Comment on "NTIA AI Open Model Weights RFC" [Docket No. NTIA–2023–0009]

Date:

March 25, 2024

By Electronic Delivery To:

Information Technology Laboratory
ATTN: AI E.O. RFI Comments,
National Institute of Standards and Technology
100 Bureau Drive, Mail Stop 8900
Gaithersburg, MD 20899-8900

303 2nd Street, Ste N460
San Francisco, CA 94107
info@unlearn.ai
415 455 3005

To Whom It May Concern:

Unlearn.AI, Inc. (Unlearn) is submitting these comments in response to the February 25, 2024 National Telecommunications and Information Administration (NTIA) AI Open Model Weights Request for Comment. Unlearn is innovating advanced machine learning (ML) methods to leverage generative artificial intelligence (AI) in forecasting patient health outcomes, starting with the domain of randomized clinical trials (RCTs). We produce a distribution of longitudinal forecasts for individual trial participants (e.g., their "digital twins"), enabling smaller and more efficient clinical trials to bring effective medicines to patients sooner. As part of our shared goal to improve patient health outcomes through more efficient clinical trials, Unlearn has developed an EMA-qualified method for prognostic covariate adjustment (PROCOVA) that leverages AI to enable increases in power in RCTs with continuous outcomes.

We applaud NTIA for taking action to understand the benefits and risks of open foundation models and agree that these models have the potential to transform research and advance scientific knowledge. While there are important considerations for mitigating the risks of rights- and safety-impacting AI systems, such as in healthcare, the evaluation of AI models should not be approached in a one-size-fits-all manner. Rather, AI technologies should be regulated according to risk level, which can only be determined by the specific context of use. As existing regulatory bodies would have the best understanding of associated risks in specific AI use cases, methods involving AI should be evaluated in a sector-dependent manner. For example, AI models used in drug development should be regulated by the U.S. Food and Drug Administration (FDA), where a risk-based framework should be developed to include requirements that are commensurate with the potential risk associated with the specific context of use.

The following addresses the questions put forth by the NTIA on dual use foundation AI models with widely available model weights:

1a. Is there evidence or historical examples suggesting that weights of models similar to currently-closed AI systems will, or will not, likely become widely available? If so, what are they?

While it is difficult to predict whether the weights of all currently-closed models will become widely available, there are some historical examples to consider. Several AI systems that were initially closed or proprietary have eventually been made available to the public, either in full or in part. These examples include Google's BERT (Bidirectional Encoder Representations from Transformers), LLaMA by Meta, as well as GPT-2 by OpenAI. The growing movement towards open science and open source has encouraged developers to enhance collaboration and shared knowledge, and advocate for the transparency and accessibility of scientific knowledge. However, many models have remained closed and proprietary, such as OpenAI's more advanced GPT models, Anthropic AI's Claude, and Google's Gemini.

There are various reasons that cause developers to keep their AI systems closed, such as commercial interests to maintain a competitive advantage, privacy and security, and to control the model's usage.

The history of AI model performance has driven the innovation of new AI models and the evolution of existing ones. In fact, comparisons between models of the same size category (see: LMSYS Chatbot Arena) have shown favorable ranking of open source models that are consistently improving over time. While the scale of capital has created a gap in performance between truly open-source and closed-source foundation models, improvements in algorithmic and computational efficiency, advances in hardware, and better quality of training data may make it possible for open-source models to reach the performance level of current closed-source models rapidly. Once open-source models are able to match the capabilities of closed models, the developers of closed models may be encouraged to open the weights of these foundational models.

1b. Is it possible to generally estimate the timeframe between the deployment of a closed model and the deployment of an open foundation model of similar performance on relevant tasks? How do you expect that timeframe to change? Based on what variables? How do you expect those variables to change in the coming months and years?

Estimating the timeframe between the deployment of a closed model and the deployment of an open foundation model of similar performance on relevant tasks is possible by looking at the gaps in human-evaluated performance between open foundation models and closed counterparts. While this is highly dependent on the specific AI model and its application domain, we can look towards a few examples. At the moment, it takes about 6 months to 1 year for similarly performing open models to be successfully deployed after the deployment of OpenAI's closed models. The time gap between proprietary image recognition models and high-quality open-source alternatives has narrowed relatively quickly due to robust community engagement and significant public interest. In contrast, more niche or complex applications, such as those requiring extensive domain-specific knowledge or data, might see longer timeframes before competitive open models emerge.

With continued research into making the training of foundation models more efficient, both in terms of training data requirements and overall compute cost and time, this timeframe will gradually decrease. Because of the significant utility of these models, and their great cost, there is huge market incentive to further optimize the efficiency of their production. The discovery of more data-efficient architectural approaches to building foundation models, the further development of task-specific hardware, advances in the field of resource efficient training (e.g. quantized or single bit model parameters), and focused improvements to firmware and software engineering for training libraries will contribute to making the production of foundation models with today's capabilities comparatively cheap over the next few years.

