The Dilemma of Uncertainty Estimation and Systemic Risk in the EU AI Act

Anonymous Authors[1]

Abstract

The AI act is a new European Union-wide regulation of AI system. It includes specific provisions for general-purpose AI models which however need to be further interpreted in terms of technical standards and state-of-art studies to ensure practical compliance solutions. This paper examines the AI act requirements for providers and deployers of general-purpose AI and further focuses on uncertainty estimation as a suitable measure for legal compliance and quality assurance in training of such models. We argue that uncertainty estimation should be a required component for deploying models in the real world, and under the EU AI Act, it could fulfill some requirements for transparency and trustworthiness. However, generally using uncertainty estimation methods increases the amount of computation, producing a dilemma, as computation might go over the threshold (10^{25} FLOPS) to classify the model as systemic risk.

1. Introduction

In this paper, we argue that to enable GenAI model evaluation and human oversight, it would be useful for providers and deployers of AI systems to be informed on the model confidence in output. For a human it is natural that she can express full confidence, partial confidence, or reply with a simple ”I don't know” (Barrett et al., 2023). Similarly, it should be expected that AI models can perform the task of confidence estimation themselves as well, as this information is useful for developers and deployers to give a weight to AI model responses, and as a proxy measure about really trusting the prediction, same as with other human-generated opinions and documents.

As development of trustworthy AI models is a core principle in AI act, this paper proposes and studies the feasibility of estimation as a mandatory component of training and evaluation of AI models, as it is currently not widely adopted and AI model developers do not generally build models with these advanced capabilities.

However, GenAI providers might be reluctant to implement such measure as its use during training will increase the amount of computation for the model development which will require compliance with the more stringent regime of GenAI with systemic risk. Therefore, we further examine the benefits, computational costs, and limitations of the measure as a GenAI quality benchmark.

2. Uncertainty in Artificial Intelligence

The overall concept of estimating AI confidence is the field of uncertainty estimation in machine learning, and there are many techniques for this purpose, relying on different assumptions (Gawlikowski et al., 2023). The overall issue with this field is that estimating AI model confidence usually requires additional computational resources, and it needs to be explicitly considered during the training process.

Large AI models like Large Language Models and Vision-Language Models often do not have proper confidence estimation capabilities (Groot & Valdenegro-Toro, 2024), by outputting confidences that are not a reflection of true confidence, as correct and incorrect answers have similar high confidences, and this prevents discrimination of correct and incorrect predictions (Huang et al., 2023b) (Xiong et al., 2023). The overall concept of confidence estimation requires that incorrect predictions have lower confidence than correct predictions, ideally with incorrect predictions having 0% confidence, and correct predictions having 100% confidence.

This relates to GenAI and other kinds of models, in that they produce incorrect predictions, colloquially called ”Hallucinations”. The most common definition for hallucination is ”generated content that is nonsensical or unfaithful to the provided source content” (Ji et al., 2023) (Huang et al., 2023a), indicating that they are not useful predictions to a human given some context or prompt.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

3. The European Artificial Intelligence Act

The AI act (AIA) is the first comprehensive regulation of AI systems in the European Union. (Parliament, 2024) It is in its final stage of adoption and is expected to enter into force in 2025. In its final stage of negotiation the act was extended, with specific transparency and model evaluation obligations for providers and deployers of general-purpose AI models (GenAI). According to Art 3 (63), an AI model is general purpose if: *(1)* it is trained with a large amount of data *(2)* uses self-supervision at scale *(3)* displays significant generality and *(4)* is capable of competently performing a wide range of distinct tasks. This definition encompasses not only generative AI but also foundational AI models. Further the AI act classifies two groups based on compute threshold: GenAI and GenAI with systemic risks. The GenAI is considered with systemic risk or high impact capabilities if the cumulative amount of compute used for model‘s training is greater than 10^{25} floating point operations (FLOPs) (Art. 51 (2) AIA). In this paper we focus on interpreting the new regulatory requirements for GenAI and specifically on large language models like ChatGPT v.4 or multi-modal models (audio, video, text, etc) which fit the definition of GenAI with systemic risk.

