HUGGING FACE

GPAI Guidelines Consultation Feedback Surveys

Context:
https://digital-strategy.ec.europa.eu/en/news/commission-seeks-input-clarify-rules-general-purpose-ai-models

Preliminary approach for the content of the guidelines

General-purpose AI model

The definition of "general-purpose AI model" is key to understanding whether an entity must comply with the AI Act's rules for general-purpose AI models. See section 3.1 of the working document.

Conditions for sufficient generality and capabilities

See section 3.1.1 of the working document.

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?

- ☐ Yes
- ☑ No

Please explain why and which alternatives may be preferable.

An AI mode becomes "general purpose" based on its training objective and methodology. While a certain amount of compute is necessary for these techniques, models start being "general-purpose" significantly under the proposed threshold. The limits of FLOP thresholds have been previously discussed, e.g., [1]. Especially in recent months, we have seen highly capable general-purpose models, such as the models by DeepSeek [2] and other various providers that share their models openly, with various sizes of models [3,4]. Small models such as Phi [5] show that it is possible to create small models that learn

very efficiently from training data created by larger models.

While we appreciate and support the stated goal of limiting the burden on smaller entities, this means that any approach based on compute thresholds that covers all intended GPAIs will be unlikely to be sufficiently discriminating to achieve that goal.

Instead, we recommend ensuring that the open source exemption and definition of what constitutes “putting on the market” are properly scoped to limit this burden; especially given that most smaller entities “release” models by making them broadly available for download, often on platforms like GitHub or Hugging Face.

[1] https://cohere.com/research/papers/The-Limits-of-Thresholds.pdf
[2] https://huggingface.co/deepseek-ai
[3] https://huggingface.co/allenai
[4] https://huggingface.co/PleIAs
[5] https://huggingface.co/microsoft/phi-2

Is 10^22 FLOP a reasonable threshold for presuming that a model is a general-purpose AI model?

- ☐ Yes
- ☑ No

With the proposed threshold of 10^22 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?

Currently, we see a large number of openly shared models trained at or near the threshold. Dozens of small and medium developers from the EU and the rest of the world, including start-ups, non-profits, and universities, have independently trained models of billions of parameters on trillions of tokens. SmolLM v1 [1] and v2 [2], OLMo [3] and derivatives of the model, Pleias [4], LLM360 [5], Marin-8b [6] are examples of such models available on Hugging Face.

Training models that can be described as GPAI is increasingly accessible, and we argue that an even greater variety of models that are designed with different values, constraints, and high-level goals is desirable to maintain a thriving ecosystem that meets the requirements of all its stakeholders. Proportionate requirements on smaller developers, through the implementation of the proposed open source exemption and proper scoping of what constitutes putting a model on the EU market, will make this outcome much more likely.

[1] https://huggingface.co/HuggingFaceTB/SmolLM-1.7B

[2] https://huggingface.co/HuggingFaceTB/SmolLM2-1.7B
[3] https://huggingface.co/allenai/OLMo-7B
[4] https://huggingface.co/PleIAs
[5] https://huggingface.co/LLM360
[6] https://huggingface.co/marin-community/marin-8b-base

Differentiation between new models and model versions

See section 3.1.2 of the working document.

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?

The amount and type of additional data used to create a derivative of a model would be a much more insightful metric and better to make qualified judgement on whether the model should be further documented. As discussed above, a threshold in terms of FLOPS should be secondary to an evaluation of the impact of the model in terms of access to the model and its social impact [1,2]. This extends to derivatives of a model. A smaller FLOPS threshold as proposed here can be an additional indicator.

[1] https://arxiv.org/pdf/2502.16701
[2] https://ui.adsabs.harvard.edu/abs/2023arXiv230605949S/abstract

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?

The current draft only speaks about the scenario of the same entity creating and modifying a model. In the interest of science and fostering a healthy open source AI development community, we would urge the AI Office to apply the same rules for documentation of fine-tuned models to both models that have been developed by the same entity, compared to models fine-tuned by a different developer than the original model provider. That ensures a fair playing field for all model developers and avoids loopholes such as the acquisition of models or developing companies to avoid documentation of fine-tuned models, as they can now be claimed to be developed by the same entity. At the current draft, it is not clear how the application of the rules would differ between models developed by the same and different entities.

Downstream modifiers as providers of general-purpose AI models

See section 3.2.2 of the working document.

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?

- ☐ Yes
- ☑ No

Training compute is an incomplete metric to assess the impact of an AI model [1], which should be the goal of setting the threshold. Therefore, we would argue the changes to the model in terms of which training data was used is a more reliable metrics. These are still reasonably accessible to smaller providers and can serve as a better base for the AI Office to estimate which impact the model will have.

