Statement of the Consultation for the Commission Guidelines to clarify the scope of the obligations of providers of General-Purpose AI Models in the AI Act

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

22. Mai 2025

Contact

- Patrick Brunner, patrick.brunner@fiz-karlsruhe.de

About FIZ Karlsruhe

FIZ Karlsruhe - Leibniz Institute for Information Infrastructure is one of the leading addresses for scientific information and services and a member of the Leibniz Association. Our core tasks are to provide science and industry with professional research and patent information and to develop innovative information infrastructures, with, for example, a focus on research data management, knowledge graphs and digital platforms. To this end, we conduct our own research, cooperate with renowned universities and research associations and are internationally and interdisciplinarily networked. FIZ Karlsruhe is a non-profit limited liability company and one of the largest non-academic institutions of its kind.

Question 1)

Question: Many entities will have to assess the general-purpose nature of their models to determine whether they need to follow the obligations for providers of general-purpose AI models. A pragmatic metric is thus highly desirable to limit the burden, especially on smaller entities. Do you agree that training compute is currently the best metric for assessing generality and capabilities, despite its various shortcomings?

Answer: No.

The use of FLOPs as a potentially sole criterion for determining whether a model qualifies as a GPAI raises diverse concerns. Not at least for this reason, Annex XIII of the AI Act provides a list of criteria on the basis of which a model is classified as a model with systemic risks after a balancing of interests. Nothing different should apply to the definition of a model as GPAI. The following text is mainly based on Large Language Models (LLMs), as they are paradigmatic of generally usable models. Specifically:

First, in practice, the FLOP threshold is already undercut, even though a model is generally considered capable. For example, the o4-mini model is said to have around 8 billion parameters, which means that at least 4_10^11 tokens would have had to be used for its training in order to fall under the term GPAI at all based on the commonly used formula. It is provided by OpenAI, whose products are paradigmatic for GPAI.

Secondly, the practical feasibility of the metric is questionable. This follows, on the one hand, from the (un)availability of the necessary metrics and, on the other hand, from everything that has to potentially be added to the used compute. A sample revealed, for example, that the number of parameters is usually easy to determine (since it is often stated as a proxy for model quality), but the number of used tokens is already more difficult to determine. This is also shown, for example, by the leading model on the leaderboard of the popular machine learning platform Huggingface before the leaderboard was closed (https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/ last access 16.05.2025).

Furthermore, often used formula for training compute can be inaccurate. This is because it is unclear how far preparatory steps count towards the limit like scraping or Knowledge Distillation. In this technique, a large, computationally intensive model is used to train a smaller model specifically. This results in high performance even of the smaller model (Hinton, G. et al., Distilling the Knowledge in a Neural Network, https://doi.org/10.48550/arXiv.1503.02531). An example of this is the Llama 4 “Behemoth” model, which exists specifically for this purpose (https://www.llama.com/models/llama-4/, last access: 16.05.2025). In this aspect, the question already arises as to whether the performance used for training these large models should also be attributed to the training of the small models. This question has further-reaching implications, as many providers scrape the internet to collect training data. The output of LLMs and other generative models is increasingly found there, which raises the question as to whether their training performance should be attributed to the models trained or refined on these

data sets. Nothing different applies if models are used for data preparation, for example to annotate the training data.

Another problem also emerged with the technique of distillation. Because thirdly, there are techniques that can significantly increase the performance of models without the performance used being directly attributed to the trained model. In addition to the already discussed distillation and its implication, Retrieval Augmented Generation (RAG) should be mentioned. Here, LLMs are passed relevant information or data in the context of a prompt, which significantly improves the result (Lewis, P. et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, https://doi.org/10.48550/arXiv.2005.11401). This can also allow models that do not meet the proposed definition of GPAI to achieve similar or even better performance. The same applies to n-shot learning, in which the model receives relevant examples for a prompt (Brown, T. B. et al., Language Models are Few-Shot Learners, https://doi.org/10.48550/arXiv.2005.14165).

