sup>, Article 4(1)(a) and (b) of the Software Directive<sup>70</sup> and Article 15 of the DSM Directive itself. The exception of Article 4 is broader than the TDM exception of Article 3 of the DSM Directive, as this exception is not limited to research organisations and cultural heritage institutions.

This exception permits TDM provided that (i) the person conducting the mining has lawful access to the work<sup>71</sup>, and (ii) the copyright holder has not expressly reserved their rights in an appropriate manner.<sup>72</sup> This means that TDM activities are permitted for users with lawful access to data, unless the rights holder has implemented an opt-out.

Article 4 (3) of the DSM Directive states that *“the exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightsholders in an appropriate manner (...)”*. This implies that there must be an option to opt-out for the rightsholders.

### 3.2.3. Applicability of TDM exceptions to training of generative AI

It is to be noted that generative AI operates in a different manner compared to TDM systems. While TDM systems extract factual or semantic insights from data, generative AI models, such as those based on transformer or diffusion architectures, are trained to synthesise new outputs by encoding and internalising expressive features of the input content.<sup>73</sup> During training, these models build multi-dimensional parameter spaces that allow them to reproduce style, structure, and composition, enabling outputs that closely resemble original creative works. This process goes beyond the analytical purpose envisaged by the TDM exceptions, moving instead toward expressive reproduction.<sup>74</sup> In legal terms, it challenges the applicability of Articles 3 and 4 to generative AI training and calls for precise clarification or reform. The TDM exceptions were not designed to

<sup>69</sup> Directive 2001/29/EC on the harmonization of certain aspects of copyright and related rights in the information society.

<sup>70</sup> Directive 2009/24/EC on the legal protection of computer programs.

<sup>71</sup> Article 4 paragraph 1 DSM Directive; Recital 14 DSM Directive.

<sup>72</sup> Article 4 paragraph 3 DSM Directive; E. Valk, I. Toepoel (IViR), ‘DPG Media et al vs. HowardsHome – A National Ruling on DSM’s Press Publishers’ Rights and TDM Exceptions’ (16 January 2025) *Kluwer Copyright Blog*, available at: <<https://copyrightblog.kluweriplaw.com/2025/01/16/dpg-media-et-al-vs-howardshome-a-national-ruling-on-dsms-press-publishers-rights-and-tdm-exceptions/>>.

<sup>73</sup> Lucchi N. And Hunter S., ‘Generative AI and Copyright - Training, Creation, Regulation’ (9 July 2025) Study by the European Parliament (Think Tank) PE 774.095.

<sup>74</sup> Lucchi N. And Hunter S., ‘Generative AI and Copyright - Training, Creation, Regulation’ (9 July 2025) Study by the European Parliament (Think Tank) PE 774.095, p. 33.

accommodate machine-based replication of creative forms on this scale or with this degree of fidelity, nor to serve commercial uses detached from scientific investigation.<sup>75</sup>

While the TDM exceptions under Articles 3 and 4 of the DSM Directive have been interpreted as the primary legal tools enabling AI developers to access and analyse copyrighted material, it is increasingly clear that these provisions were not designed with the scale, purpose, or economic impact of generative AI systems in mind.

When invoking upon the TDM Exception of Article 4 DSM, a three-step test must be complied with. This test is codified in Article 5(5) of the Infosoc Directive, and implies that exceptions: i) only apply to certain special cases, ii) do not conflict with the normal exploitation of the work, and iii) do not unreasonably prejudice the legitimate interests of the rightsholders.<sup>76</sup> When applying this test to generative AI, some concerns arise. First of all because large-scale data scraping cannot be defined as ‘special cases’. Gathering training data for generative AI is a systematic practice.<sup>77</sup> Second, generative AI models are able to replicate protected works in such a way that this can undermine normal exploitation.<sup>78</sup> Lastly, unpaid use of works can unreasonably harm the legitimate interests of the authors. Therefore, it is sometimes argued in literature that Article 4 DSM on itself does not provide a legal basis for the ingestion of protected works for generative AI training purposes. Options with regard to licensing structures or permission must therefore be considered.<sup>79</sup>

However, in literature, others argue that the exceptions for text and data mining in the DSM Directive include AI training as specified in the AI Act, by referring to Article 53 (1)(c) in combination with recitals 105 and 107 of the AI Act.<sup>80</sup> How current legislation on text and data mining exceptions should be interpreted in relation to training Generative AI models therefore does not yet seem to be unequivocally established. Future case law on this subject will need to provide clarity.

Moreover, for the LLMs4EU project, it is essential to verify prior to any TDM activity whether the data sources used as input are subject to any rights reservations of authors. In case a rights reservation is detected, the use of the specific data to which the rights reservation applies to cannot be used without an explicit licence or permission from the

<sup>75</sup> Lucchi N. And Hunter S., ‘Generative AI and Copyright - Training, Creation, Regulation’ (9 July 2025) Study by the European Parliament (Think Tank) PE 774.095, p. 33.

<sup>76</sup> Lucchi N. And Hunter S., ‘Generative AI and Copyright - Training, Creation, Regulation’ (9 July 2025) Study by the European Parliament (Think Tank) PE 774.095, 43.

<sup>77</sup> Longpre S. et al., ‘A Large-Scale Audit of Dataset Licensing and Attribution’ (2024) 6 *Nature Machine Intelligence* 975 – 987.

<sup>78</sup> Lucchi N. And Hunter S., ‘Generative AI and Copyright - Training, Creation, Regulation’ (9 July 2025) Study by the European Parliament (Think Tank) PE 774.095, 43.

<sup>79</sup> Lucchi N. And Hunter S., ‘Generative AI and Copyright - Training, Creation, Regulation’ (9 July 2025) Study by the European Parliament (Think Tank) PE 774.095, 44.

<sup>80</sup> Guadamuz A., ‘The EU’s Artificial Intelligence Act and Copyright’ (November 10, 2024) 28 *The Journal of World Intellectual Property* 213, 216.

rightsholders. If the rights reservation is not respected and the data is used as input or training data nevertheless, this may result in an infringement of IP right(s).

### 3.2.4. Rights reservation mechanisms and technical implementation

Article 4 of the DSM Directive determines that TDM is permitted unless right holders actively opt-out in an appropriate manner, such as machine-readable means. In the recital (18) of the DSM Directive is explained that rights can be reserved in relation to publicly available online content through machine-readable methods including metadata and website terms. In other situations, rights can be reserved through contractual agreement and unilateral declarations. Additionally, rights holders should be able to implement measures ensuring their opt-outs are respected.

The European Union Intellectual Property Office (EUIPO)<sup>81</sup> distinguishes between two types of opt-out measures:

- 1) Legally driven opt-out measures (unilateral declarations by right holders, licensing constraints and website terms and conditions)
- 2) Technically driven opt-out measures (robot exclusion protocol, TDM reservation protocols, tools adapted for content provenance and authenticity and broader copyright management solutions still in development). Technically driven measures can be differentiated in certain ways. Some measures can be categorised by location, for instance linked to a location where the content is hosted or found online, or asset-based, meaning they are tied directly to the content itself, regardless of its online location.<sup>82</sup> Examples of technical measures already in use that serve an opt-out function include opt-out tools by AI model providers like Google (Google-Extended)<sup>83</sup> the TDM reservation protocol,<sup>84</sup> and third-party solutions like Spawning AI's tools (e.g. the Do Not Train Tool Suite).<sup>85</sup>

<sup>81</sup> European Union Intellectual Property Office (EUIPO), Development of Generative Artificial Intelligence from a Copyright Perspective (EUIPO, 2025). Examples of the latter include the Liccium Trust Engine Infrastructure or Valuenode's Open Rights Data Exchange platform.

<sup>82</sup> European Union Intellectual Property Office (EUIPO), Development of Generative Artificial Intelligence from a Copyright Perspective (EUIPO, 2025).

<sup>83</sup> See these and other examples provided in Keller P. and Warso Z, 'Defining Best Practices for Opting out of ML Training' (28 September 2023) *Open Future Foundation*, available at: <<https://openfuture.eu/publication/defining-best-practices-for-opting-out-of-ml-training/>>; Quintais J.P., 'Generative AI, Copyright and the AI Act' (30 January 2025) 56 *Computer Law & Security Review* p. 106107.

<sup>84</sup> W3C Community Group, TDM Reservation Protocol (TDMRep) (W3C Community Group, 2024).

