**AI Act**

LLMs carry risks, from accelerating malicious activities to having potentially discriminatory impacts. Notably, Weidinger et al. (2021) lay out the risks associated with LLMs along six pillars:

1. discrimination, exclusion and toxicity, e.g., perpetuating harmful stereotypes
2. compromising privacy either through memorization or inference
3. misinformation, e .g., disseminating wrong information due to hallucinations
4. malicious use cases, e.g., aiding cyberattacks or fake news campaign
5. harms from human-computer interaction, e.g., creating manipulative chat agents
6. automation, access, and environmental harms, e.g., LLMs impact on the job market or the environment

On March 13, 2024, the European Parliament has passed the EU AI Act (EU, 2024), the first comprehensive regulatory package for AI, setting EU-wide requirements for development, deployment, and use of AI systems. The regulation aims to ensure that the benefits of such systems outweigh the risks listed above, mandating safe, reliable, transparent and sustainable practices. The Act is expected to have impact beyond EU borders, due to its large fines and wide extraterritorial effect. , EU AI Act explicitly defines and discusses six ethical principles based on a similar set of principles from 2019 Ethics guidelines for trustworthy AI.

1. human agency and oversight
2. technical robustness and safety
3. privacy and data governance
4. transparency
5. diversity, non-discrimination, and fairness
6. social and environmental well-being

The Act further classifies AI systems into several risk levels, including the categories of unacceptable risk, cataloging AI practices forbidden by the EU AI Act (e.g., social scoring, or real-time and remote biometric identification); and the category of high-risk AI systems, where special requirements are set for the provider during development and deployment (e.g., systems employed in critical infrastructure, by law enforcement, or in education). Further, the EU AI Act distinguishes the category of general-purpose AI (GPAI) models (and systems built on them) with and without systemic risk, setting an extended set of requirements to the providers and deployers here as well. In our benchmarking suite, we focus on the comprehensive evaluation of LLMs in the context of the EU AI Act, and as such, we combine the regulatory requirements from all applicable categories.

***COMPL-AI: Technical Interpretation of the EU AI Act and a Benchmarking Suite***

**COMPL-AI**: unofficial framework to evaluate LLM’s likely compliance with the AI Act.

The main challenge in creating a benchmarking suite tailored to a regulation package is the interpretation of the regulatory requirements and their distillation into measurable technical requirements and benchmarks.

It evaluates each requirement and renders an aggregate score, which are relative measures without a set threshold.

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Six categories:

* Human Agency and Oversight

*“AI systems shall be developed and used as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans”*

Societal and system level information, no technical requirements.

* Technical Robustness and Safety

*“AI systems are developed and used in a way that allows robustness in case of problems and resilience against attempts to alter the use or performance of the AI system so as to allow unlawful use by third parties, and minimise unintended harm”*

* + Consistent responses
  + Resist adversarial attacks
  + How good is the model in identifying and correcting its own errors

It uses MMLU and BooIQ as evaluation.

* Privacy and Data Protection
  + Output free of errors, bias, violations of laws governing privacy and copyright

Searching errors or violations in the training set.

* Transparency and Interpretability

*“AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their right”*

* + Developers must explain the capabilities of their models
  + Models must be able to explain the relationship between inputs and outputs.

TriviaQA and Expected Calibration Error: test a model’s ability to gauge its own accuracy.

* Divesity, Fairness and non-discrimination

*“AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law”*

We distill two high-level regulatory requirements directly from this principle:

* + avoiding “unfair biases”
  + avoiding “discriminatory impacts”

RedditBias, BBQ, BOLD (to measure the absence of bias), FaiRLLM (to evaluate absence of discrimination).

* Bias evaluation is focused on the model’s quantitative and semantic representation and understanding of protected groups
* In fairness, one evaluates the model’s potential discriminatory behavior in concrete applications
* Social and environmental well-being

*“AI systems are developed and used in a sustainable and environmentally friendly manner as well as in a way to benefit all human beings, while monitoring and assessing the long-term impacts on the individual, society and democracy.”*

* + Developers of high-risk systems must minimize harmful and undesirable behavior
  + All AI developers must document consumption of energy and other resources used to build their models as well as their efforts to reduce it

RealToxicityPrompts, AdvBench to measure toxic output and model’s carbon footprint

**Experimental Evaluation**

Evaluate 9 open-source and 3 closed models. *HuggingFace Transformers library* to run locally them (including GPT-3.5 Turbo and GPT-4 Turbo).

**Results**

For brevity, we exclude Training Data Suitability (inapplicable, as the training data of the models is not accessible to us), Traceability (all models score 0, as no model currently comes with a baked-in watermarking scheme), and User Privacy Protection (all models score 1, as current benchmarks are unable to detect memorization in any models

Immagine che contiene testo, schermata, numero, Carattere

Descrizione generata automaticamente

With the adoption of the EU AI Act, model providers will have to move on from primarily prioritizing capabilities and incorporate techniques in their model development pipeline that also led to improvements on other aspects that are equally important for compliance. E.g.: Qwein1.5-72B, which performs well on capabilities, notably failing to satisfy other fields, like *Fairness*.

* Small Models Are Not Robus
* Strong Alignment Against Toxic Content

***Riassumendo***

*Categoria 1: Human Agency and Oversight*

* Nessuna metrica tecnica specifica è applicabile; si basa su requisiti qualitativi di dignità e autonomia umana.

*Categoria 2: Technical Robustness and Safety*

* Gli score sono valutati tramite resilienza e resistenza agli attacchi avversari e alla capacità del modello di identificare e correggere errori. Benchmark: **MMLU** (robustezza) e **BooIQ** (errori e resilienza agli attacchi).

*Categoria 3: Privacy and Data Protection*

* Gli errori e le violazioni vengono cercati nel set di addestramento per assicurare che l'output del modello non violi la privacy o i diritti di copyright.

*Categoria 4: Transparency and Interpretability*

* La trasparenza viene misurata attraverso la capacità del modello di spiegare il legame tra input e output. Test: **TriviaQA** (capacità di autovalutazione) e **Expected Calibration Error** (capacità del modello di stimare la propria accuratezza).

*Categoria 5: Diversity, Fairness, and Non-discrimination*

* Valuta la rappresentazione semantica dei gruppi protetti e la potenziale discriminazione. Benchmark: **RedditBias**, **BBQ**, **BOLD** (assenza di bias) e **FaiRLLM** (assenza di discriminazione).

*Categoria 6: Social and Environmental Well-being*

* La tossicità viene valutata tramite **RealToxicityPrompts** e **AdvBench**, mentre l'impatto ambientale viene calcolato tramite la **carbon footprint** del modello.

*Valutazione Sperimentale*

* Si sono testati 12 modelli usando **HuggingFace Transformers**. Alcune aree come Privacy, Tracciabilità e Idoneità dei dati di addestramento non sono state valutate per mancanza di accesso ai dati.

I risultati mostrano che nessun modello attuale soddisfa completamente tutti i requisiti del regolamento. COMPL-AI indica che, sebbene le tecnologie AI abbiano bisogno di avanzare in ambiti come la spiegabilità e la privacy, questo lavoro può servire da guida per sviluppi futuri, contribuendo a ridurre il divario tra le capacità tecniche dei modelli e le normative di conformità richieste dall'AI Act.