Fourthly, models could and probably will be optimized by providers in such a way that they fall below the GPAI limit. For such optimizations exist various techniques, as already mentioned (also, see Pistillo, M.; Villalobos, P, Defending Compute Thresholds Against Legal Loopholes, https://doi.org/10.48550/arXiv.2502.00003) Consequently, fifthly, ongoing adjustments to the limit would be necessary, even if only in the short term, with regard to a metric which is costly to determine, especially by supervisory authorities and the courts. This leads to the conclusion that the metric of used FLOPs is ineffective and disproportionately time-consuming compared to other approaches. This also results in a deficit in enforcement of the AI Act. Based on this we conclude that training compute as (sole) base for the definition of GPAI is not only imperfect, but should not be used. It is hard to determine while being imprecise, already practically obsolete for what should be archieved and creates incentives to avoid regulation while the models are still capable of creating the risk the AI Act is supposed to protect from.

Question 2)

Question: Is 10ˆ22 FLOP a reasonable threshold for presuming that a model is a general-purpose AI model?

Answer: No.

We propose three alternative criteria or metrics for GPAI: From our point of view, an alignment with the possible application areas and the risk of uses that the AI Act is specifically intended to protect from would be preferable. Especially, since the risks associated do not only result from the capabilities of the model, but also due to context it is deployed in. This approach would be similar to the concept of identifiability in the sense of Art. 4(1) GDPR. In practice, this would mean a risk assessment by the providers as to whether their model is regulated or not, as in data protection law. This approach would also be similar to the regulation in product safety, where the effects and dangers

of a product matter, and not, for example, how long the plastic was kept warm before casting. Because that is not what matters if the resulting shape has a sharp edge or the raw material contains toxic impurities. A second approach would be to draw on industry benchmarks for various abilities of models, such as image generation or LLMs. These are also not perfect, as also mentioned in the proposal. However, providers and developers regularly highlight these benchmarks when they release a new model. So in practice these benchmarks seem to be good enough, at least in the view of the industry. This approach has at least three advantages: The metrics are easily available, the benchmarks at least approximately indicate the capabilities of the models and, moreover, developers hardly have an incentive to optimize this metric with regard to (non-)regulation. Because if they did, their model would appear less powerful compared to the rest of the market, making them less attractive.

Since models are sometimes also optimized for high values in corresponding benchmarks, new benchmarks and evaluation approaches are constantly appearing. This will make it necessary in the future to redefine the benchmarks to be drawn upon and the minimum score to be achieved. It is advisable to refer to a holistic view from several benchmarks, not least because these usually only measure certain abilities (e.g. MMLU Pro or COCOcap). Such benchmarks will also be available for other types of models or architectures, not least because providers want to advertise their products meaningfully on the one hand and, on the other hand, be able to empirically determine whether the effort for training the respective model was worth it. The benchmarks are usually set up for specific application areas, on which the selection and evaluation of such benchmarks for the determination if a model is GPAI could be based.

While the proposal addresses such benchmarks, it rejects them as being too immature. However, it argues against this view that they seem mature enough for the various providers to use them to advertise their models and further develop them on this basis. In comparison, the use of training compute appears to be much less reliable and, as with the latter approach, benchmarks and scores would also have to be reassessed and adjusted repeatedly. In result, this leads to no difference in practice. Moreover, the benchmarks used are characterized by being standardized, which means that the values of individual models are very easy to determine - and, as already mentioned, probably already exist due to the advertising effect. Usually, there also exists a standard framework to execute the tests connected to a benchmark easily. Compared with the effort required to create a model in the first place the effort needed to run the benchmarks seems to be neglible. There also exist benchmarks which specifically try to assess risks connected to LLMs (Guldimann, P. et al., COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act, https://doi.org/10.48550/arXiv.2410.07959). Those are relatively new, but could be either used instead of the industry benchmarks or be used alongside those. At least such benchmarks could be a base to iterate upon. A similiar approach could draw on the marketed use cases for models and whether they in sum are an indicator for generality of the model.

Thirdly: If the use of training compute as a criterion appears unavoidable, we would recom-

mend a graded approach. This would simultaneously result in sufficient legal certainty, but also flexibility. For example, it could be stipulated that models below a certain (very low) threshold would not fall under the term GPAI in any case and above a certain threshold in any case. Between these two thresholds, a balancing of interests would take place, which could be based on the criteria in Annex XIII of the AI Act (regarding taking in other factors besides training compute see Heim, L.; Koessler, L„ Training Compute Thresholds: Features and Functions in AI Regulation, https://doi.org/10.48550/arXiv.2405.10799). This would make the approach more transparent and align it with the goal of legal clarity stated in the proposal.