[1] https://cohere.com/research/papers/The-Limits-of-Thresholds.pdf

Exemptions from certain obligations for certain open-source releases

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

Free and open-source software (FOSS) licenses require that the model and its components be freely usable, modifiable, and redistributable, including for commercial purposes. As such, it is not compatible with FOSS licensing to restrict access to the model itself based on payment or to grant access exclusively through, e.g., monetised licensing arrangements.

That said, monetisation remains possible through downstream services that do not affect the model's open-source status. For instance, charging for hosted inference, enterprise support, or fine-tuning services built on top of the model is generally permissible, as these do not restrict access to the model weights or source code. Under the EU AI Act, such models would still be considered open-source general-purpose AI models, provided the model itself remains freely and openly available under a license that complies with

open-source principles.

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

Information on usage could be, e.g., download statistics, as they are provided for example on Hugging Face for the past month [1]. Mandating the collection of more information on downstream model use would in most cases conflict with open source licenses.

[1] https://huggingface.co/HuggingFaceTB/SmolLM-1.7B ("Downloads last month")

Estimating the cumulative amount of computational resources used for training

What are examples of activities and methods used during training or directly feeding into training that are intended to enhance the capabilities of the model prior to its deployment, beyond pre-training, fine-tuning and synthetic data generation?

Methods to train models are constantly changing, and different methods are applied for different modalities. While data curation and filtering is a method that is fundamental, the other methods might change and improve with time and research development. We list various examples from literature below that enhance model capabilities dependent on the task, modality, and devloper goals.

- Data curation and filtering: curating high-quality, diverse, and instruction-relevant datasets using heuristics, data filtering, or clustering techniques; preference-driven sampling, see https://en.wikipedia.org/wiki/Data_curation
- Curriculum Learning: structuring training data in increasing order of difficulty, allowing the model to learn basic patterns before more complex ones, e.g., https://arxiv.org/abs/2101.10382
- Contrastive learning or similarity alignment: used to better align representations or compress knowledge (e.g., SimCLR-style training, CLIP), e.g., https://arxiv.org/abs/2002.05709
- Knowledge distillation and teacher-student pretraining: distillation from larger models (or ensembles) into smaller models before fine-tuning, e.g., https://arxiv.org/pdf/2203.05180

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?

Most methods which do not change model capabilities but aim to make models safer are on the system level, such as system design [1] or output filtering and safety classifier [2].

[1] https://huggingface.co/blog/giadap/ai-personas
[2] https://ai.google.dev/responsible/docs/safeguards

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?

It is currently not feasible for providers to meaningfully estimate the compute used for synthetic data generation when the data originates from third-party or closed-source models accessed via API. Commercial providers typically withhold essential information about the model architecture, sampling processes, rejection rates, or the total number of generations involved. As a result, any estimate would be speculative at best and offer little legal or technical certainty. Without transparency obligations or standardised documentation practices, such estimations remain impractical and unreliable. Hence, the size, type of data, and purpose of the data would be more meaningful than calculating the compute the resources for generating the data.

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

A provider can reasonably be expected to know the compute they will use in post-training only when they also control the deployment of the model, as post-training compute depends not just on model size, but also on access patterns and the nature of user requests. While some post-training techniques, such as supervised fine-tuning, may have fixed compute budgets that can be planned in advance, others do not, making accurate calculations difficult unless the provider manages the full deployment context.

Is further clarification required regarding any of the aspects discussed in section 3.4.2 of the working document?

The current draft only considers compute as a threshold for systemic risk, however Article 51(1), point (a) details “high impact capabilities evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks”. The current exclusive focus on the compute threshold dismisses the crucial question of all other indicators and evaluations such as benchmarks.

We recommend devising thresholds based on reach (or accessibility) of the model, i.e., a prediction of how many people will interact with the model in some form such as a user interface [1]. It would be crucial to include past examples of global cybersecurity outages (CrowdStrike) or data breaches (Cambridge Analytica). We believe this would be the most efficient way to follow a precautionary principle; since one of the main characteristics of GPAIs, especially those classified as systemic risks, is their ability to drastically scale up behaviors with minimal human supervision, this targets both the most likely harm mechanisms and the ones best aligned with the general properties of these systems. Whatever the thresholds, handling risks that correspond to a high likelihood of reproducing or intensifying known and verified technological hazards should be given priority.

Further, the context of models and their deployment mechanisms are crucial to evaluate whether a model is harmful. Previously, knowledge about where the type of model has been deployed and the consequences of that deployment can inform whether the downstream consequences of different risk management approaches.

[1] https://arxiv.org/abs/2502.16701

Submitted by:
Lucie-Aimée Kaffee, EU Policy Lead & Applied Researcher, Hugging Face
Yacine Jernite, ML and Society Lead, Hugging Face