<sup>85</sup> Other third-party solutions, such as Cloudflare's crawl blocker – which allows content owners to block AI crawlers and prevent their works from being mined – perform a similar function but may not be considered opt-out tools within the meaning of Art. 4(3) DSM Directive. See European Union Intellectual Property Office (EUIPO); Development of Generative Artificial Intelligence from a Copyright Perspective (EUIPO, 2025); Quintais J.P., 'Generative AI, Copyright and the AI Act' (30 January 2025) 56 *Computer Law & Security Review* p. 106107. The EUIPO study considers that these measures fail to meet "the legal criteria for right reservation."

Legal uncertainty remains regarding how the opt-out requirements apply to training of LLMs. Some of the relevant legal questions are mentioned below:

- Are the different location-based, asset-based, or hybrid opt-outs all consistent with Art. 4 DSM Directive? How should we deal with potential conflicts?
- Can an opt-out in natural language integrated in terms and conditions on a website constitute an “adequate” machine-readable opt-out?
- At what moment can rights holders exercise their opt-out? At what moment during the training of a LLM model can opt-outs be enforced? Is this before, during and after pre-training and training and also during fine tuning? Or do they only apply when content is publicly available online during web scraping?

### 3.3. Licensing strategies for training data

The TDM framework offers no solutions for consent or remuneration for proprietors. Art. 4 DSM Directive provides a fee-free exception that is applicable unless proprietors actively opt out. If the exception is applicable to the training of LLMs, the exception imposes no duty to inform, credit, or remunerate authors when their work is used for training of LLMs. The TDM framework leaves proprietors without any enforceable mechanism to license the use of their work for training of LLMs.

The above-mentioned legal vacuum leads to a “value gap”. Providers of LLMs gain substantial commercial rewards by leveraging content in their models, yet proprietors gain no monetary remuneration for the use of their works for the training of LLMs. Proprietors have requested to ensure fair remuneration for their works from collective licensing schemes and revenue sharing agreements to the creation of a remuneration right for TDM.<sup>86</sup> AI providers on the other hand are of the opinion that requiring *ex ante* licensing or remuneration to proprietors would render AI research logistically and economically infeasible.<sup>87</sup>

Several potential licensing mechanisms are considered:

i) Voluntary licensing: private licensing agreements between AI developers and right holders associated in content platforms could contain remuneration for right holders. There are successful examples of these licensing agreements.<sup>88</sup> Downside is that less commercially visible works remain unlicensed.

<sup>86</sup> Initiative Urheberrecht (Copyright Initiative), ‘Authors and Performers Call for Safeguards Around Generative AI’ (19 April 2023).

<sup>87</sup> Lucchi N. And Hunter S., ‘Generative AI and Copyright - Training, Creation, Regulation’ (9 July 2025) Study by the European Parliament (Think Tank) PE 774.095, p. 81

<sup>88</sup> See Kretschmer M. et al., ‘Copyright Law and the Lifecycle of Machine Learning Models’ (February 2024) 55 *II/C – International Review of Intellectual Property and Competition Law* 110 (2024) (acknowledging the potential of collective licensing to reduce market entry barriers, but highlighting significant challenges).

ii) Extended collective licensing: In this scenario collective societies would negotiate standard TDM licenses on behalf of their members. There are standard licenses (radio, television and music) available to base the TDM licenses on. This scenario would require new legislation as the mandate of collective societies and equitable distribution would need to be regulated. Recently, the Dutch initiative GPT-NL was developed in which a group of publishers together with research institute TNO launched an AI language model trained on legally obtained Dutch data in which copyrights are protected and remuneration for authors is safeguarded.<sup>89</sup>

iii) Statutory Remuneration Right for TDM: This option would involve the creation of a new exclusive right that would require AI developers to pay a fee or share of profits when copyright protected works are used in training of LLMs. This option raises difficult question as to what the impact of a work is on an AI model and how to make sure that no double payments are made when the content is available in multiple datasets.

#### IPR aspects in a nutshell

**Prioritisation of non-IPR-restricted data:** Technical teams are instructed to prioritise the identification and use of datasets that are not subject to copyright and *sui generis* protections. This includes proprietary data owned or licensed by consortium members where full rights for use in LLM training are secured. Additionally, open data compliant with the EU Open Data Directive and similar public domain or Creative Commons licensed resources should be preferred.

**Mandatory copyright assessment:** Technical teams must conduct initial copyright evaluation of all datasets to identify protected works under the DSM Directive. Any collection of independent works, data, or materials arranged systematically qualifies as a database under Directive 96/9/EC and requires assessment for both copyright protection of contents and *sui generis* database rights. This evaluation needs to determine whether a TDM exception applies. Documentation of this assessment is mandatory before any data processing commences.

**Rights reservation verification:** When article 4 DSM Directive is relied upon, technical teams shall implement detection of rights reservations expressed under the same article of the directive before scraping. This includes scanning for machine-readable opt-out mechanisms (robots.txt files, TDM reservation protocols), reviewing website terms and conditions, and checking for unilateral declarations by rightsholders. In this context, data subject to active rights reservations cannot be processed without explicit licensing agreements. Detection failures may result in copyright and database right infringement.

<sup>89</sup> Website GPT-NL, available at: <<https://gpt-nl.nl/>>

**TDM exception compliance framework:** Technical teams must ensure adherence to the three-step test under Article 5(5) InfoSoc Directive when invoking Article 4 DSM exceptions. Data collection must constitute "certain special cases," avoid conflict with normal exploitation of works, and not unreasonably prejudice legitimate interests of rightsholders. Large-scale systematic scraping operations face heightened scrutiny under this framework. Where Article 3 scientific research exceptions apply, teams must demonstrate methodology consistent with "systematic pursuit of new knowledge" as established in Kneschke/LAION precedent.

**Licensing strategy:** Technical teams shall coordinate with WP6 to assess whether voluntary licensing, extended collective licensing, or statutory arrangements are required for specific data categories. Commercial partnership agreements must specify copyright clearance responsibilities and include warranties regarding lawful data sharing. Licensing documentation must accompany all proprietary datasets and be maintained throughout the project life cycle.

**Source documentation requirements:** Technical teams must maintain source documentation allowing identification of rightsholders where technically feasible. For limited source collections, precise identity of sources must be recorded. For large-scale collections, categories of sources with representative examples satisfy documentation requirements. This information supports both transparency obligations and enables rightsholder contact for rights exercise purposes.

**Output liability management:** Technical teams must implement filtering mechanisms preventing generation of content that closely resembles protected works from scraped fine-tuning datasets. Direct reproduction safeguards must be integrated at the model output level, such as content matching systems. Regular auditing of model outputs for potential copyright infringement is required, coordinated with legal compliance monitoring under WP6.

## 4. Trade secrets

Confidential information that is valuable is often protected as trade secret<sup>90</sup>. Trade secrets, also referred to as know-how, are of a diverse nature. They may relate to the

<sup>90</sup> Shira Perlmutter, *One Year On: Developments in the Protection of Trade Secrets*, U.S. PATENT & TRADEMARK OFFICE (June 19, 2017), <https://www.uspto.gov/subscription-center/2017/one-year-developments-protection-trade-secrets> (last visited Mar. 2, 2024).<sup>90</sup> Website GPT-NL, available at: <<https://gpt-nl.nl/>>

composition of products, the method of production, the organisation of the company, lists of customers, the treatment of customers, etc. In the context of AI<sup>91</sup>, elements such as algorithms and training processes, training data, parameters (including parameter weights), applications, source code and computers or other hardware devices may be protected as trade secrets.

The difference between the protection of intellectual property rights, analysed previously, and trade secrets is that the protection of IP rights is limited in time (for example patents with a maximum of 20 years) and there are specific exceptions applicable (also in relation to training of LLMs) to the exclusive rights of the proprietors of these IP rights. IP rights are most of the time publicly known and thereby others can build on these IP Rights. Trade secret protection is not limited in time and a little to no exceptions are available to the exclusive rights of the proprietor of the trade secrets. Trade secrets are by nature not publicly known and therefore not always easy to understand if data could be protected as trade secret.

The AI Act<sup>92</sup> also refers to trade secrets. The AI Act introduces the obligation for providers of general-purpose AI models to draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose AI model<sup>93</sup>. In Recital 167 it is stated that all parties involved should carry out their tasks and activities in such a manner that the trade secret legal framework is respected.<sup>94</sup> Also in Articles 25(5), 52(6), 53(1)(b), 55(3), 78(1)(a) reference is made to trade secrets.

AI, and more specifically LLMs, may jeopardise the protection of trade secrets. The training of LLMs involves large amounts of training data collected from different sources, usually using scraping methods. Depending on the types of sources involved, these scraped datasets may also contain trade secrets, which is not always compliant with the applicable trade secret legal framework. The training of the LLM with datasets that contain trade secrets is potentially infringing the rights of the proprietors of these trade secrets and the consequences of the use of these trade secrets in LLMs are most of the time irreversible. In the situation trade secrets are used to train LLMs, the LLM can reproduce this valuable information on request in seconds with little to no human intervention.

