Key concerns with the General-Purpose AI Guidelines under the EU AI Act

Support compute as an imperfect quantitative proxy indicator for now (instead of data): Data is far less reliable as a proxy for capability since data quality, not volume, is what drives model risk or performance. Since model providers will already need to track training compute for other purposes under the AI Act (e.g., deriving energy consumption), compute is for now the more practical and readily trackable metric.

Introduce a capabilities-based alternative to compute thresholds:

- Given the lack of legal certainty today in relying fully on benchmarks (due to high saturation and benchmark turnover, and lack of standardized approaches), we believe a capabilities-based approach should be included as an **optional alternative to compute thresholds**.
- This also aligns with the notion of "safely derived models" under the GPAI Code, which exempts providers of GPAI models with systemic risk (GPAISR) from safety and security requirements for models they derive from GPAISRs that adhere to the Code, if certain conditions are met. One of those conditions is benchmark parity: latest draft of the Code would require a modified model to score no higher than the base GPAISR on all state-of-the-art (SOTA) general capability benchmarks. However, it should be sufficient for the modified model to perform no higher than the base GPAISR, across **a majority of relevant categories of SOTA general capability benchmarks**, as opposed to all benchmarks.

Remove the requirement to update the transparency template for new model versions: Providers should be required to update the transparency template **only for "distinct models"** (i.e., any modification of first-party models above the thresholds), and not for every new "model version" (i.e., *any* modification of first-party models below the thresholds). Mandating updates for every minor fine-tune could discourage safety improvements and risk disclosing highly sensitive data. Moreover, third-party modifications below the thresholds do not face the same requirement, implying an uneven application to first-party modifications.

Scope the application of model requirements appropriately for proprietary platform services: The understanding of "online repositories" should include proprietary platform services. Platform and model providers should also be able to contractually agree on which entity is the model provider. In the latter case, we seek clarity that compliance with model requirements should be the responsibility of the original model provider, and the **platform service provider should be treated as a provider of "AI tools, processes and/or services"**. The responsibilities of downstream entities modifying models would depend on whether they surpass the compute thresholds.

Remove the requirement to track compute used for generating synthetic data: There is no direct connection between compute used to generate synthetic data and data that ends up in the final data mixture used in a large pre-training run. It is unlikely that an entire synthetically generated dataset ends up in a final training corpus, and it is technically infeasible to isolate compute used for portions of a dataset that may have been synthetically generated.

Provide a grace period for implementing the Code: The finalization of the GPAI Code is delayed despite the AI Act's intention to have a three-month gap between the Code's adoption (intended to be 2 May 2025) and the start of model requirements (2 August 2025). We recommend providing model providers that sign the Code with a sufficient amount of time beyond 2 August 2025, to implement the Code's provisions.

Increase the compute threshold for GPAISR designation to align with current technological practices: The AI Act's current compute threshold (10^25 FLOP) for designation of GPAI models with systemic risk (GPAISR) risks being both too blunt, bringing in a broader range of less capable models than the regulation originally intended. We therefore recommend accelerating the issuance of the delegated act to **raise the current compute threshold to 10^26 FLOP** in the near term, ideally before model requirements start to apply on 2 August 2025.

Clarify scope of what constitutes monetization: "Monetization", which is not allowed under the AI Act's free and open-source exception, should refer solely to revenue-generating uses by the original provider that fall outside the open-source licence, and not to any licence-permitted activities (e.g., paid hosting, fine-tuning, or adaptation).

Clarify which open-source licenses meet the AI Act's open-source definition: OSI-approved licences (e.g., MIT, Apache) meet the Act's open-source definition, whereas restrictive licenses (e.g., Llama 4, RAIL) do not.

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?

We consider capability evaluations, which directly assess a model's generality and capabilities, to be the best metric. In the near term, we also recognize that leveraging training compute to indirectly assess generality and capabilities could help establish important legal certainty.

We welcome the AI Office's acknowledgement that training compute is an "imperfect proxy" for assessing a model's generality and capabilities. Relying on a compute threshold runs the risk of both scoping in an overly broad set of models and overlooking substantial performance gains that can be achieved with relatively less compute.

However, amongst various options being considered for quantitative thresholds (e.g., amount of data used for or amount of investment in model training), compute is the most appropriate quantitative proxy indicator. Currently, there are more significant gaps across methods for measuring the amount of data or for indicating its quality, and amount of investment is even further removed than training compute or data as a proxy for generality or capabilities.