Question 3)

Question: With the proposed threshold of 10ˆ22 FLOP, or your alternative threshold suggested above, how many models and how many entities do you expect to be in scope of the AI Act, and why?

Answer: Given the simplicity of finetuning and smaller adjustments, as well as the open availability of very powerful models such as Deepseek R1, a reliable estimate of the number of regulated models and entities cannot be made. An illustrative example is the collection of text generators on the popular platform huggingface (https://huggingface.co/models?pipeline_tag=text-generation) where a single entry can represent multiple variants of a model that have been adapted or compressed in various ways. Huggingface is one of the most popular, but not necessarily the only platform for sharing models. In addition to text generators, image generators are also available, which, for example, may have a similar risk potential due to misuse with regard to disinformation.

Question 4)

Question: In addition to the examples presented in section 3.1.1 of the working document, are there other examples for which it would be important to clarify whether the presumption of being a general-purpose AI model based on the training compute threshold may be rebutted?

Answer: We would like to reiterate that we consider training compute to be a poor metric for determining performance, especially if it is the sole metric used. For a detailed explanation, please refer to our answer to question 1. In addition to the cases mentioned in the proposal, which are related to the highly limited use case, models could be excluded from regulation due to their limited user base, for example, because of a purely internal use of a model inside a company without any external impact. An example could be a model used for information retrieval. It should be noted that the use of the field of

application for an exemption from the applicability of the GPAI concept raises questions. This is because it presupposes that fields of application can be defined, which is also currently rejected in the proposal for the applicability of the definition. Thus, the proposal is contradictory at this point. In addition, the benchmarks already mentioned in our answer to Question 1 show that there are indeed case groups of application areas for which the definition of GPAI could be based upon. It should also be noted that when relying on a limited use case for an exemption, it must be considered how this restriction is archieved and how robust this restriction is. Robustness refers to the circumventability of protection mechanisms associated with the model. Thus, restrictions "built into" models can be partially circumvented or removed through additional training, for example, in the case of LLMs. In such cases, models are referred to as "uncensored" or "abliterated" on platforms such as Huggingface.

Question 5)

Question: Besides the criteria presented in Section 3.1.2 of the working document, are there other criteria that can be used to determine whether iterations, instances, or derivates of a model constitute distinct models for the purposes of the AI Act?

Answer: Model or knowledge distillation. That is the use of a large(r) model to specifically train a smaller model for higher capabilities than it would reach by "traditional" training. If techniques of model distillation are used we would suggest to view the distilled model as an derivate of the larger model, which does not constitute a distinct model. We would suggest to use the same approach for quantized models. For details refer to our answer to Question 1.

Question 6)

Question: In addition to the considerations presented in section 3.1.2 of the working document, are there other examples where it is unclear whether iterations, instances, or derivates of a model developed by the same entity constitute distinct models within the context of the AI Act?

Answer: See answer to Question 5.

Question 7)

Question: In addition to the considerations presented in section 3.2.1 of the paper, are there other examples for which it may not be clear which entity is the provider of a given general-purpose AI model?

Answer: None that we are aware of.

Question 8)

Question: Many downstream modifiers will have to assess whether they need to comply with the obligations for all providers of general-purpose AI models and the obligations for providers of general-purpose AI models with systemic risk. A pragmatic metric is thus highly desirable to limit the burden on downstream modifiers having to make this assessment, especially on smaller entities. Do you agree that training compute is currently the best metric for quantifying the amount of modification, despite its various shortcomings?

Answer: No. For reasons see the answers to Question 1.

Question 9)

Question: Are there examples of modifications of general-purpose AI models that meet the proposed training compute threshold of 3x10^21 FLOP, yet which should not result in the downstream modifier being considered a provider?

Answer: None that we know of

Question 10)

Question: Are there examples of modifications of general-purpose AI models with systemic risk that do not meet the proposed training compute threshold of one third of 10^24 FLOP, yet which significantly change the systemic risk profile of the model in ways that could not have been reasonably foreseen by the upstream model provider?