<sup>90</sup> Shira Perlmutter, *One Year On: Developments in the Protection of Trade Secrets*, U.S. PATENT & TRADEMARK OFFICE (June 19, 2017), <https://www.uspto.gov/subscription-center/2017/one-year-developments-protection-trade-secrets> (last visited Mar. 2, 2024).

<sup>91</sup> Lee N., 'Protection for Artificial Intelligence in Personalised Medicine – The Patent/Trade Secret Trade Off' p. 7, in Schovsbo J., Minssen T. & Riis T. (Eds.), *The harmonization and protection of trade secrets in the EU: An appraisal of the EU directive* (pp. 267-294).

<sup>92</sup> Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (**AI Act**).

<sup>93</sup> Recital 108 AI Act.

<sup>94</sup> Recital 167 AI Act.

Therefore, if trade secrets are used for training of LLMs without the permission of the proprietor of the trade secret it could potentially lead to damages and liabilities. This could also lead to a loss of trust.

## 4.1. International protection of trade secrets

The Agreement on Trade-Related Aspects of Intellectual Property Rights ('TRIPs Agreement')<sup>95</sup> of the World Trade Organisation ('WTO') is one of the first international treaties that addresses the protection of trade secrets. A total of 164 countries are signatories to the TRIPs Agreement which gives it a global coverage.<sup>96</sup>

Article 39 of the TRIPs Agreement prohibits the use of undisclosed information and thus obliges the protection of undisclosed information against disclosure, acquisition or use by third parties in a manner that is contrary to honest commercial practices. In a footnote of the TRIPs Agreement is explained in which situation this may be the case, such as breach of contract, abuse of confidence and inducement to infringe. Third parties who knew or were seriously negligent in not knowing that the know-how in question had been acquired through such conduct are also acting contrary to honest commercial practices.<sup>97</sup> Conditions for protection are that the know-how is not generally known or easily accessible to those who are normally engaged in the relevant field of information, that it has commercial value because it is secret, and that the owner of the know-how has taken reasonable measures to keep it secret.

Article 39 of the TRIPs Agreement has no direct effect in the states that are parties to the TRIPs Agreement. Despite the TRIPs Agreement, there are considerable differences between the laws of the various states regarding the protection of trade secrets against their unlawful acquisition or disclosure or their unlawful use by third parties. For example, there are differences in the definitions of trade secrets and unlawful acquisition, disclosure or use of a trade secret.<sup>98</sup> This has led to differences in the legislation of the various states, with the result that trade secrets do not enjoy the same level of protection throughout the different states, which in turn may act as a brake on cross-border innovation-related economic activities. In view of this fragmented protection, the European Union decided to adopt the Trade Secrets Directive.

<sup>95</sup> Trb. 1995, 130 (TRIPs: pp. 361- 400). Also see Dutch Explanatory memorandum to the proposal for a national law approving the WTO Agreement, 'Memorie van Toelichting bij het voorstel van rijkswet tot goedkeuring van het WTO-verdrag' (Kamerstukken II 1994/95, 23 961 (R 1525), nr. 3).

<sup>96</sup> Website WTO, available at: <[https://www.wto.org/english/thewto\\_e/whatis\\_e/tif\\_e/org6\\_e.htm](https://www.wto.org/english/thewto_e/whatis_e/tif_e/org6_e.htm)>.

<sup>97</sup> Geerts P. et al, *Kort Begrip van het intellectuele eigendomsrecht* en anderen, *Kort Begrip van het Intellectuele eigendomsrecht*, 2017 12<sup>e</sup> druk, p. 722.

<sup>98</sup> Recital 6 of the Directive (EU) 2016/943.

## 4.2. Trade Secrets Directive (EU) 2016/943

On 14 April 2016, Directive (EU) 2016/943 of the European Parliament and of the Council on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure (the “Directive”) entered into force. The Directive aims to establish rules on the protection against the unlawful acquisition, use and disclosure of trade secrets. The Directive predominantly seeks minimum harmonisation, leaving Member States of the European Union free to provide further protection in their national legislation than the protection provided by the Directive. This has resulted in different national regulations in the Member States of the European Union that are applicable on trade secrets.

The Directive contains four chapters: i) scope, ii) acquisition, use and disclosure of trade secrets, iii) measures, procedures and remedies, and iv) penalties, reporting and final provisions). We will discuss the material articles in subsections i) and ii) below.

### *i) Scope of application*

Article 2 (1) of the Directive contains a number of definitions, including the definition of trade secret. A trade secret must meet the following cumulative conditions:

1/ The information must be secret. This means that the information is not generally known or easily accessible to persons within the circles that are normally involved in the relevant type of information. This includes competitors. The test is not based on the knowledge of general public, but on the level of knowledge of experts. Experience and skills acquired by employees in the normal course of their employment are also excluded from the definition.

2/ the information must have commercial value because it is secret. This concerns both actual and potential commercial value. Information is quickly deemed to have “commercial value”. This is the case when the unlawful use or disclosure of the information could be detrimental to the right holder. Such harm may include damage to the right holder's business or financial interests, strategic positions or competitive advantage.

3/ the information has been subject to reasonable measures by the person lawfully in possession of it to keep it secret, given the circumstances. These measures must be reasonable, acceptable, plausible or appropriate in the circumstances of the case. A deliberately broad definition of “reasonable measures” has been chosen rather than a list of measures that must be taken. The following measures are considered to be reasonable:

- Including confidentiality clauses in commercial contracts and employment contracts;

- Applying a need-to-know policy: only employees who need to know have access to the secret;
- Monitoring the company premises;
- Explicitly identifying and registering trade secrets, for example by means of an i-depot; and
- Encrypting digital information.

*ii) the acquisition, use and disclosure of trade secrets*

In Article 2(1)(4) of the Directive “infringer” is defined as any natural or legal person who has unlawfully obtained, used or disclosed a trade secret.

It follows from Article 4(2) that the acquisition of a trade secret without the consent of the trade secret holder is considered unlawful if the acquisition was obtained by means of:

- unauthorised access to or unauthorised appropriation or copying of documents, objects, materials, substances or electronic files lawfully in the possession of the holder of the trade secret and containing the trade secret or from which the trade secret can be derived;
- other conduct which, given the circumstances, is considered contrary to honest commercial practices.

In Article 4(3) of the Directive, it is stated that the use and disclosure of a trade secret is considered unlawful if it is carried out without the consent of the holder of the trade secret by a person who meets one of the following conditions:

- has obtained the trade secret unlawfully;
- is in breach of a confidentiality agreement or other obligation not to disclose the trade secret;
- is in breach of a contractual or other obligation to restrict the use of the trade secret.

It follows from Article 4(4) that the acquisition, use or disclosure of a trade secret is also considered unlawful if, at the time of obtaining, use or disclosure, a person knew or, given the circumstances, should have known that the trade secret was obtained directly or indirectly from another person who unlawfully used or disclosed the trade secret within the meaning of Article 4(3).

Article 4(5) of the Directive provides that the production, offering or placing on the market of infringing goods, or the import, export or storage of infringing goods for those purposes, shall also be considered an unlawful use of a trade secret if the person carrying out such activities knew or, given the circumstances, should have known that the trade secret was being unlawfully used within the meaning of Article 4(3).

Article 5 contains several exceptions to the exclusive right of the proprietor of the trade secret. The article stipulates that there is no infringement of a trade secret in the following cases:

- the exercise of the right to freedom of expression and information as laid down in the Charter of Human Rights of the European Union.
- the disclosure of misconduct, errors or illegal activities, provided that the defendant acted with the intention of protecting the public interest.
- the disclosure of the trade secrets by employees to their representatives in the course of the legitimate exercise of their representative functions in accordance with Union or national law, provided that such disclosure was necessary for the exercise of those functions;
- for the protection of a legitimate interest recognised by Union or national law.

### 4.3. Ensuring trade secrets are protected in the fine-tuning of LLMs

The development or fine-tuning of high-quality generative AI systems based on machine learning often requires access to big amounts of data. AI training data can include personal information, facts, creative content, software, audio files, videos, digital images, and virtually any other digital content.<sup>99</sup> All of this data could potentially fall within the scope of the definition of a trade secret of the Directive.