Gaps also remain in methods for measuring the amount of training compute, though, and the AI Office could further improve legal certainty by recognizing the need for flexibility as standards continue to develop. There is currently no standardized methodology for measuring the amount of compute used to train a model, potentially impacting comparability among measurements. To avoid divergences with emerging global industry best practices, the AI Office should support and leverage international efforts to standardize methodologies to calculate training compute. In the immediate term, the GPAI Guidelines could also clarify in section 3.1 that companies may rely on robust estimation methods even as they should be prepared to provide context regarding how they have calculated compute.
As methods, tools, and standards for more directly assessing model capabilities (i.e. via capability-based evaluations) continue to advance, we recommend relying on them as the preferred approach to assessing both generality and capabilities. While evaluation instruments like benchmarks may need to be regularly updated, growing uncertainty regarding the reliability of indirect proxies for assessing advanced capabilities or downstream risks, along with increasing investment in standardized evaluation tools, make capability-based evaluations more reliable and future-proof.

We recommend aligning the thresholds for first- and third-party modifications that lead to new GPAISR by removing the scenario whereby a modified model could be designated as GPAISR if the sum of the compute for the base model and the compute for modifications surpasses 10^25 FLOP. This approach would also necessitate a detailed estimate of compute used for modifications, which currently risks inconsistent impacts to downstream providers given the lack of standardized approaches to measuring the amount of training compute. We also note downstream modifiers may not have the basis to know the amount of compute used for training the base GPAI model; under the GPAI Code v3, GPAI model providers will only be expected to disclose the amount of compute they used for training, which could be highly sensitive information, towards the AI Office or national competent authorities, and not towards downstream providers.

The cumulative scenario could also create perverse incentives, such as avoiding minor safety modifications to stay below the GPAISR threshold. It could capture a GPAI model just below the 10^25 FLOP threshold that is modified using a very small amount of compute, even making a smaller downstream entity a GPAISR provider even though the minor modifications it made are unlikely to materially increase risk. This scenario also emphasizes the imperfect nature of compute as an indirect proxy for capability or risk increase and the opportunity for a capabilities-based approach.

Allowing for a **capabilities-based alternative to the compute threshold** would also bring the GPAI Guidelines into alignment with the approach to GPAISR modification included in the third draft of the GPAI Code of Practice, which introduced the concept of "safely derived models." The Code's Measure II.4.2 suggests a signatory to the Code would not need to repeat measures under the Safety & Security section of the Code for any modified models derived from “safe originator models,” which refer to designated GPAISRs provided the following conditions are met:

1. **The base GPAISR must already be safe.** The provider of the original GPAISR must have completed the Code’s systemic-risk process. The GPAISR must have already been placed on the market, have no unresolved safety or security flaws, and be transparent enough for the signatory to understand its architecture, mitigations, capabilities, and risk profile. The condition on sufficient transparency towards the modifying entity is assumed to be true automatically for open-source or first-party models.
2. **The modified model must be *derived* from a GPAISR.** It is produced directly from the originator by distillation, quantisation, fine-tuning, post-training, or by adding or improving safety and security measures.
3. **Equal-or-lower risk must be a *reasonable* assumption.** That assumption holds when:
   - **Benchmark parity:** The safely derived model’s scores on state-of-the-art benchmarks that measure general capabilities are all lower than or equal (within a negligible margin of error) to the scores of the safe originator model, provided that the evaluations are carried out by qualified testers.
   - **No capability boost:** The modification process does not aim to or could not reasonably be expected to increase general capabilities, introduce new features relevant for systemic risks, alter risk-relevant propensities, or weaken existing safeguards.
   - **No new systemic risk scenarios:** After performing the systemic risk identification under the GPAI Code for the safely derived model, the modifying entity does not reasonably foresee any new systemic risk scenarios for the safely derived model compared to the safe originator model.

Regarding the benchmark parity requirement for a modified GPAISR to qualify as a safely derived model under the GPAI Code, or to insignificantly change a model’s systemic risk profile under the GPAI Guidelines, it should be sufficient for the modified model to perform **at or below the original model across a majority of relevant categories of advanced general-purpose capability benchmarks**. Focusing on relevant categories of state-of-the-art (SOTA) benchmarks, which could be defined as meeting qualifications such as low saturation levels, helps avoid overreliance on any individual benchmark while ensuring meaningful evaluation of uplift. Relevant categories of advanced general-purpose capabilities for SOTA benchmarks could include the following:

- General reasoning capability: ability to understand language and demonstrate knowledge across a number of different domains and tasks;
- Scientific and mathematical reasoning: ability to reason and solve challenging, expert-level questions;
- Spatial understanding and awareness: ability to understand video and imagery and spatial relationship between different objects;
- Long context reasoning: ability to reason over long context inputs;
- Autonomy, planning and tool use: ability to complete actions autonomously, plan and use tools;
- Advanced software engineering: Ability to perform software engineering tasks over whole code repositories.