Answer: Finetuning a model for specific tasks or scenarios like disinformation. Also see Answer to Question 1 regarding the use of training compute as a metric for capabilities of a model.

Question 11)

Question: In addition to the examples presented in section 3.3.1 of the working document, are there other examples of when a general-purpose AI model should be considered as being placed on the market?

Answer: We would add that a GPAI model should be considered as being placed on the market, when it is integrated into an AI system made available on the market or put into service. In that regard, a GPAI model should be considered as having been placed on the market, if it is integrated into its providers own (general-purpose) AI system and is not used for purely internal processes that do not affect natural persons (cf. Recital 97 AI Act). Furthermore, the examples mentioned in section 3.3.1 are missing some essential parts of the AI Act's notion of a "placing on the market"-activity. First, a placing on the market-activity presupposes that a GPAI model is for the first time made available on the Union market. By that notion, any subsequent supply of the model, e.g., by copying it "onto a customer's own infrastructure", is not considered as placing the GPAI model on the market. Secondly, a "placing on the market"-activity requires the supply of a GPAI model "in the course of a commercial activity" Art. 3(10) AI Act). The notion of "commercial activity" is rather critical, also with regard to the exemption for AI systems and GPAI models released under free and open-source licenses (Art. 2(12) AI Act). Noteworthy, under other EU legal acts that are interrelated with the AI Act, the legislator specified that the supply of a product that is not monetised in any way should in principle not constitute a commercial activity (cf. Recital 18 of Regulation (EU) 2024/2847 [CRA] and Recital 14 of Directive (EU) 2024/2853 [PLD]). Taking account of a similar interpretation under the AI Act, we think that the supply of a GPAI model without any monetisation should in principle not constitute a commercial activity. The result of this interpretation is that the burden of regulatory compliance lies with providers of AI systems and GPAI models that are economic operators as these operators can include the compliance costs in the marketing of their AI systems and models.

Question 12)

Question: What are examples of ways in which open-source general-purpose AI models can be monetised?

Answer: There are various ways in which an open-source GPAI model can be monetised, such as:

- Providing access to the GPAI model via a platform that is monetised, e.g., through a subscription fee, the provision of data by the user or platform advertisement.
- Using the GPAI model via Cloud services that are monetised, e.g., with regard to the necessary computing resources (AIaaS).
- Distributing the GPAI model together with additional services that are monetised, such as support services, consultation on customization/specific use cases, etc.
- Distributing the GPAI model together with add-ons that are monetised, e.g. tools for simplified integration and adaptation for products, data sets for fine tuning, etc.
- Integrating the GPAI model into AI systems, products or services that are made available on the market against remuneration.

- Based on open-source licenses, a direct fee can also be charged.

Question 13)

Question: What are examples of 'information on usage' as stated in Articles 53(2) and 54(6) AI Act for open-source models?

Answer: We consider that the information on model usage should cover the information stipulated by Annex XI of the AI Act in so far as such information is necessary to have a good understanding of the GPAI model's capabilities and limitations in use. In that regard information should be provided on how the model can interact with software and hardware (Annex XII (1)(d)); on versions of relevant software for the use of the model (Annex XII(1)(e)).; the modality and format of inputs and outputs (Annex XII(1)(g)). We would further argue that information on the GPAI model's energy consumption should also be provided. Our reason for this is that the provision of information on energy consumption would allow downstream AI operators to assess the environmental impact of GPAI models that have been released and opt for more sustainable models. There are existing methods for the measurement of the energy consumption by LLM inference (see for example: M. F. Argerich and M. Patiño-Martínez, "Measuring and Improving the Energy Efficiency of Large Language Models Inference," in IEEE Access, vol. 12, pp. 80194-80207, 2024, doi: 10.1109/ACCESS.2024.3409745). For the case that training compute is used as a metric for the capabilities of a model (which we would not recommend, see our Answer to Question 1), the FLOPs used for training of the open-source model should be included. This information is already in the hands of the developer of the model, so that this information would not place an additional burden on them. Also it allows downstream users or providers to evaluate whether a model is regulated.