As previously mentioned, data scraping is used in the development of LLMs. In the situation that the AI training data contains trade secrets, all of the techniques mentioned above used for scraping AI training data without the permission of the proprietor of the trade secret are considered to be infringing the exclusive rights of the proprietor of the trade secrets as defined in Art. 4(3) and 4(4) of the Directive. Further the scraping of data that violates the terms of use could also result in a breach of contract.<sup>100</sup>

#### *Trade secrets in Targeted Source Data vs Scraped Data*

In general, data sources used to train and fine-tune LLMs can be put under two categories.<sup>101</sup> These are 1) targeted sourced data ('Targeted Source Data') and 2) large amounts of scraped data ('Scraped Data').<sup>102</sup> These data sources correspond to the data to be collected in LLMs4EU.

<sup>99</sup> Clark & Perrault, 2022.

<sup>100</sup> CJEU, 15 January 2015, C-3/14 (Ryanair/PR Aviation)

<sup>101</sup> OECD, 'Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data' (OECD Artificial Intelligence Papers, February 2025) 33, 14.

<sup>102</sup> Baack S., 'A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl' (5 June 2024) *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Target Source Data are characterised by its smaller size, well-defined parameters, and selection based on specific quality criteria (accuracy, completeness, consistency, reliability, validity, and timeliness). A crucial advantage of Targeted Source Data is that its origin can be clearly identified, allowing developers to verify the quality and reliability of the data. This is advantageous for compliance with the trade secrets legal framework. For example, by requesting permission from holders of trade secrets or invoking one of the exceptions of the Directive.

Advantages of Targeted Source Data include higher quality of the data, compliance with laws and regulations, accuracy of the selected data for specific purposes, and consultation and cooperation with data providers. Disadvantages of Targeted Source Data are its limited size and associated costs (for licences, access to the data, and service agreements). In the context of Targeted Source Data, various licensing structures are available for free use. Free licences allow free use of Targeted Source Data but may impose restrictions on modifications or redistribution of the Targeted Source Data. In addition, there are also 'libre licenses' that offer more freedom to modify, share, or redistribute the Targeted Source Data subject to specific license conditions. Much Targeted Source Data contains protected content, including trade secrets, which is subject to specific conditions requiring compliance with the Directive and additional contractual obligations. Such license agreements are considered a form of legal certainty that helps both developers and rights holders to manage risks.

Given the size constraints of targeted datasets, developers frequently supplement their training data with large-scale Scraped Data.<sup>103</sup> This Scraped Data provides the extensive and contextually rich corpus necessary for effective models.<sup>104</sup> Advantages of Scraped Data include large amount of data, limited costs, and scalability. Disadvantages of Scraped Data include the risk of low quality, ethical and legal risks, potential for overload (overwhelming websites with excessive requests leading to blocking or legal repercussions), and maintenance (scraping scripts must be constantly updated).<sup>105</sup>

#### *Trade secret protection in the fine-tuning process in LLMs4EU*

If the scraped data falls within the scope of protection of the trade secret legal framework, compliance is mandatory as infringement on the rights of proprietor of trade secrets could potentially lead to damages, liability and loss of trust.

<sup>103</sup> Baack S., 'A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl' (5 June 2024) *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

<sup>104</sup> Soldaini, L. et al., 'Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research' (2024) <<https://arxiv.org/abs/2402.00159>>.

<sup>105</sup> OECD, 'Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data' (OECD Artificial Intelligence Papers, February 2025) 33, 14.

To be compliant with trade secret legal framework such as the Directive and TRIPs Agreement the following questions should be answered, and risk analyses should be taken prior to scraping of data for the fine-tuning of LLMs:

- a. Which dataset are collected? (Targeted Source Data or Scraped Data?)
- b. Is there an opt-out applicable on the data stating that the proprietor of that dataset does not allow data scraping?
- c. How likely is that these datasets are protected by trade secret legal framework? Targeted Source Data is more likely to contain trade secrets than Scraped Data as Scraped Data (such as data on websites or social media platforms)<sup>106</sup> is most of the time publicly available and thereby does not have a confidential nature. But be aware that even Scraped Data that seems publicly available is not always fully completely available and could still qualify as trade secret.<sup>107</sup>
- d. To determine whether a specific dataset is protected as a trade secret it is of importance to understand if the following cumulative conditions are fulfilled;
  - i.) Does the data have confidential nature?
  - ii.) Is the data valuable? and
  - iii.) Are there reasonable measures taken to maintain the confidential nature of the data?

If one of the aforementioned cumulative criteria is not fulfilled the data is most likely not protected as trade secret. It could still be that this data is protected by other rights such as IP rights, but it can be assumed that this data is not protected as trade secret.

- e. In the situation the proprietor of the data is known, it is advisable to ask specific permission to use this data for the fine-tuning of LLMs. If permission is given by the owner of the data, the use of data is compliant with trade secret legal framework.

<sup>106</sup> J. E. Barbee, *Scratching the Surface of IP Rights: Data scraping and Artificial Intelligence*, NYSBA, December 22, 2023.

<sup>107</sup> See for example US case 959 F.3d 1288 (11th Cir. 2020) *Compulife Software Inc v. Newman*. Compulife, a company that produced a database of life insurance quotes, sued a competitor for misappropriating trade secrets and violating a state anti-hacking statute, among other things. While Compulife's database was available to the public, individual users were only able to retrieve individual quotes from the database – with the database structured that way, it would simply take too long for a person to retrieve every quote in the database. In that sense, the public did not have access to the entire database; the public only had access to the individual set of quotes requested. Even though the database was technically available to the public, it could not practically be recreated or copied in its entirety by any single person. Based on that reasoning, the 11th Circuit concluded that a data scraping program could misappropriate trade secrets by copying an entire database of information, even when that database is technically available to the public. It is far from clear whether other courts will follow the same rationale, but the 11th Circuit established one principle that other courts may consider: even if an AI tool can access data, which does not mean that the AI tool should access that data.

- f. If the owner of the data is not known or not responding on the request to give permission, determine if an exception in the Directive is applicable.
- g. If all of the above still leads to uncertainty about whether the use of data is compliant with the trade secret legal framework, it is important to determine if the data is absolutely necessary for the fine tuning of the LLM. If that answer is positive, the following actions can be taken to decrease the risks involved with the use of this data:
  - a. Determine if it is possible to shield the data instead of making it publicly available;
  - b. Create models for restoring and removing data from a dataset if a problem on the basis of the trade secret legal framework is identified.
  - c. Ensure transparency. Make it easy for proprietors of trade secrets without a technical background to check whether their trade secrets are included in a dataset

### Trade secret aspects in a nutshell

**Trade secret identification and risk assessment:** Technical teams must identify and assess trade secret risks embedded in datasets considered for LLM fine-tuning. Trade secrets include confidential know-how, algorithms, parameters, source code, and hardware designs that possess commercial value and reasonable secrecy safeguards under Directive (EU) 2016/943 and the TRIPs Agreement. Prior to data collection, teams shall evaluate whether the data satisfies the cumulative criteria of secrecy, commercial value, and protective measures. This risk assessment is mandatory to prevent unlawful acquisition, use, or disclosure.

**Scraping compliance and permission verification:** All data scraping activities must ensure explicit permission from trade secret proprietors where protected information is involved. Technical teams must distinguish between Targeted Source Data, typically higher risk but legally cleared through licenses, and large-scale Scraped Data which generally pose lesser trade secret risks but still require due diligence.

**Protective measures and transparency protocols:** Technical teams shall implement robust confidentiality measures including encryption, need-to-know policies, and monitoring protocols to respect, among other considerations, trade secret protection throughout data processing and storage. Transparent documentation must be maintained and where necessary, data shielding techniques or selective model retraining/removal processes should be developed to mitigate identified trade secret exposures.

**Legal exception evaluation:** In cases of uncertain permission status or unknown data proprietors, teams must assess applicability of Directive exceptions (e.g. public interest) and evaluate whether the data is strictly necessary for model fine-tuning. Alternative

approaches to decrease trade secret risks include shielding sensitive data from training sets and employing technical methods for data removal or obfuscation post-training.

**Coordination with legal governance:** Technical teams must work closely with WP6 legal governance for comprehensive evaluations, licensing negotiations, and timely resolution of trade secret-related compliance issues. WP6 partners should be consulted proactively to guide adherence to evolving trade secret frameworks and reconcile conflicts between proprietary information protection and AI development objectives.

## 5. Aspects related to the Artificial Intelligence Act

This section examines the provisions of the AI Act that are particularly relevant for the LLMs4EU project, in particular obligations that may affect the development and deployment of fine-tuned LLMs. The adoption of the AI Act represents a pivotal development in the regulation of artificial intelligence within the European Union. As the first comprehensive legislative framework governing AI, it establishes harmonised rules that affect both the design and deployment of AI systems, including General-Purpose AI (GPAI) models and fine-tuned LLMs.

