

The AI Act and its impact on Large Language Models and the CLARIN Infrastructure

Pawel Kamocki

IDS Mannheim, Germany

kamocki@ids-mannheim.de

Anna Goslawska

Wroclaw University of Science and Technology, Poland

anna.goslawska@pwr.edu.pl

Henk van den Heuvel

Radboud University, Nijmegen, the Netherlands
h.vandenheuvel@let.ru.nl

Erik Ketzan

King's College London, UK

erik.ketzan@kcl.ac.uk

Krister Linden

University of Helsinki, Finland

krister.linden@helsinki.fi

Costanza Navarretta

University of Copenhagen, Denmark

costanza@hum.ku.dk

Andrius Puksas

Vytautas Magnus University Lithuania

andrius.puksas@vdu.lt

German Rigau

University of the Basque Country, Spain

german.rigau@ehu.eus

Abstract

The EU Artificial Intelligence Act (Regulation 2024/1689), which entered into force in August 2024, represents the world's first comprehensive regulatory framework for AI. While it explicitly excludes AI systems developed solely for scientific research and development, its implications for the CLARIN community are far-reaching, especially given the growing reliance on large language models (LLMs) and general-purpose AI (GPAI). This extended abstract examines the scope of the research exemption in the AI Act (Section 2), and provides a detailed overview of the obligations imposed on providers of GPAI models, including those with systemic risk (Section 3). Given CLARIN's role as a data provider, special attention is paid to the requirement for dataset documentation under the upcoming GPAI Code of Practice.

1 Introduction

EU Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, better known as the AI Act, is a recent addition to the legal framework affecting language data and language technology. The Act was officially proposed in 2021; the proposal was then substantially modified to better address the issues related to General-Purpose AI (Bertuzzi, 2023), which delayed its adoption. The Act entered into force on 1 August 2024, and it will become applicable, in principle, 24 months after that date, but with many exceptions (in particular, the rules on General-Purpose AI have entered into force on 2 August 2025).

The AI Act is, globally, the first comprehensive attempt to legally regulate AI (see e.g. Ruschemeier 2022, Wörsdörfer 2025). As such, it cannot go unnoticed by the CLARIN community. For the typical academic researcher employed by a university or research institute, the impact of the AI Act may, at first glance, appear to be negligible, as the Act “exclude[s] from its scope AI systems and models specifically developed and put into service for the sole purpose of scientific research and development”¹. But even with a clear exception for scientific research, the AI Act has the potential to impact the CLARIN community's sources of tools and methods, funding, collaborations with industry, corporations that may be hiring academics and students in the future, stakeholders beyond the typical academic researcher, and a myriad of end-users of CLARIN community members' research. An exploration of the impact of the AI Act on CLARIN-related activities is thus important.

This extended abstract will discuss the general exclusion of research from the scope of the AI Act (Section 2), before analysing the rules on General Purpose AI (a category in which LLMs fall) in greater detail (Section 3).

¹ Recital 16 and Article 2(6) of the AI Act.

2 General Exclusion of Research and Development Activities from the AI Act

One of the stated objectives of the AI Act is to promote innovation, safeguard academic freedom, and avoid hindering research and development activities. From the perspective of the CLARIN community, a particularly important provision of the AI Act is that it does not apply to AI systems and models—including their outputs—that are specifically developed and used exclusively for scientific research and development purposes. Furthermore, even in cases where systems or models are developed for broader applications, the AI Act does not extend to research, testing, and development activities carried out prior to their placement on the market. It is important to note, however, that this exemption does not include testing conducted under real-world conditions.

However, models and systems that are primarily designed for other purposes but are employed within research and development contexts still fall under the scope of the AI Act. Given that research does not operate in isolation from broader societal and technological developments, it is reasonable to expect that the AI Act will influence researchers—particularly when they develop or deploy models and systems intended for general use, or when they test such technologies in real-world conditions. Since CLARIN serves not only the academic community but also a broader range of stakeholders, it is essential for the CLARIN community to remain informed about the evolving regulatory landscape. The rapidly advancing capabilities of AI in solving complex tasks (see, e.g., Kaur et al., 2024) suggest that general-purpose or fine-tuned large language models (LLMs) may increasingly replace many of the specialized NLP tools and packages traditionally used by CLARIN researchers (see, e.g., Min et al., 2021). Therefore, understanding and analysing the implications of the AI Act for CLARIN-related activities is both timely and crucial.

3 General-Purpose AI Models under the AI Act

A general-purpose AI model (GPAIM) is defined 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” (Article 3, (63) of the AI Act).

