

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

Paweł Kamocki
IDS Mannheim, Germany
kamocki@ids-mannheim.de

Joanna Blochowiak
University of Zurich,
Switzerland
joanna.blochowiak@uzh.ch

Anna Gosławska
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 paper discusses the scope of research exemptions in the AI Act (Section 2), and provides an overview of the obligations related to transparency of AI-generated text outputs (Section 3), as well as those imposed on providers of General-Purpose AI models, including those with systemic risk (Section 4). Given CLARIN's role as a data provider, special attention is paid to the requirement for dataset documentation.

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 AI Act entered into force on 1 August 2024. While most of its provisions will apply from 2 August 2026 (i.e. 24 months after entry into force), the Act provides for a phased application with several earlier deadlines. In particular, the obligations relating to general-purpose AI apply from 2 August 2025.

The AI Act is, globally, the first comprehensive attempt to legally regulate AI (see e.g. Ruschmeier 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 exemption for scientific research, the AI Act has the potential to impact the CLARIN community's tools and methods, funding, collaborations with industry, corporations that may be hiring academics and students in the

¹ Recital 25 and Article 2(6) of the AI Act.

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 paper will discuss the general exclusion of research from the scope of the AI Act (Section 2), before analysing the rules on Transparency (Section 3) General Purpose AI (a category in which LLMs fall) in greater detail (Section 4).

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

One of the objectives of the AI Act, expressly stated in its Recital 25, is to promote innovation, safeguard freedom of science, and avoid hindering research and development activities. One of the provisions aimed at this objective, particularly important for the CLARIN community (albeit not included in the draft – see Smuha et al. 2021), can be found in Article 2(6) of the AI Act, which contains a general scientific research exemption. According to this provision, 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”. In addition to this, Article 2(8) provides that the AI Act “does not apply to any research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service”. This indicates that models and systems specifically developed for “scientific research” (as well as their outputs) are fully exempted from the GDPR, whereas commercial “research and development” activities are exempted only until the system or model concerned is “placed on the market” (i.e., first made available on the EU market – cf. Article 3(9) of the AI Act) or “put into service” (i.e. first supplied for use on the EU market – cf. Article 3(11) of the AI Act), i.e. exploited commercially. One can easily imagine that this two-speed solution may be difficult to apply in practice, especially in public-private partnerships, or within research departments of commercial companies (Finck, 2025).

In practice, AI 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 Transparency of AI-generated Text Outputs under the AI Act

The rules of transparency of AI outputs are among the most interesting provisions of the AI Act, both for theoreticians and for practitioners. The language community should pay particular attention to Articles 50(2) and 50(4), as they both involve text outputs. According to Article 50(2), providers of AI systems (including general-purpose AI systems) generating synthetic audio, image, video or *text* content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Article 50(4) provides *inter alia* that deployers of an AI system that generates or manipulates text which is published with the purpose of informing the public on matters of public interest (hereinafter: “news”) shall disclose that the text has been artificially generated or manipulated. The latter obligation does not apply when the text has undergone human review or editorial control, and is published under the editorial responsibility of a legal or natural person.

First of all, it should be stressed that the transparency obligation concerning all AI-generated content weighs on the providers, whereas the more specific obligation concerning “news” weighs on deployers. It is therefore crucial to explain the difference between the two categories of stakeholders: providers, according to Article 3(3) of the AI Act, are the entities who develop an AI system or have it developed and who exploit it under their own name (e.g., OpenAI

is the provider of ChatGPT). Deployers, on the other hand, are the persons or entities who use an AI system in the course of their activity, except when this activity is both personal and non-commercial. In other words, a deployer can be described as a “professional user”, e.g., a university or a publishing house using ChatGPT are deployers, and so is an individual researcher using ChatGPT for purposes other than strictly personal.

It is therefore the provider that has to make sure, under Article 50(2), that the outputs generated by an AI system are marked in a machine-readable format; in addition, under Article 50(4), the user (excluding personal non-commercial users) that uses an AI system to generate or manipulate “news” shall disclose this information to the public.

Both provisions enter into application in August 2026, raising significant practical challenges. In anticipation of these difficulties, in September 2025 the AI Office (a part of the European Commission) launched a process of drafting a Code of Practice on Transparency of AI-Generated Content. The first draft of the Code was published on 17 December 2025². Members of two dedicated working groups admit that at this stage the Code remains “high-level and broad”, but it proposes some interesting measures. Once approved by the Commission, the Code will serve as a voluntary tool for providers and deployers of generative AI systems to demonstrate compliance with their respective obligations regarding transparency.