The emergence of GPAI models, characterised by their generality, versatility, and ability to perform a wide range of tasks, raises significant compliance considerations for developers and downstream modifiers. These include obligations relating to transparency, technical documentation, and public disclosure of training content, as well as the identification of circumstances under which a fine-tuned LLM may itself be classified as a GPAI model.

In addition to GPAI-related provisions, the AI Act also introduces specific rules for high-risk AI systems, defined on the basis of their intended use and potential impact on health, safety, or fundamental rights. Determining whether any of the use cases in LLMs4EU fall under this category is essential for ensuring full compliance with the regulation.

This section therefore provides an overview of the key concepts and obligations introduced by the AI Act, focusing on their relevance to the LLMs4EU project. It first examines the definition and scope of GPAI models (Section 5.1.1), the applicability of this framework to fine-tuned LLMs (Section 5.1.2) and the obligations for providers of GPAI models (Section 5.1.3). It then examines the classification criteria for high-risk systems (Section 5.2.1), and provides a preliminary analysis of potential high-risk classification of use-cases developed in the project (Section 5.2.2) followed by an overview of the obligations concerning data in the context of a high-risk AI system (Section 5.2.3). A final section outlines the implementation timeline of the AI Act as well as its evolving nature (Section 5.3). The analysis aims to clarify the regulatory implications for the project's

activities, ensuring that all partners are equipped to meet the Act's compliance requirements while continuing to advance trustworthy and legally sound AI solutions.

## 5.1. General-Purpose AI models

### 5.1.1. Definition of General-Purpose AI model

Article 3(63) of the AI Act defines a General-Purpose AI (GPAI) model as “an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market”.

The key criteria are the significant generality and the capability to perform a wide range of tasks. According to recital 98 of the AI Act, these criteria are presumed to be met when the model displays at least 1 billion of parameters and was trained with a large amount of data using self-supervision at scale<sup>108</sup>.

Given the wide margin of interpretation that these conditions still allow, the Commission has adopted a different approach in recent guidelines<sup>109</sup>, establishing a training compute threshold measured in FLOP<sup>110</sup>, above which an AI model should be considered a GPAI model, mirroring what is already provided for the classification of GPAI models with systemic risks<sup>111</sup>. Here, the minimum required training compute to qualify as a GPAI model is set at  $10^{23}$  FLOP.

In addition to this threshold, the Commission specifies that the model's capabilities must be assessed based on the modalities it offers, assigning decisive importance to the ability to generate language, whether in text or speech. The Commission states that “no other modality confers such a wide range of capabilities”, and as a consequence, considers that “models that generate language are typically more capable of competently performing a wider range of tasks than other models”<sup>112</sup>.

In other words, any language-generating model trained with more than  $10^{23}$  FLOP is presumed to be a GPAI model.

---

<sup>108</sup> Recital 98, AI Act

<sup>109</sup> European Commission, Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act, 18 July 2025, <https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act>

<sup>110</sup> Floating-point operations

<sup>111</sup> Article 51(2), AI Act

<sup>112</sup> European Commission, Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act, 18 July 2025, §19.

The core definition provided in Article 3(63) echoes the pre-existing notion of foundation model<sup>113</sup>, and inherently includes LLMs that are presented as highly versatile tools. Nevertheless, the clarifications provided by the Commission regarding the criterion of “a wide range of distinct tasks” help clarify the status of LLMs tailored for specific tasks within a specific domain. Given that the nature of language generation typically entails the ability to perform a wide range of distinct tasks, regardless of any intended specialisation, such LLMs are, in principle, still considered GPAI models.

### 5.1.2. Applicability of obligations for providers of GPAI models to fine-tuned LLMs

Under the AI Act, fine-tuning an LLM qualifies as a modification of a GPAI model<sup>114</sup>. The entity placing the GPAI model on the market<sup>115</sup> is a provider of GPAI model, and any other entity modifying it afterwards is, in practice, a downstream modifier. However, the latter could also be regarded as a provider of a new GPAI model, as according to Recital 97 of the AI Act “models may be modified or fine-tuned into new models”. This would subject the downstream modifier to the obligations set out in Chapter V of the AI Act, including in particular those described in section 5.1.3 of this report.

Although fine-tuning has played a key role in the rise of LLMs from the outset, and the definition of GPAI models itself refers to their potential for downstream task adaptation, the AI Act does not address certain aspects that are yet necessary to fully understand the regulatory framework that applies. Most notably, it does not specify what level or type of modification would qualify the downstream modifier as a provider of a new GPAI model.

On this point, the European Commission’s recently published guidelines offer important clarifications. In a section specifically addressing the issue of downstream modifiers<sup>116</sup>, two key statements are made:

- “The Commission deems that it is not necessary for every modification of a general-purpose AI model to lead to the downstream modifier being considered the provider of the modified general-purpose AI model.”

<sup>113</sup> “Any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g. fine-tuned) to a wide range of downstream tasks” Bommasani et al., 18 August 2021, On the Opportunities and Risks of Foundation Models.

<sup>114</sup> “The Commission considers ‘fine-tuning’ to be one way of ‘modifying’ a general-purpose AI model.” Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act, European Commission, 18 July 2025, §60.

<sup>115</sup> “General-purpose AI models may be placed on the market in various ways, including through libraries, application programming interfaces (APIs), as direct download, or as physical copy.” Recital 97 AI Act.

<sup>116</sup> Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act, European Commission, 18 July 2025, section 3.2.

- “The Commission considers a downstream modifier to become the provider of the modified general-purpose AI model only if the modification leads to a significant change in the model’s generality, capabilities, or systemic risk.”

To further facilitate the interpretation of these conditions, the Commission uses the same approach as for defining a GPAI model, and again sets an indicative threshold based on training compute. Two distinct situations are identified, each leading to a different threshold, as follows:

- Where the downstream modifier knows or is able to determine the training compute used for the original model, they will be considered the provider of a GPAI model if the training compute used for the modification is greater than a third of the training compute of the original model.
- Where the downstream modifier cannot be expected to know this value, they will be considered the provider of a GPAI model if the training compute used for the modification is greater than a third of the threshold for a model to be presumed to be a GPAI model (i.e.  $10^{23}$  FLOP).

A result, GPAI models that already offer detailed documentation should be favoured when selecting a foundation model to fine-tune, so that the applied threshold can best reflect the actual impact of the modification.

Generally speaking, the concept of “training compute” and the methodologies for estimating the different thresholds are detailed in the annex attached to the Commission’s guidelines<sup>117</sup>. This annex is a key tool that will be directly used in the LLMs4EU project to determine whether each of the fine-tuned LLMs must comply with the obligations for providers of GPAI models.

### 5.1.3. Obligations for providers of GPAI models

#### Public summary of training content

Article 53(1)(d) of the AI Act obliges providers of GPAI models to draw up and make publicly available a sufficiently detailed summary of the content used for training, according to a template provided by the AI Office. Recital 107 further clarifies that this summary must be “generally comprehensive” while balancing transparency with the protection of trade secrets and confidential business information.

The European Commission<sup>118</sup> provides the official explanatory notice and template for this summary. The preparation of a public summary of training data serves several

<sup>117</sup> Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act, European Commission, 18 July 2025, Annex: Training compute of general-purpose AI models.

<sup>118</sup> European Commission, Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models required by Article 53 (1)(d) of Regulation (EU) 2024/1689 (AI Act) - <https://digital->

purposes. It enhances transparency and accountability by providing meaningful information about the data used for training or fine-tuning models. It also enables stakeholders, such as rights-holders and data subjects, to verify compliance with copyright and data protection obligations and, when necessary, to take steps to enforce their rights. The summary is equally important for downstream developers and deployers, as it offers insight into the nature and diversity of the training data and allows them to assess potential risks and biases associated with the models. By publishing this information, the project aligns with the transparency obligations of the AI Act while safeguarding commercially sensitive information and trade secrets.

In the LLMs4EU project, no foundation model will be trained. Instead, existing GPAI models will be fine-tuned for specific use cases. In line with Section 6 of the Commission explanatory notice<sup>118</sup>, only the additional datasets used for fine-tuning must be reported, with a clear reference to the base model and a link to its existing public summary.

The template is divided into three main sections. The first section will include general information about the provider, the fine-tuned model, and its dependencies, explicitly referencing the base GPAI model and providing a link to its public summary whenever available. It will also describe the modalities present in the fine-tuning data (e.g. Text, Image, Audio, etc.), indicate the approximate data size within the broad ranges established by the template (e.g. tokens for text, hours for audio, etc.), and provide a brief description of the types of content and linguistic characteristics involved.