Question 14)

Question: What are examples of free and open-source licenses in the sense of the AI Act that allow for the access, usage, modification, and distribution of general-purpose AI models, and also require the release of publicly available information on the model parameters, including the weights, the model architecture, and model usage?

Answer: According to the wording of Art. 53(2) and 54(6) AI Act it is not necessary that free and open-source licenses (FOSS licenses) for GPAI models also require the release of publicly available information on the model parameters, including the weights, the model architecture, and model usage. Instead the release of such information seems to be an additional requirement. This is indicated by the notion of FOSS licenses under the AI Act (Recital 102) and by the use of the word "and" by the pertaining provisions, indicating

an enumeration of requirements for the exception to apply. Furthermore, the legislator also includes FOSS licenses that provide the licensee with the relevant rights of use for the GPAI model under the condition "that the original provider of the model is credited, the identical or comparable terms of distribution are respected", thereby including FOSS licenses with so-called copyleft clauses (Recital 102 AI Act). It is then necessary to consider, whether and how an AI model that has been developed with machine learning techniques is protected under IP law (evaluated in detail by: Gozalez Otero, Begoña, Machine Learning Models Under the Copyright Microscope: Is EU Copyright Fit for Purpose? (December 14, 2020). Forthcoming in: Nordic Law Review (NLR), Max Planck Institute for Innovation & Competition Research Paper No. 21-02, Available at SSRN: https://ssrn.com/abstract=3749233). With a view to the aforementioned literature, we would consider that the structure of an AI model, consisting of different algorithms may be protected as computer programs under copyright law. We would also argue that the parameters of such an AI model are not copyright protected works as they lack the necessary originality coming from a human author. With a view to the components of an AI model that qualify as computer programs, we consider the following prominent FOSS licenses for the exemptions under the AI Act:

- EUPL v1.2, SPDX identifier: EUPL-1.2 (https://eupl.eu/1.2/en/);
- Apache License, Version 2.0, SPDX identifier: Apache-2.0 (https://opensource.org/license/apache-2-0);
- The 2-Clause BSD License, SPDX identifier: BSD-2-Clause (https://opensource.org/license/bsd-2-clause);
- The 3-Clause BSD License, SPDX identifier: BSD-3-Clause (https://opensource.org/license/bsd-3-clause);
- The MIT License, SPDX identifier: MIT (https://opensource.org/license/mit)
- Mozilla Public License 2.0, SPDX identifier: MPL-2.0 (https://opensource.org/license/mpl-2-0).

Question 16)

Question: Are there any cases where a potential provider or downstream modifier would be unable to estimate the relevant amount of compute using any of the formula provided in this section? If so, why?

Answer: This could be the case, when federated learning approaches towards model training are utilized to protect the privacy of data subjects. In this case, instances of a model would be trained on different devices and by different entities and the updated model parameters shared and combined into the GPAI model. The compute used for the training of countless model instances by different entities may not be available to the provider of the GPAI model. Also, Information about the relevant values to estimate training compute is not always available. For example, the number of parameters and

the amount of tokens used for training for the popular Models of OpenAI is not public, A possible aid could be to include the relevant values in the metadata about a model.

Question 17)

Question: In addition to the approaches presented in the respective section of the paper, are there other ways for providers to estimate the amount of computational resources used for training?

Answer: In the spirit of sustainable AI, a further metric could be based on energy usage. This can easily physically be measured or virtually be estimated with specific functions of f.e. the typically used Nvidia GPUs. If a whole rack of servers or data center is measured, all other components are included as well, but these are also necessary for the main compute to do their work. This could at the same time provide an incentive for providers to be more efficient, though it should also be updated to keep step with technological progress. In general though, it does work similarly to other metrics, in that bigger and more expensive models user more energy.

Question 19)

Question: Are there examples of activities and methods that are specifically aimed at making the model safer, but which do not at the same time change the model's capabilities, and what would represent a rigorous justification that this is the case?

Answer: One such example is a second model that filters or applies the output of the main model. This is for example used in coding agents, where a second model formats the changes generated by the main model in such a way that the can be applied to code, checking for mistakes.

Question 20)

Question: What other activities and methods used during training should not be counted as part of cumulative training compute, and why?

