🇪🇺

The EU AI Act - Key takeaways for LLM builders

The law differentiates **several categories of AI systems.** It primarily targets providers of AI models (rather than deployers or users), so unless mentioned otherwise, all below points apply to them.

⛔ Forbidden usages

Social scoring, manipulative AI systems, targeted surveillance by authorities (except for offenses passible of 4+ years in prison).

⚠ High risk systems

This is critical infrastructure (like roads or water supply), law enforcement, education and professional training, administration, access to essential public and private (for instance access to credit) services.

- They will have to conform specific compliance measures and transparency obligations
- Transparency disclosure involve publishing a list of points on a public database, including a "Fundamental rights assessment" that should certify that the model does not have discriminative biases against certain groups
- The individuals significantly affected by the decision of an AI system have a right to request from the deployer clear

explanations on the decision.

🤖 General Purpose AI (GPAI)

The law will have quite a strong impact on GPAI model providers ⇒ e.g. LLM providers. They have to conform to the following obligations:

- **For any content that could create a risk of impersonation or deception: the content should bear a watermark ⇒** This seems good, although technically hard to do since existing watermarking techniques do not stay unbroken for long or degrade the output a lot.
- **Set up a policy to respect copyright law**
- Build detailed technical documentation for the EU AI Office (template not defined yet) ⇒ this part of the disclosure will stay confidential
- Build less detailed documentation, to be made available to downstream users. **Some of the required elements are really sensible:**
    - Architecture and number of parameters
    - Data used for training, testing and validation, including type and provenance of data + curation methodologies
    - Regarding these two points, we still have not solid idea what went into GPT4, more than 1 year after its release! **So disclosing these publicly is probably a difficult thing to ask AI companies.**

☢️ GPAI models (=LLMs) with “systemic risk”

- 😶‍🌫️ **The concept of “systemic risk” is not well-defined.** Some criteria are mentioned, including a threshold of 10e25 FLOPs, number of parameters, or performance on certain benchmarks (cf Annex IXa).

  - The final definition will be decided by the AI Office.
  - For me anything that relies solely on FLOPs or number of parameters is meaningless, as for instance Yi-34B is more powerful by many benchmarks than the 10 times bigger Groq-1.

- "Since systemic risks result from particularly high capabilities, a general-purpose AI models **should be considered to present systemic risks if it has high-impact capabilities**, evaluated on the basis of appropriate technical tools and methodologies, or significant impact on the internal market due to its reach".
  - Judging by market impact, **GPT4 already bears systemic risk, and Claude-3 or Mistral-Large may soon**.
- Additional obligations for providers of LLMs with systemic risk:
  - Assess and mitigate possible systemic risks: **Build evaluations methodologies and metrics, report results** for each identified risk
  - **Red teaming:** set up adversarial testing.
  - Track, document and report serious incidents

👮 **Possible fines** for GPAI providers: 3% of their global turnover or up to 15M EUR, whichever is higher.

🤗 **What about Open source models?**

- To encourage open source model builders, **they will not be subjected to all disclosure requirements**. **Still they should produce a summary** about the training content and respecting copyright law.
  - Notwithstanding this simplification, the effort to publish a model you trained or finetuned is still going from close to 0

now to “some paperwork to do” in the future. **This will probably have a strong deterrent effect on small organisations or individuals** - who for now are really important in OS models.

My thoughts 👇

About the impact of this law on European startups that would want to build LLMs.

✅ To me, **these requirements are mostly good**: for instance, respecting copyright laws, and setting up watermarking when possible will protect content creators from seeing their style or content copied by inappropriately trained LLMs.

🤔 What I’m more concerned about are the public reporting obligations, which may **have a risk to advantage big players, thus creating additional consolidation on the market**.

Let me explain:

A lot of the advantage already resides in computing power. For instance, Meta has 600K H100 GPUs, while one costs approximately 30k$. You need thousands for pre-training large models, for instance our StarCoder2-7B used 145k hours of H100 (cf the paper). Even if you use rental as a service like our https://huggingface.co/training-cluster, prices are still really high: 7M$ for a 70B model.

This compute requirement makes it harder for startups to develop incremental innovations fast enough. Thus, **radical changes in architecture or data are key for startups to compete** with big names. The problem is that **if these changes must be publicly disclosed, the largest companies can leverage their immense GPU wealth to copy your changes and improve on them in a few weeks**. Thus reinforcing the consolidation effect.

What the European champions say 👇

Existing European AI Champions Mistral and AlephAlpha had been vocal opponents of the first drafts. But after some of their concerns have been addressed (with some examples of how lobbying can be virtuous), they now agree with the final version. For instance, in an interview to Le Monde, Arthur Mensch, CEO of Mistral, mentions that we should rather regulate the usage of models rather than their production (like we do for computer languages for instance), “**However, in its final form, the AI Act is quite manageable for us**.”

Aymeric Roucher - March 2024