The second section will include information about the specific data sources used in fine-tuning. This includes publicly available datasets, private licensed datasets, data obtained through web scraping, user data where applicable, and synthetic data if generated for the purpose of fine-tuning.

Finally, the summary will describe relevant data-processing aspects, such as the measures taken to respect copyright reservations, procedures to ensure the removal of illegal content, and safeguards implemented for the protection of personal data and trade secrets.

The explanatory notice specifies that the information should be provided in narrative form, and that only high-level information is required, without disclosing the exact composition of the datasets. If the same summary applies to several models or versions, this must be indicated.

Finally, the summary must be made publicly available at the latest when the model is placed on the Union market and must be published on the provider's official website and

---

[strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models](https://strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models)

in all public distribution channels. For models placed on the market before 2 August 2025, providers have until 2 August 2027 to publish the corresponding summary.

According to Article 53 of the AI Act, providers of all general-purpose AI models placed on the Union market must fulfil the obligation of providing a summary, including providers of general-purpose AI models released under free and open-source licenses<sup>119</sup>.

### **Model card and technical documentation**

Article 53 of the AI Act establishes the obligations for providers of general-purpose AI models. It requires providers to prepare and keep up to date both the technical documentation of the model and the information necessary for downstream providers to understand the capabilities and limitations of the model. Specifically, Article 53(1)(a) refers to the technical documentation containing at least the elements set out in Annex XI, while Article 53(1)(b) requires making available transparency information to downstream providers as listed in Annex XII.

The European Commission guidelines<sup>120</sup> stress that these transparency and documentation obligations apply throughout the entire life cycle of the model, from pre-training to fine-tuning and beyond. As outlined in Section 5.1.2 of this report, when a model is modified or fine-tuned by another actor, that actor becomes the provider of the modified version and may be required to comply with the relevant obligations proportionate to the modification. In the context of the current project, as no foundation models are being created but rather existing models are fine-tuned, the documentation will have to focus on the modifications and adaptations made during fine-tuning rather than on the full pre-training process.

The Transparency Chapter of the Code of Practice for General-Purpose AI Models<sup>121</sup> provides a practical framework for fulfilling these obligations, including a Model Documentation Form aligned with Annexes XI and XII of the AI Act.

The requirement for a model card and associated technical documentation is grounded in the objectives of promoting transparency, accountability, and trust in AI. By making information on model design, training data, capabilities, and limitations available, downstream providers can better integrate models into their systems and fulfil their own obligations under the AI Act. This is particularly important in a scenario where models are

<sup>119</sup> According to Recital 102 AI Act “General-purpose AI models released under free and open-source licences should be considered to ensure high levels of transparency and openness if their parameters, including the weights, the information on the model architecture, and the information on model usage are made publicly available. The licence should be considered to be free and open-source also when it allows users to run, copy, distribute, study, change and improve software and data, including models under the condition that the original provider of the model is credited, the identical or comparable terms of distribution are respected.”

<sup>120</sup> European Commission, Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act, 18 July 2025

<sup>121</sup> [2] Code of Practice for General-Purpose AI Models, Transparency Chapter

fine-tuned for specific applications: understanding the provenance and characteristics of the base model is essential to assess compliance, risks, and ethical considerations in deployment.

The model card and technical documentation must contain a general description of the model, and a detailed description of the development process.

The transparency information for downstream providers further specifies how the model interacts with external hardware or software, versions of relevant software, and additional technical means required for integration. For fine-tuned models, it is essential to indicate the dependencies on the base model, specifying which model was fine-tuned and how it was modified.

The template for the model card can be derived from the Model Documentation Form in the Transparency Chapter. The document should include the following sections further divided into corresponding fields:

- General Information: Legal name of the model provider, model name and version, evidence of authenticity, release dates, and details of dependencies if the model is a fine-tuned version of an existing one.
- Model Properties: Description of the architecture, design specifications, rationale for key design choices, input and output modalities (e.g. text, audio, image, etc.), and total number of parameters.
- Methods of Distribution and Licenses: Channels through which the model is made available, the type of license, and any additional assets such as training or inference code.
- Intended Use: Link or copy of the acceptable use policy, description of intended uses (e.g. productivity, enhancement, translation, creative content generation, data analysis, data, visualisation, programming assistance, scheduling, customer support, variety of natural language tasks, etc.), description of the AI systems in which the GPAI model might be integrated, instructions on how to integrate the GPAI model in AI systems; description of hardware and software needed to use the model.
- Training/Fine-Tuning process: a general description of the main steps or stages involved in the training/fine-tuning process, including methodologies and techniques, the key design choices and assumptions made.
- Data and resources: information on data used for training/fine-tuning, validation and testing, including data type and data provenance; a description of the process to obtain data, including the size of dataset together with a description of the main characteristic of test/validation such as domain, geography, language; a description of data processing involved in transforming the acquired data into training/fine-tuning, testing, and validation data for the model curation methods;

a description of strategies implemented in data acquisition so to filter unsuitable data (if any); a description of strategies implemented in data collection to bias detection (if any).

- Computational resources: a description of the time needed for training/fine-tuning together with hardware specifications; the measured or estimated amount of computation (e.g. number of floating-point operations) together with the description of the adopted measurement methodology.
- Energy consumption: a description of measured or estimated amount of energy used for training/fine-tuning, reported in Megawatt-hours recorded with at least two significant figures, together with the description of the adopted measurement methodology; a description of measured or estimated amount of energy used for benchmarking, reported in Megawatt-hours recorded with at least two significant figures, together with the description of the adopted measurement methodology.

## 5.2. High-Risk AI systems

### 5.2.1. Classification rules for high-risk AI systems

Unlike GPAI models, the applicability of obligations for high-risk AI systems does not depend on general criteria, but rather on specific use-case descriptions identified as high-risk.

Article 6 of the AI Act outlines two classification mechanisms relying on Annex I and Annex III. Accordingly, high-risk AI systems are:

- a) AI systems that are a product or a safety component of a product covered by the Union harmonisation legislation listed in Annex I, when a third-party conformity assessment is required under the listed legislation.

For instance, Annex I includes the Medical Device Regulation<sup>122</sup> and the Directive on the safety of toys<sup>123</sup>.

- b) AI systems referred to in Annex III.

This annex comprises 8 areas, under which specific use cases to be considered as high-risk are described.

While the high-risk systems in relation to Annex I belong to sectors where the EU market surveillance framework already applies, the areas designated in Annex III are newly integrated into this framework through the AI Act. It is therefore relevant to restate them in this report, particularly to assess whether the use cases envisioned in LLMs4EU may be subject to the AI Act requirements for high-risk AI systems. The 8 areas are:

<sup>122</sup> Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices.

<sup>123</sup> Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys.

- Biometrics
- Critical infrastructure
- Education and vocational training
- Employment, workers management and access to self-employment
- Access to and enjoyment of essential private services and essential public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Administration of justice and democratic processes

It is important to clarify that not all AI systems used in these areas are considered high-risk. In fact, the use cases identified for each area are quite specific. High-risk AI systems typically perform a task that materially influences the outcome of a decision-making process, in a situation where any failure or malfunctioning may pose a significant risk of harm to the health, safety and fundamental rights of natural persons<sup>124</sup>. The requirements for high-risk AI systems are therefore expected to concern only a narrow subset of AI systems, considering the broad range of possibilities AI offers for low-value and repetitive tasks. Nonetheless, it remains essential to assess each development project on a case-by-case basis, considering the high-risk classification from the outset, to ensure compliance requirements are anticipated and integrated into the project work plan.

### **5.2.2. Preliminary analysis of the potential high-risk classification of use cases to be developed in LLMs4EU**

The purpose of the LLMs4EU project is to adapt LLMs to serve the specific needs of the following five sectors, referred to as use-case domains:

- Science
- Public services
- Tourism
- Telecommunications
- Energy

The Science and Tourism use-case domains appear to be the least likely to fall under a high-risk classification. The Tourism domain's goal is to leverage cultural heritage for sustainable tourism development, which is distinct from any of the high-risk scenarios mentioned in the previous section. In the case of Science, not only does it not evidently

---

<sup>124</sup> This stems from the description provided in Article 6(3) of what shall not be considered a high-risk AI system.

correspond to any high-risk AI scenario, but the use cases to be developed may also fall outside the scope of the AI Act. As provided in Article 2(6), the AI Act “does not apply to AI systems or AI models, including their output, specifically developed and put into service for the sole purpose of scientific research and development”.