The power of GenAI models is that they can make predictions of many kinds, but these predictions are only useful to humans if they are approximately correct (Campos & Laurent, 2023). Incorrect predictions, or popularly known as ”hallucinations” result from data or modelling problems are generally not useful to humans due to being simply wrong (Huang et al., 2023a). Determining truth in AI models is very difficult as these models are not trained to produce an objective ”truth”, but to reproduce tokens from the training set, which make more or less meaningful answers, but there are no guarantees for correctness.

The AI act specifies concrete transparency documentation and model evaluation requirements for providers and deployers of general-purpose AI models (GenAI) in Art. 53 with specific focus on metrics to evaluate the GenAI model such as accuracy and performance evaluation metrics, quality of datasets assurance, and robustness against errors (explicitly listed in Annex XI AIA). Moreover, the law requires human oversight measures, which enable humans to interpret the AI system output and if needed to intervene in order to avoid negative consequences or risks, or stop the system if it does not perform as intended. This can significantly improve their integration in AI systems for specific tasks as it requires also close cooperation with downstream providers (those who implement the GenAI model in their own AI systems).

GenAI with systemic risk is a model that has high-impact capabilities in the sense that it is characterised with unpredictability, emerging capabilities, etc. Therefore, the legislator considers that GenAI with systemic risk can have actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale (Art. 3(65) AIA).

Providers of such models must follow a stringent regime with additional obligations to perform mandatory model evaluation and adversarial testing according to specific standards, which ensure assessment and mitigation of systemic risks. (Art. 55-56 AIA) The newly established AI Office is a body that will facilitate the development of standards and codes of practice. Therefore, research and discussion on what practical measures are necessary to assess the risks of GenAI is of crucial importance in this regulatory initiative. Despite the strict legal regime, existing guidelines on risk management in GenAI report issues with high uncertainty and lack of standards to mitigate it (Barrett et al., 2023), while an initial assessment by Stanford on the most common Gen AI models show that they suffer from data quality and governance issues, lack of transparency and low robustness. (Bommasani et al., 2023)

4. Discussion and Conclusions

The new AIA act is a brave fist step towards a comprehensive accountability regime for AI systems. However, the act is a framework that requires its interpretation with respect to each AI model or system on case-by-case bases. Some controversies emerge in the field of GenAI since integration of legal compliance measures like uncertainty estimation also increases the FLOPs for model training. Consequently, in legal terms the AI model becomes with higher risk, while in reality the implemented measures are aimed to reduce such risk. To decide if systems are risky based on amount of compute is a good starting point, but it is a simplistic view, as computation can be used for different purposes that might not imply emerging or unexpected properties of a model.

Users of AI systems, specially in GenAI, often want to know if the output they obtain from the system is correct or they should trust the prediction, but current GenAI systems do not give confidence estimates. This paper provided arguments that GenAI models should be trained with proper uncertainty estimation methods, and provide confidence estimates to the end user. This presents a dilemma, as such methods increase the computational requirements, and under the EU AI Act, might make the system being classified as systemic risk, which might discourage developers from implementing uncertainty estimation methods.

References

Barrett, A. M., Newman, J., Nonnecke, B., Hendrycks, D., Murphy, E. R., and Jackson, K. Ai risk-management standards profile for general-purpose ai systems (gpais) and foundation models. *Center for Long-Term Cybersecurity, UC Berkeley. https://perma. cc/8W6P-2UUK*, 2023.

Bommasani, R., Klyman, K., Zhang, D., and Liang, P. Do foundation model providers comply with the eu ai act?, 2023. URL https://crfm.stanford.edu/2023/06/15/eu-ai-act.html.

Campos, S. and Laurent, R. A definition of general-purpose ai systems: Mitigating risks from the most generally capable models. *Available at SSRN 4423706*, 2023.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Groot, T. and Valdenegro-Toro, M. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. *arXiv preprint arXiv:2405.02917*, 2024.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023a.

Huang, Y., Song, J., Wang, Z., Chen, H., and Ma, L. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*, 2023b.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Parliament, E. CORRIGENDUM to the position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU) 2024/ ...... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) P9$_TA(2024)0138(COM(2021)0206–C9-0146/2021–2021/0106(COD)), April2024.$

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.