The key characteristic of a GPAIM is its capability to perform a wide range of distinct tasks. Typically, such models are trained on large amounts of data using machine learning methods and can be made available for direct download, as physical copies, or through libraries or APIs. LLMs, and other large generative models, are considered GPAIMs (Recital 99 of the AI Act). In case of doubt, the generality of a model can also be determined by the number of its parameters – models with a billion or more parameters are presumed to be GPAIMs (recital 98 of the AI Act). Most modern Large Language Models, at least those released after 2019, meet this criterion; remarkably, BERT, originally released in 2018 is still significantly smaller (despite being considered an LLM)².

In the spirit of the general exclusion of research and development activities from the AI Act (see above), models used for research, development and prototyping and not available on the market are not considered GPAIMs.

The rules on high-risk AI systems and GPAIM are not mutually exclusive and can be cumulated: a GPAIM can be part of a high-risk AI system.

Within the category of GPAIM, the AI Act distinguishes two additional sub-categories: GPAIM with systemic risks (which are subject to a stricter framework) and “free open source” GPAIM.

The providers of all GPAIM need to comply with the four rather generally phrased obligations listed in [Article 53](#) of the AI Act, i.e.:

1. Prepare and keep up to date the technical documentation of the model, including its training and testing process and the results of its evaluation. The documentation should at the very least contain the information listed in Annex XI of the AI Act, and upon request be provided to the competent national authorities and the AI Office.
2. Prepare and keep up to date information and documentation that should be made available to the systems providers who want to integrate the GPAIM in their AI systems. It should give said providers a “good understanding of the capabilities and limitations” of the model, and enable them

² https://en.wikipedia.org/wiki/List_of_large_language_models -- last visit 31.03.2025.
CLARIN Annual Conference Proceedings, 2025

to comply with their obligations under the AI Act (e.g., for providers of high-risk AI systems, the obligation to risk management, transparency, and their own documentation obligations). At the very least, the documentation should contain the information listed in Annex XII of the AI Act;

3. Put in place a policy to comply with copyright and related rights, and in particular to respect the opt-out (reservation of rights) under the “general” TDM exception (Article 4(3) of the [DSM Directive](#));
4. Prepare and make publicly available a sufficiently detailed summary about the datasets used for training the model.

Obligations 1. and 2. do not apply to GPAIM released under free and open source licenses, and whose parameters (including the weights), information about their architecture, and the information on model usage are publicly available (Article 53(2) of the AI Act), unless these models qualify as GPAIMs with systemic risks (see below).

A GPAIM with systemic risks is a GPAIM that meets one of the following conditions:

- a) it has high impact capabilities evaluated on the basis of appropriate technical tools and methodologies OR
- b) It has been officially recognised as such in a Commission decision.

The criteria mentioned in a) will be fine-tuned by the Commission in a delegated act. The AI Act only mentions one criterion: a GPAIM is presumed as carrying a systemic risk “when the cumulative amount of computation used for its training measured in floating point operations is greater than 10^{25} ”. Over 25 modern GPAIMs, including GPT-4, are believed to meet this threshold³.

Providers of GPAIM with systemic risk need to comply with additional obligations under Article 55 of the AI Act:

1. Perform model evaluations, including adversarial testing to identify and mitigate systemic risk.
2. Assess and mitigate possible systemic risks, including their sources.
3. Track, document and report serious incidents and possible corrective measures to the AI Office and national competent authorities;
4. Ensure an adequate level of cybersecurity protection.

One can assume that currently only very large LLMs will be considered as GPAIMs with systemic risks; to what extent this framework will be relevant for CLARIN remains to be seen.

The rules regarding GPAIM became applicable on 2 August 2025. However, according to Article 111 of the AI Act Providers of general-purpose AI models that have been placed on the market before 2 August 2025 shall take the necessary steps in order to comply with the obligations laid down in the AI Act by 2 August 2027.

In recent weeks and months, these rather abstractly formulated lists of obligations have gradually been filled up with practical meaning - some elements are already in place, others are still being developed. To a large extent, this is due to the adoption of the GPAI Code of Practice, whose final version was published on July 10, 2025 (<https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>). It is complemented by Commission guidelines on key concepts related to GPAIM (<https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act>). The Code contains, among other interesting elements, a Model Documentation form for GPAIM, which may seem surprisingly simple and straightforward, totaling only 3 pages.

CLARIN, as a provider of data for GPAIM (LLM) training, should also pay special attention to the details concerning the obligation laid down in Article 53(1)(d) of the AI Act, i.e. to provide a sufficiently detailed summary of the training data. In July 2025, the European Commission published an Explanatory Notice and a standardized Template for the Public Summary of Training Content for general-purpose AI models to support providers in meeting their transparency obligations under Article 53(1)(d) of the AI Act (<https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models?utm>).