The Telecommunications and Energy use-case domains, however, must be examined in light of the “critical infrastructure” high-risk area, which more precisely designates “AI systems intended to be used as safety components in the management and operation of critical digital infrastructure, road traffic, or in the supply of water, gas, heating or electricity”<sup>125</sup>. While these two use-case domains may cover, respectively, the management of critical digital infrastructures and the supply of electricity, the condition that the system be used as a safety component considerably narrows the probability that the use-cases developed within the project will be concerned. Safety components of critical infrastructure are defined in recital 55 as “systems used to directly protect the physical integrity of critical infrastructure or the health and safety of persons and property, but which are not necessary in order for the system to function”. This recital clarifies that components intended to be used solely for cybersecurity purposes are not included in this definition and cites as examples systems for monitoring water pressure and fire alarm controlling systems in cloud computing centres. The fine-tuned models developed for these two use-case domains should therefore not fall under the high-risk classification as the intended use cases primarily aim to improve accessibility and interpretation of domain-specific regulatory and normative content across the EU, which constitutes a well-suited and low-risk application for an LLM.

The Public services use-case domain appears to be potentially more challenging, as multiple high-risk areas refer to activities that are usually performed by public bodies. It will therefore be necessary to verify, for each intended use-case, whether it aims to perform any tasks classified as high risk, especially those listed in Annex III under the areas of “access to and enjoyment of essential private services and essential public services and benefits”, “law enforcement”, “migration, asylum and border control management”, and “administration of justice and democratic processes”.

### 5.2.3. Obligations for high-risk AI systems regarding data

The provisions set out in Chapter III of the AI Act (Articles 6 to 49) apply **exclusively** to AI systems classified as high-risk. These include requirements such as establishing a risk

---

<sup>125</sup> Annex III (2), AI Act

management system<sup>126</sup>, and ensuring that the system achieves appropriate levels of accuracy, robustness, and cybersecurity<sup>127</sup>.

Regarding data, which is the main topic of this report, Article 10 of the AI Act requires the adoption of appropriate data governance and management practices, and sets quality criteria that training, validation, and testing datasets must meet.

Providers of high-risk AI systems must, for instance, be able to explain and justify the data collection processes used, the origin of the data, the original purpose of the collection if personal data is involved, and any data preparation operations that have been applied. Various assessments must be carried out, including regarding the availability, quantity, and suitability of the datasets needed, as well as potential biases that could negatively impact the health and safety of persons or fundamental rights.<sup>128</sup>

As for the required quality criteria, datasets must be “relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose”<sup>129</sup>. This means that any decision concerning data must be made with a thorough understanding of the context in which the system will operate, and that a very high standard of data quality is required.

The case of a high-risk AI system resulting from the fine-tuning of a GPT model presents additional challenges in complying with these requirements. The underlying foundation model is often developed by an initial provider who was not required to consider the specific obligations applicable to high-risk AI systems. Yet, the provider of the high-risk AI system must ensure compliance with all applicable obligations for the system as a whole, including its underlying components. On this matter, Article 25(2) provides that the “initial provider shall closely cooperate with new providers and shall make available the necessary information and provide the reasonably expected technical access and other assistance that are required for the fulfilment of the obligations set out in this Regulation, in particular regarding the compliance with the conformity assessment of high-risk AI systems”. Although this provision does attempt to address the need, it remains difficult to

---

<sup>126</sup> Article 9 AI Act

<sup>127</sup> Article 15 AI Act

<sup>128</sup> Article 10(2) AI Act “Those practices shall concern in particular: (a) the relevant design choices; (b) data collection processes and the origin of data, and in the case of personal data, the original purpose of the data collection; (c) relevant data-preparation processing operations, such as annotation, labelling, cleaning, updating, enrichment and aggregation; (d) the formulation of assumptions, in particular with respect to the information that the data are supposed to measure and represent; (e) an assessment of the availability, quantity and suitability of the data sets that are needed; (f) examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations; (g) appropriate measures to detect, prevent and mitigate possible biases identified according to point (f); (h) the identification of relevant data gaps or shortcomings that prevent compliance with this Regulation, and how those gaps and shortcomings can be addressed.”

<sup>129</sup> Article 10(3) AI Act

envision in practice, given that the initial provider may be exempted from this requirement simply by explicitly excluding the use of the GPAI system in a high-risk scenario<sup>130</sup>. As a result, such exclusions are likely to become common in terms of use and licenses, which considerably hinders the possibility to rely on a foundation model when developing a high-risk AI system.

### 5.3. Implementation timeline and evolving framework

The AI Act provisions applicable to the matters addressed in the previous sections take effect as follows:

- 2 August 2025
  - Obligations for providers of GPAI models placed on the market or put into service after that date.
- 2 August 2026
  - Provisions for High-Risk AI systems referred to in Annex III<sup>131</sup>.
- 2 August 2027
  - Obligations for providers of GPAI models placed on the market or put into service before 2 August 2025.
  - Provisions for High-Risk AI systems referred to in Annex I.

A key characteristic of this regulation is that the Commission is empowered to amend certain provisions through delegated acts. Notably, Annex III (list of high-risk AI systems), as well as Annexes XI and XII (technical documentation for GPAI models), may evolve over time in light of technological developments or to cover new high-risk scenarios identified.

This regulation also relies on various guidelines and templates to be provided by the Commission. These documents, which are being gradually published, are crucial in ensuring a consistent interpretation and implementation of the regulation, despite being non-binding<sup>132</sup> and also subject to change over time. While the guidelines on the scope of obligations for providers of GPAI models and the template for the public summary of training content for GPAI models were recently published and could be taken into account in this report, further guidelines are expected in the coming months, particularly to clarify provisions relating to High-Risk AI systems.

In addition, the GPAI Code of Practice also appears as a useful reference framework for GPAI model providers seeking to align with their obligations under the AI Act. Recently validated by the Commission and the AI Board, the voluntary Code offers practical

<sup>130</sup> Article 25(2) and Recital 86

<sup>131</sup> For High-Risk AI systems that have been placed on the market or put into service before 2 August 2026, the regulation applies only if, as from that date, those systems are subject to significant changes in their designs.

<sup>132</sup> An authoritative interpretation of the AI Act may only be given by the Court of Justice of the European Union.

guidance on transparency, copyright, and safety. The approaches described therein will shape the development steps undertaken in the project, regardless of whether they relate to applicable obligations, as the Code can be regarded as a standard for best practices in LLM training and deployment.

### AI act aspects in a nutshell

AI Act obligations may apply to LLMs under two main qualifications:

- General-Purpose AI model
- High-Risk AI system

The latter introducing, where applicable, an additional set of obligations.

**Conditions for an LLM to qualify as a GPAI model:** An AI model that is capable of competently performing a wide range of distinct tasks and displays significant generality qualifies as a GPAI model. The ability to generate language is presumed by nature to meet the first condition, whereas the criterion of significant generality is fulfilled if the compute required for training the model exceeds the threshold of  $10^{23}$  FLOP. Typically, an LLM with more than 1 billion parameters qualifies as a GPAI model under the AI Act.

**When do obligations for GPAI model providers apply to downstream modifiers?** A downstream modifier becomes the provider of the modified GPAI model only if the modification leads to a significant change in the model's generality, capabilities, or systemic risk. These conditions are met if the training compute used for the modification is greater than 1/3 of the training compute of the original model.

**What obligations apply to downstream modifiers that are considered GPAI model providers?** Obligations for providers of GPAI models are set out in Chapter V of the AI Act, in particular in Article 53. These notably include drawing up technical documentation and publishing a summary of training content. Templates are available to facilitate compliance with these requirements.

These obligations apply to downstream modifiers solely within the scope of the modifications they have carried out. For example, in the case of a fine-tuned LLM, the existing documentation of the foundation model must be complemented with information on the data used as part of the fine-tuning.

**High-risk AI systems:** AI systems can be classified as high-risk either because they are intended to be used in a context where a Union harmonisation legislation listed in Annex I of the AI Act applies, or because their intended use case is expressly designated as such among the 8 areas described in Annex III. Therefore, the situations in which the extensive requirements provided in Chapter III of the AI Act apply remain limited.

Achieving compliance of a fine-tuned LLM with these additional requirements poses a significant challenge: unlike obligations applicable to GPAI models, which mainly focus on transparency, high-risk AI systems must demonstrate a consistently high level of

quality on all aspects. Data quality, for example, has to be evidenced against various criteria that are not yet translated into practical expectations. Moreover, these quality requirements must be met by the system as a whole and therefore cannot be limited to the modifications introduced by the fine-tuning process, which complicates the selection of a foundation model.

**Key takeaways:**

- The amount of training compute needed is a decisive factor in determining when AI Act obligations apply. Estimating it as early as possible allows for better anticipation.
- Selecting a foundation model that already provides the required documentation for GPAI models makes compliance efforts easier for downstream modifiers.
- When specialising an LLM for a given use case, it is necessary to verify whether the use case falls within those listed and classified as high-risk. If so, a specific development process must be established to consider all additional obligations linked to this classification from the outset.