Answer: We would argue that the utilization of privacy preserving machine learning (PPML) techniques facilitate model safety and can contribute to defence in depth against (privacy)attacks as an additional layer. Additional resources necessitated by the application of such methods should therefore not count towards the cumulative training compute. A survey of various risks to privacy in relation to machine learning as well as privacy preserving techniques for machine learning is provided by Feretzakis, G., Papaspyridis,

K., Gkoulalas-Divanis, A., & Verykios, V. S. (2024). Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review. Information, 15(11), 697. https://doi.org/10.3390/info15110697. Mentioned techniques include inter alia:

- Differential privacy;
- Federated learning;
- Homomorphic encryption;
- Secure Multi-Party Computation;
- Selective Forgetting.

These techniques aim to reduce the sharing of personal data and prevent or mitigate risks arising from privacy attacks on models and unintended information disclosure by models. These techniques can change the models capabilities, for example with regard to accuracy. However, the inclusion of these techniques in model training in this regard fosters the protection of fundamental rights to privacy (Art. 7 CFR) and protection of personal data (Art. 8 CFR). Their application should be considered by data protection authorities when evaluating the residual likelihood of identification of natural persons via AI models (cf. EDPB, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, adopted on 17 December 2024, p. 17 available at: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf (last visited: 13.05.2025)). With a view to the AI Acts objective of ensuring a high level of protection for inter alia fundamental rights of persons, the interpretation of requirements and obligations pertaining to the mitigation of AI related risks should include the application of privacy preserving techniques.

Question 21)

Question: How may providers reasonably and in a practically feasible way estimate the amount of computational resources used for synthetic data generation when the generating model is not their own model (for example a closed-source model accessed via API) or when the synthetic data set has been obtained from a third party (taking into account the possibility that the data set may not represent the entirety of the synthetic data generated to produce the data set, for example if a selection process was conducted), and how accurate would these estimations be?

Answer: They can only rely on data provided by the model provider or by the dataset provider. To our knowledge, without more information, it is impossible to estimate.

Question 22)

Question: When might a provider reasonably be expected to know how much compute they will use in post-training?

Answer: Assuming they do it themselves and have reasonable software and hardware access, they should be able to track it. If they do not, the service provider should be able to provide this data.

Question 23)

Question: s further clarification required regarding any of the aspects discussed in section 3.4.2 of the working document?

Answer: Section 3.4.2 inter alia explains when the provider of a GPAI model is required to estimate the cumulative amount of training compute for the development of their model and notify the Commission, if the stipulated threshold under the AI Act for GPAI models with systemic risks is met or it becomes known that it will be met. The AI Office's approach is that the provider should estimate "the amount of pre-training compute that they will use ahead of commencing their large pre-training run" and "to notify the Commission "without delay and in any event within two weeks" if the estimated value meet the threshold specified in Article 51(2) AI Act, following Article 52(1) AI Act" (cf. p. 16 of the Commission's Guidelines). This interpretation of the providers obligations raises questions towards the applicability of the pertaining provisions under the AI Act to such activities. According to Art. 2(8) AI Act research, testing or development activities regarding AI systems or AI models prior to their being placed on the market or put into service are not regulated by the AI Act (also cf. recital 25 AI Act). Recital 97 AI Act explicitly confirms the legislators understanding that the definition of GPAI models should not cover AI models used for research, development and prototyping activities before they are being placed on the market . Noteworthy, the legal terms "development" and "training" of AI systems and models are sometimes used together under the AI Act (see for example Recital 105 AI Act). In other cases, it seems the legislators understanding of "development" includes the training of AI systems and models. For example, Annex XI, section 1 describes the information that must be included in the technical documentation for GPAI models. Annex XI, section 2 stipulates "elements of the model [. . .] and relevant information of the process for the development", which inter alia include the training process for the GPAI model (Annex XI, section 1, point 2, b AI Act). This suggests that the training of GPAI models is part of their development. Should this be the case, the AI Office's approach towards estimation and notification obligations of providers of GPAI models that apply during the development of the model would be out of the regulations scope with a view to Art. 2(8) AI Act. Therefore, further clarification on the notion of the legal term "development" under the AI Act and its relation to the "training" of AI systems and models is needed, especially considering the exemption form the regulation's applicable scope under Art. 2(8) AI Act.