It may be also worthwhile to consider providing descriptions of CLARIN datasets following the template in the final version of the GPAI Code of Practice:

³ <https://epoch.ai/data-insights/models-over-1e25-flop> (last visit: 11.03.2025)
CLARIN Annual Conference Proceedings, 2025

Information on the data used for training, testing, and validation				AIO	NCAs	DPs			
Data type/modality: <i>Select all that apply.</i>	☐ Text	☐ Images	☐ Audio	☐ Video	If any other please specify:		☒	☒	☒
Data provenance: <i>Select all that apply</i>	☐ Web crawling		☐ Private non-publicly available datasets obtained from third parties		☐ User data		☒	☒	☒
For definitions of each listed category, see the Template for the Public Summary of the Training Content of General-Purpose AI models provided by the AI Office	☐ Publicly available datasets		☐ Data collected through other means		If any other please specify:				
	☐ Synthetic data that is not publicly accessible (when created directly by or on behalf of the provider)								
How data was obtained and selected:	A description of the methods used to obtain and select training, testing, and validation data, including methods and resources used to annotate data, and models and methods used to generate synthetic data where applicable. For data previously obtained from third parties, a description of how the provider obtained the rights to the data if not already disclosed in the public summary of training data published in accordance with Article 53(1), point (d). <i>[Recommended 300 words]</i>						☒	☒	☐
Number of data points:	The size (in number of data points) of the training, testing, and validation data respectively, together with the definition of the unit of data points (e.g. tokens or documents, images, hours of video or frames), recorded with at least one significant figure (e.g. 3×10^{13} tokens).						☐	☒	☐
	The size (in number of data points) of the training, testing, and validation data respectively, together with the definition of the unit of data points (e.g. tokens or documents, images, hours of video or frames), recorded with at least two significant figures (e.g. 1.5×10^{13} tokens).						☒	☐	☐
Scope and main characteristics:	A general description of the scope and main characteristics of the training, testing and validation data, such as domain (e.g. healthcare, science, <u>law</u>), geography (e.g. global, restricted to a certain <u>region</u>), language, modality coverage, where applicable. <i>[Recommended 200 words]</i>						☒	☒	☐
Data curation methodologies:	General description of the data processing involved in transforming the acquired data into training, testing, and validation data for the model, such as cleaning (e.g. filtering out irrelevant content such as advertisements), <u>normalisation</u> (e.g. tokenizing), augmentation (e.g. back-translation). <i>[Recommended 300 words]</i>						☒	☒	☒
Measures to detect unsuitability of data sources:	A description of any methods implemented in data acquisition or processing, if any, to detect the presence of unsuitable data sources considering the model's intended uses, including but not limited to illegal content, child sexual abuse material (CSAM), non-consensual intimate imagery (NCII), and personal data leading to its unlawful processing. <i>[Recommended 400 words]</i>						☒	☒	☐
Measures to detect identifiable biases:	A description of any methods implemented in data acquisition or processing, if any, to address the prevalence of identifiable biases in the training data. <i>[Recommended 200 words]</i>						☒	☒	☐

References

- Bertuzzi, Luca. 2023. "Leading EU lawmakers propose obligations for General Purpose AI". Euractiv.com, <https://www.euractiv.com/section/artificial-intelligence/news/leading-eu-lawmakers-propose-obligations-for-general-purpose-ai/> (last visit: 11.04.2025).
- Floridi, Luciano, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. 2022. "capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act." *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4064091>.
- Kaur, Pravneet, Gautam Siddharth Kashyap, Ankit Kumar, Md Tabrez Nafis, Sandeep Kumar, and Vikrant Shokeen. 2024. "From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility." arXiv. <https://doi.org/10.48550/ARXIV.2402.16142>.
- Min, Bonan, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. "Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey." arXiv. <https://doi.org/10.48550/ARXIV.2111.01243>.
- Ruscheimer, Hannah. 2023. "AI as a Challenge for Legal Regulation – the Scope of Application of the Artificial Intelligence Act Proposal." *ERA Forum* 23 (3): 361–76. <https://doi.org/10.1007/s12027-022-00725-6>.
- Wörsdörfer, Manuel. 2025. "The E.U.'s Artificial Intelligence Act: An Ordoliberal Assessment." *AI and Ethics* 5 (1): 263–78. <https://doi.org/10.1007/s43681-023-00337-x>.