## 6. Conclusion

This deliverable provides guidance for data usage in the fine-tuning of large language models within the LLMs4EU project. The analysis identifies the regulatory obligations that project partners must navigate.

Regarding data protection compliance, partners must implement measures including the data minimisation and the storage limitation principle. The deliverable highlights that these requirements pose challenges in the context of model fine-tuning where personal data used for the training of the model may persist in the final model. Partners must additionally ensure compliance with data subject rights, including rights to access, rectification, and erasure, which present technical complexities when personal data has been incorporated into the models.

Regarding intellectual property rights, the deliverable analyses the legal foundation for using text and data mining exceptions under Articles 3 and 4 of the Copyright in the Digital Single Market Directive. The analysis also explores licensing strategies encompassing voluntary licensing arrangements and extended collective licensing schemes, which facilitate usage of copyrighted materials while providing fair remuneration to rights holders.

Regarding trade secret protection, the framework addresses obligations arising where data falls within the scope of the Trade Secrets Directive. The analysis identifies critical considerations for fine-tuning practitioners to ensure compliance with confidentiality requirements and avoid unlawful acquisition, use, or disclosure of protected information.

The AI Act implications are also highlighted, mainly regarding obligations for general-purpose AI model providers. The analysis further evaluates potential classification of

project use cases as high-risk AI systems, which may activate additional compliance requirements.

This deliverable is meant as a theoretical framework throughout the project life cycle. The analysis is non-exhaustive and should be contextualised by taking into account the regulatory evolutions that will occur during the project period.

Subsequent WP6 deliverables will address dataset-specific considerations and will provide detailed legal guidelines for data services and model production at Month 12. These documents will build upon this theoretical foundation to deliver practical implementation guidance tailored to the project's specific data sources and processing activities.

## 7. Bibliography

### 7.1. Academic references

- A. Guadamuz, 'The EU's Artificial Intelligence Act on Copyright' (10 November 2024) *The Journal of World Intellectual Property*, pp. 213-219.
- A. Guadamuz, "A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs," *GRUR International*, vol. 73, no. 2, 2024
- C. Geiger and V. Iaia, "The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI," *Computer Law & Security Review*, vol. 52, 2024,
- C. Novelli, et al. "Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity." *Computer Law & Security Review* 55 (2024)
- E. B., Van Veen, M. Boeckhout, I. Schlünder et al. "Joint controllers in large research consortia: a funnel model to distinguish controllers in the sense of the GDPR from other partners in the consortium"
- F. Cooper and James Grimmelmann, "The Files Are in the Computer: On Copyright, Memorization, and Generative AI," *arXiv*, 18 July 2024
- F. Cooper and others, "Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models," *arXiv*, 18 May 2025
- G. Cifrodelli and L. Edwards, "Copyright and Generative AI: What Can We Learn from Model Terms and Conditions?," *Kluwer Copyright Blog*, 24 July 2024
- G. Feretzakis, G.; Vagena, E.; Kalodanis, K.; Peristera, P.; Kalles, D.; Anastasiou, A. "GDPR and Large Language Models: Technical and Legal Obstacles." *Future Internet* 2025, 17, 151.
- H. Ruschemeier "Generative AI and data protection. Cambridge Forum on AI: Law and Governance" 2025
- I. Emanuilov and T. Margoni, "Forget Me Not: Memorisation in Generative Sequence Models Trained on Open-Source Licensed Code," *Zenodo*, February 2024
- J. E. Barbee, "Scratching the Surface of IP Rights: Data scraping and Artificial Intelligence", *NYSBA*, December 22, 2023.
- J. P. Quintais, "Generative AI, Copyright and the AI Act," *Computer Law & Security Review*, vol. 56, 2025
- L. Edwards and others, "Private Ordering and Generative AI: What Can We Learn from Model Terms and Conditions?" *CREATE*, 24 May 2024
- L. Soldaini, et al., 'Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research' (2024) <<https://arxiv.org/abs/2402.00159>>.
- M. Kretschmer, et al., "Copyright Law and the Lifecycle of Machine Learning Models", 55 *IIC – International Review of Intellectual Property and Competition Law* 110 (2024)

M. Sag, "Fairness and Fair Use in Generative AI," Fordham Law Review, vol. 92, 2024, p. 1887.

M. Senftleben, "Generative AI and Author Remuneration," IIC – International Review of Intellectual Property and Competition Law, vol. 54, 2023

N. Lee, "Protection for Artificial Intelligence in Personalised Medicine – The Patent / Trade Secret trade off",

N. Lucchi and others, "Generative AI and copyright" 2025,

N. Sousa e Silva, "Are AI Models' Weights Protected Databases?", Kluwer Copyright Blog, 18 January 2024.

O. Bulayenko and others, "AI Music Outputs: Challenges to the Copyright Legal Framework" Social Science Research Network, 28 February 2022

P. Keller and Z. Warso, "Defining Best Practices for Opting out of ML Training" (Open Future Foundation, 2023);

R. Chandrasekhar. "Legal frictions for data openness: Reflections from a case-study on re-use of the open web for AI training". Centre Internet et Société - CNRS; Inno3; Open Knowledge Foundation. 2025

S. Baack, 'A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl' (2024) The 2024 ACM Conference on Fairness, Accountability, and Transparency.

S. Longpre et al., 'A Large-Scale Audit of Dataset Licensing and Attribution' (2024) 6 Nature Machine Intelligence

S. Perlmutter, "One Year On: Developments in the Protection of Trade Secrets", U.S. Patent & Trademark Office (June 19, 2017)

T. Margoni, 'Saving Research: Lawful access to unlawful sources under Art. 3 CDSM Directive?' (December 22, 2023) Kluwer Copyright Blog.

T. W. Dornis, "The Training of Generative AI Is Not Text and Data Mining," European Intellectual Property Review, no. 2, 2025

## 7.2. Institutional references

CJEU - C-252/21 - Meta Platforms and Others (General terms of use of a social network)

CJEU, 15 January 2015, C-3/14 (Ryanair/PR Aviation)

CNIL, Ensuring the lawfulness of the data processing - In case of re-use of data, carrying out the necessary additional tests and verifications, 07 June 2024

CNIL, Fiches pratiques de l'IA, 2025, <https://www.cnil.fr/fr/les-fiches-pratiques-ia>

Copyright Initiative, Authors and Performers Call for Safeguards Around Generative AI (April 20, 2023), <https://urheber.info/media/pages/diskurs/call-for-safeguards-around->

generative-ai/069a7d264a-1697140342/authors-and-performers-call-for-safeguards-around-generative-ai\_20.4.2023.pdf

EDPB, Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR

EDPB, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models

European Commission, 2019, Ethics guidelines for trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

European Commission, Code of Practice for General-Purpose AI Models

European Commission, Commission Implementing Decision (EU) 2024/458 of 1 February 2024 on setting up the European Digital Infrastructure Consortium for the Alliance for Language Technologies (ALT-EDIC) [https://eur-lex.europa.eu/eli/dec\\_impl/2024/458/oj/eng](https://eur-lex.europa.eu/eli/dec_impl/2024/458/oj/eng)

European Commission, Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models required by Article 53 (1)(d) of Regulation (EU) 2024/1689 (AI Act) Annex C (2025) 5235 to the Communication to the Commission

European Commission, Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act, 18 July 2025

European Parliament, 'Generative AI and Copyright – Training, Creation, Regulation' (July 2025) Study, Requested by the JURI Committee PE 774.095, 43.

European Union Intellectual Property Office (EUIPO), Development of Generative Artificial Intelligence from a Copyright Perspective (EUIPO, 2025).

European Union, Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030

European Union, Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market (DSM Directive).

European Union, Directive 2001/29/EC on the harmonization of certain aspects of copyright and related rights in the information society.

European Union, Directive 2009/24/EC on the legal protection of computer programs.

European Union, Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys.

European Union, Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (GDPR).

European Union, Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices.

European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) (AI Act).

Knowledge Rights 21, 'Why the EU TDM Exceptions (Art. 3 & 4) Can Be Used for AI – Including Generative AI' (June 3, 2025) <<https://www.knowledgerights21.org/news-story/eu-tdm-exceptions-can-be-used-for-ai/>> accessed on July 4, 2025.

OECD, 'Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data' (OECD Artificial Intelligence Papers, February 2025)

United States District Court Northern District of California 25 June 2025, Case No. 23-cv-03417-VC (Kadrey / Meta Platforms).

W3C Community Group, TDM Reservation Protocol (TDMRep) (W3C Community Group, 2024).