The draft Code obliges AI providers who adhere to it to ensure that AI-generated or manipulated text is marked with an “imperceptible watermark”, directly interwoven within the content in a manner that is difficult for it to be separated from the content, and that withstands typical processing steps that may be applied to the content (Section 1, Sub-measure 1.1.2.). The watermark can be embedded during model training, model inference, or within the output of an AI model or system. To address the deficiencies of this method, AI providers will also establish and maintain logging facilities that allow for checking whether a text output has been generated or manipulated by their generative AI system (Section 1, Sub-measure 1.1.3.). “Logging” is defined in the draft Code as “verbatim recording and indexing of content” which “can be used for a fast lookup of known content”. As an additional measure, AI providers who sign the Code will be obliged to implement a digitally signed manifest, allowing deployers to obtain a certified version of the AI output generated or manipulated by their AI system or model to formally guarantee the origin of content that does not allow secure embedding of metadata. This provenance certificate will enable deployers and end-users to provide third parties with guarantees that the content is AI-generated or manipulated, linking it back to the specific generative AI system or model” (Section 1, Sub-measure 1.2.1.).

Regarding the transparency obligations of deployers, they are also included in the draft Code. In particular, the deployers should agree on a common taxonomy, including clear definitions of what constitutes AI-generated and AI-assisted “news”. According to the draft Code (Section 2, Measure 1.1), AI-generated “news” are “fully and autonomously generated by the AI system without human authored authentic content” (in other words, generated from a prompt), whereas AI-assisted “news” are those with “mixed human and AI involvement (...), where the AI system substantially impacts the content” of the “news”. AI involvement may consist of, *inter alia*, “adding AI-generated or manipulated text to existing human-authored text”, or “AI-enabled alterations to description of events, actors, arguments or interpretation”. Furthermore, signatories of the Code will set up internal procedures to identify AI-generated and AI-manipulated “news” (Section 2, Measure 5.1), and documentation of the human review process (Section 2, Measure 5.2). For “news” concerned by the Article 50(4), the signatories will disclose the use of AI by placing a standardised icon in a “fixed, clear and distinguishable position”, which could include but is not limited to placing the icon “at the top of the text, beside the text, in the footer or after the closing sentence of the text”.

4 General-Purpose AI Models under the AI Act

The rules on General-Purpose AI Models (GPAIM), which potentially include many language models, are found in Chapter V of the AI Act.

² <https://digital-strategy.ec.europa.eu/en/library/first-draft-code-practice-transparency-ai-generated-content>

In recent months, the rather abstractly formulated lists of obligations of GPAIM providers have gradually been filled up with practical meaning – some elements are already in place. To a large extent, this is due to the adoption of the GPAI Code of Practice, whose final version was published on July 10, 2025³. It was complemented by Commission guidelines on key concepts related to GPAIM⁴. The Code contains, among other interesting elements, a Model Documentation Form for GPAIM, which may seem surprisingly simple and straightforward, totaling only 3 pages.

The rules regarding GPAIM became applicable on 2 August 2025. However, according to Article 111 of the AI Act, providers of GPAIMs 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.

4.1. Concept and typology of General-Purpose AI Models

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)⁵.

According to the Commission’s guidelines (para. 17), an indicative criterion for a model to be considered a general-purpose AI model is that its training compute is greater than 10²³ FLOP and it can generate language (whether in the form of text or audio), text-to-image or text-to-video.

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.

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.

4.2. Obligations of General-Purpose AI providers

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. (see below in 4.2.1)
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. (see below in 4.2.2)
3. Put in place a policy to comply with copyright and related rights (see below in 4.2.3)

³ <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>

⁴ <https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act>

⁵ https://en.wikipedia.org/wiki/List_of_large_language_models -- last visit 31.03.2025.

4. Prepare and make publicly available a sufficiently detailed summary about the datasets used for training the model. (see below in 4.2.4).

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.

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 (FLOPs) is greater than 10^{25} ”. Over 25 modern GPAIMs, including GPT-4, are believed to meet this threshold.

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.

4.2.1. Technical documentation of General-Purpose AI Models

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. The Annex XI lists the following items to be included in the documentation:

- 1. A general description of the general-purpose AI model including:
 - (a) the tasks that the model is intended to perform and the type and nature of AI systems in which it can be integrated;
 - (b) the acceptable use policies applicable;
 - (c) the date of release and methods of distribution;
 - (d) the architecture and number of parameters;
 - (e) the modality (e.g. text, image) and format of inputs and outputs;
 - (f) the licence.

2. A detailed description of the elements of the model referred to in point 1, and relevant information of the process for the development, including the following elements:
 - (a) the technical means (e.g. instructions of use, infrastructure, tools) required for the general-purpose AI model to be integrated in AI systems;
 - (b) the design specifications of the model and training process, including training methodologies and techniques, the key design choices including the rationale and assumptions made; what the model is designed to optimise for and the relevance of the different parameters, as applicable;
 - (c) information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies (e.g. cleaning, filtering, etc.), the number of data points, their scope and main characteristics; how the data was obtained and selected as well as all other measures to detect the unsuitability of data sources and methods to detect identifiable biases, where applicable;
 - (d) the computational resources used to train the model (e.g. number of floating point operations), training time, and other relevant details related to the training;
 - (e) known or estimated energy consumption of the model.

4.2.2. Documentation for AI system providers

The documentation should give said providers a “good understanding of the capabilities and limitations” of the model, and enable them 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, i.e.:

1. A general description of the general-purpose AI model including:
 - the tasks that the model is intended to perform and the type and nature of AI systems into which it can be integrated;
 - the acceptable use policies applicable;
 - the date of release and methods of distribution;
 - how the model interacts, or can be used to interact, with hardware or software that is not part of the model itself, where applicable;
 - the versions of relevant software related to the use of the general-purpose AI model, where applicable;
 - the architecture and number of parameters;
 - the modality (e.g. text, image) and format of inputs and outputs;
 - the licence for the model.
2. A description of the elements of the model and of the process for its development, including:
 - (a) the technical means (e.g. instructions for use, infrastructure, tools) required for the general-purpose AI model to be integrated into AI systems;
 - (b) the modality (e.g. text, image, etc.) and format of the inputs and outputs and their maximum size (e.g. context window length, etc.);

- (c) information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies.

4.2.3. Copyright policy

GPAIM providers are obliged to put in place a policy to comply with Union law on copyright and related rights, and in particular to respect the opt-out (reservation of rights) under the “general” text and data mining exception (Article 4(3) of the [DSM Directive](#)). According to the GPAI Code of Conduct, the policy should include a commitment to reproduce and extract only lawfully accessible copyright-protected content when crawling the World Wide Web, to identify and comply with rights reservations, and to designate a point of contact and enable the lodging of complaints.

4.2.4. Summary of the training data

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 obligations⁶.

As a part of the detailed summary of the training data, the model provider must provide, among other things, the following information (below are described only selected elements that are considered particularly relevant):

1) Modalities, overall training data size and other characteristics

This section requires general information about the overall training data after pre-processing and before the training of the model. The following information must be presented: (i) modalities present in the training data (text, image, audio, video, other), (ii) training data size - the provider must select the range within which the estimated total training data size for that modality falls (tokens for text, hours for audio and video, number for images), (iii) types of content (description of the type of content that has been included in the training data)

In this section, it is also necessary to provide inter alia the description of the linguistic characteristics of the overall training data (for example the languages covered by the training data focusing in particular on EU official languages and other relevant characteristics of the overall training data (for example national/regional or demographic specificities of the training data where readily available).

2) List of data sources

This section requires information on the specific data sources used to train the general-purpose AI model. For the purposes of this section, a “dataset” refers to a single, pre-packaged collection of data.

The data sources should be presented according to the following categories: (i) publicly available datasets, (ii) private non-publicly available datasets obtained from third parties, (iii) data crawled and scraped from online sources, (iv) user data, (v) synthetic data, and (vi) other sources of data.

With regard to publicly available datasets, the GPAI model provider is required to list only those datasets classified as large. A dataset is considered to be “large” if the total data size for any one of the modalities contained in the dataset exceeds 3% of the size of all publicly available datasets for that modality used for training. Dataset size must be based on the post-processed dataset (for example after filtering) and not on artificially split subsets. For other publicly

⁶ <https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models?utm>

available datasets that are not listed as large ones, model provider should provide a general description of their content.

3) Data processing aspects

a) Respect of reservation of rights from text and data mining exception or limitation

This section requires the model provider to describe the measures implemented to respect reservations of rights under the TDM exception or limitation before and during data collection. In particular, the provider should explain the opt-out protocols and technical or organisational solutions honoured, whether applied directly by the provider or, where applicable, by third parties from whom the datasets were obtained.

In this context, it should be noted that certain dataset providers collect data using identifiable web crawlers and publish clear, practical instructions for preventing such collection (e.g., by specifying how the crawler can be blocked through standard access-control mechanisms such as the Robot Exclusion Protocol (robots.txt), as specified in the Internet Engineering Task Force (IETF) Request for Comments No. 9309). See for example <https://commoncrawl.org/ccbot>

b) Removal of illegal content

This section concerns measures taken to avoid or remove illegal content under Union law from the training data (e.g., blacklists, keywords, and model-based classifiers), without requiring disclosure of internal business practices or trade secrets. These measures are advisable where training data is likely to include illegal or unlawful content—particularly child sexual abuse material, terrorist content, and non-authorised use of material protected by intellectual property rights. Such measures do not include data selection practices to increase the model’s capability. In the General-Purpose AI Code of Practice published by the European Commission, the Copyright chapter envisages publishing, on an EU website, a list of links to websites that have been determined by EU or EEA courts or authorities to be “persistently and repeatedly infringing copyright and related rights on a commercial scale.” As of now such a list has not been published yet.

If a dataset has already been filtered by its provider to remove illegal or unlawful content before training, it would be particularly valuable for AI model providers because it reduces legal and compliance risks and increases the reliability of the data used.

4.3. Concept of Downstream Modifiers of General-Purpose AI Models

In the context of GPAIMs, downstream modifiers constitute an interesting case, especially in scientific setups. Downstream modifiers are entities or actors that take an existing general-purpose AI model or system and modify it for a specific downstream use or application. In this context, ‘upstream’ refers to the original developer or provider of a GPAIM, such as a foundation model, while ‘downstream’ describes the later stages of the value chain, where the model is adapted, integrated, or deployed for specific, concrete tasks. Typical tasks of downstream modifiers include (i) fine-tuning or further training of a GPAIM on new or domain-specific data; (ii) adapting the model through techniques such as prompting, reinforcement learning, or parameter adjustments; (iii) integrating the model into products, services, or workflows; (iv) changing or extending functionality, performance characteristics, or intended use.

4.3.1. When a downstream modifier becomes a GPAIM

From a technical standpoint, an indicative criterion for when a downstream modifier is considered a provider of a GPAIM relates to the training compute used to perform the modification: when this modification requires more than one-third of the training compute used for the original model, the downstream modifier is considered to be a GPAIM provider. Where original compute metrics are unavailable, if the original model presents a systemic risk, the threshold should be one-third of 10^{25} FLOPs; otherwise, it should be one-third of 10^{23} FLOPs (cf. Commission Guidelines on the scope of the obligations for providers of general-purpose AI models).

From a regulatory perspective, two articles of the EU AI Act are the most relevant for downstream modifiers and for GPAI models in general – namely, Articles 2(6) and 2(8), listed below:

1) Article 2(6) – Scientific research exemption

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.

2) Article 2(8) - Pre-Market research exemption

The AI Act does not apply to any research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service.

To illustrate how these articles apply to downstream modifiers that become GPAIM providers, consider a hypothetical scenario where a team of researchers from the Swiss national consortium CLARIN-CH undertakes the development of a comprehensive multilingual model for Swiss language varieties. Building upon an existing foundation model such as Llama 3.1 70B, the researchers conduct extensive continued pre-training using their curated collections of Swiss German dialects, Romansh language corpora, and historical Swiss multilingual parliamentary proceedings. The training process incorporates approximately 2 billion tokens drawn from digitized archives, contemporary media sources, and linguistic fieldwork repositories. This computational effort requires resources equivalent to 40% of the original model's training compute, exceeding the one-third threshold established in the Commission Guidelines. Therefore, under the EU AI Act, the researchers responsible for the development of such a model no longer fall under the category of downstream modifiers, but under that of providers of a new GPAIM.

Following standard academic practice, the researchers publish their findings in a peer-reviewed venue and, in the interest of scientific reproducibility, upload the fine-tuned model weights to Hugging Face, which is a widely-used repository for machine learning models. The act of uploading to repositories such as Hugging Face or Github is not clear from a regulatory perspective because this may be interpreted either as research dissemination or as placing on the market. If the model release is interpreted as market placement, the research would no longer qualify for the Article 2(6) research exemption, despite the fact that the model was developed within an academic context and served primarily scientific objectives. This classification would imply obligations stated in Article 53, including technical documentation requirements, implementation of copyright compliance policies, and provision of training content summaries. At present, there is uncertainty surrounding this kind of case, which demonstrates the need for Commission guidance on how open science practices (particularly model sharing for reproducibility purposes) should be reconciled with the AI Act concerning the market placement of the models.

4.3.2. When a downstream modifier remains a downstream modifier

A downstream modifier is not automatically a GPAIM provider. The Commission makes clear that not every modification of a GPAIM will lead to the downstream modifier being considered a provider of GPAIM. They only become the provider of a (new) GPAIM under specific conditions regarding training compute, i.e., when the modification requires more than one-third of the training compute used for the original model (see section above).

For purposes of illustration, consider one more hypothetical scenario in which the CLARIN-CH undertakes a domain-specific adaptation of an existing large language model for humanities research purposes. The consortium research team initiates a project to develop specialized capabilities for processing historical Swiss German texts from the 17th-19th centuries, drawing upon their extensive digitized archival collections. Rather than training a model from scratch, the research team uses Llama 3.1 70B as a foundation and conducts targeted fine-tuning using approximately 500 million tokens of historical texts curated from their repositories. The computational resources required for this fine-tuning operation represent roughly 20% of the original model's training compute. Since this modification falls below the critical one-third threshold established in the Commission Guidelines, CLARIN-CH does not cross the indicative criterion that would designate them as the provider of a new general-purpose

AI model. Consequently, the consortium maintains its status as a downstream modifier rather than becoming a GPAI model provider under the EU AI Act.

As stated previously, Article 2(6) of the AI Act explicitly excludes AI models developed and put into service for the sole purpose of scientific research and development, provided they are not placed on the market. In this scenario, these conditions are satisfied, since the modified model serves exclusively academic research in humanities, distribution occurs only within the authenticated CLARIN research infrastructure, and no commercial offering or market placement takes place. Documentation clearly establishes the research-only purpose of the development. As in the previous scenario, potential complications arise if CLARIN-CH considers broader dissemination strategies, like uploading the model to repositories, such as Hugging Face. Even though the decision of uploading to such a repository is consistent with open science principles, it may still be interpreted either as research dissemination or as placing on the market (see Article 2(8) of the AI act).

Despite the lack of consensus on the legal requirements concerning the fine boundary between placing a model on the market and putting a model into service for research purposes only, CLARIN can already implement several best practices. These include maintaining comprehensive documentation of the modification's research purpose and implementing robust access controls through existing federated authentication infrastructure. Additionally, CLARIN can use its established metadata infrastructure to document training data sources, licenses, and model capabilities, as stated in the transparency objectives of Article 53 (see above). These proactive measures align with CLARIN's longstanding commitment to the FAIR principles and put the consortium in a position to smoothly adapt to regulatory requirements once they are clarified by the Commission.

5 Conclusion

While the AI Act formally excludes AI systems developed solely for scientific research from its scope, this exemption should not be read as insulating the CLARIN community from the AI Act's practical and strategic consequences. As this paper has shown, the Act's transparency obligations and its regulatory framework for general-purpose AI will shape the environment in which language data and language technologies are developed, shared, funded, and used. For CLARIN-related activities, the relevance of the AI Act lies less in immediate compliance duties for academic researchers than in its indirect effects on research infrastructures, collaborations with industry, downstream uses of research outputs, and the expectations placed on future professionals trained within the community. Engaging proactively with the AI Act is therefore not merely a matter of legal awareness, but a necessary step in safeguarding the sustainability, openness, and societal legitimacy of language resources and technologies in an increasingly regulated AI ecosystem.

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>.
- Finck, Michèle, In Search of the Lost Research Exemption: Reflections on the AI Act, GRUR International, Volume 74, Issue 10, October 2025, Pages 903–904, <https://doi.org/10.1093/grurint/ikaf100>
- 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>.

Smuha, Nathalie A. and Rengers, Emma and Harkens, Adam and Li, Wenlong and MacLaren, James and Piselli, Riccardo and Yeung, Karen, How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act (August 5, 2021). Available at SSRN: <https://ssrn.com/abstract=3899991> or <http://dx.doi.org/10.2139/ssrn.3899991>

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>.