1c. Should "wide availability" of model weights be defined by level of distribution? If so, at what level of distribution (e.g., 10,000 entities; 1 million entities; open publication; etc.) should model weights be presumed to be "widely available"? If not, how should NTIA define "wide availability?"

There is no effective way to assess how "widely available" a given set of model weights are. The existence of a single public link to a set of model weights would be sufficient to ensure their continued distribution in public or private. In the case of Mistral AI, a public torrent link was sufficient to distribute their foundation models, even without their source code for inference. Open-source contributors were able to produce that code independently within hours.

Furthermore, once a foundation model is widely available, a vast number of progenitor models can be developed from them via fine tuning, instruction tuning, reinforcement learning from human feedback (RLHF), or direct preference optimization (DPO) to adapt the models and improve their performance on any number of specific use cases. Each of these subsequent models may have their own history of distribution and further refinement in the open-source community. These patterns of distribution have been observed for both Mistral AI and LLaMA model releases. The tools to modify and finetune open foundation models have become very refined over the past six months and will continue to have their effectiveness and accessibility improved.

1d. Do certain forms of access to an open foundation model (web applications, Application Programming Interfaces (API), local hosting, edge deployment) provide more or less benefit or more or less risk than others? Are these risks dependent on other details of the system or application enabling access?

Providing models through an API offers the possibility of rejecting certain prompts or user requests based on the risk tolerance of the hosting organization. However, it risks centralization and singular control of downstream applications. A recent example of this is the backlash that Google faced after its release of their Gemma and Gemini models, which were widely criticized for reinforcing the views of their designers rather than aiding their users. Furthermore, privacy risks associated with the content of user inputs are significant for centralized models. Users inputting personally identifiable information or classified materials into the model is a risk for system maintainers.

While the burden is on the user to make sure that they are using a system appropriately, the operator must comply with government regulations with respect to such information and any subsequent audits.

Local hosting, on the other hand, offers the maximum amount of freedom of use – even if a model has been fine-tuned to a designer's policy, holding the model locally allows users to alter the model to better suit their needs or use case. Furthermore, all inputs remain local, which obviates the privacy and security risks of information sharing with a second party. Of course, one may be concerned about how such models are being used, as there is no central authority on the nature of the local hosted models. However, the U.S. government should recognize that policing the ownership of local models is a losing battle. They already exist in the open and will continue to exist in the open, whether shared publicly or privately.

The focus should not be on policing or preventing the creation and use of foundation models, but rather on ensuring that we have the tools to react quickly to their misuse. This includes having criminal statutes up to date with the modern technological landscape, properly funding and equipping state and national cybercrime divisions, and prosecuting individuals who misuse technology to harm or defraud others.

1d.i. Are there promising prospective forms or modes of access that could strike a more favorable benefit-risk balance? If so, what are they?

Because capable models are already widely accessible and currently in use, the government should focus more on how they can apply foundation models to practical problems and understand how to detect and respond to their potential misuse.

2a. What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?

The most significant risk associated with widely available model weights is that one cannot police the use of the model. While this could potentially amplify the capabilities of bad actors to operate on a larger scale, the context of use remains critical to determining the risk level associated with a model.

A lack of source code for model training does not significantly slow down the use of a model, as the open-source ML community has shown that re-engineering equivalent model training from limited information is possible on the time-scale of weeks. If there are risks in the mere production of a certain foundation model, restricting access to the source code will not materially prevent those risks.

A lack of training data, however, does impede the progress of external development of model capabilities. If the intended, risk-bearing task has a limited or somehow unique dataset which is not widely available (e.g., genomic data, clinical records, medical imaging), then the reproduction of a dual-use foundation model may not be possible without large efforts being applied to acquire such datasets, whether in private or publicly. However, if one considers dual-use models that are able to develop capabilities based on widely available text, image, or video data from the internet, then access to such datasets is not a significant impediment to the reproduction of the dual-use foundation model, or the risks it represents.

2b. Could open foundation models reduce equity in rights and safety-impacting AI systems (e.g., healthcare, education, criminal justice, housing, online platforms, etc.)?

Open foundation models would only reduce equity in rights and safety-impacting AI systems if the models were misused, either intentionally or unintentionally, by their operators. In the example cases listed, the risks associated with AI or dual-use foundation models in particular are a consequence of a lack of the operators' understanding of how to apply such systems equitably. Or, in the case of an intentional misuse, the operator wittingly uses an AI model to "launder" responsibility from the operator of the model in order to justify a biased or improper decision.

These cases highlight why it is important for regulators to focus on the use of a model rather than the model in general. Ultimately, it is impossible to enumerate all the ways an operator may improperly use a technology, but the responsibility lies with the operator to take reasonable actions to ensure that the service or good they are providing does not cause public harm.

Whether the model is open or closed, the risks remain the same. A closed dual-use foundation model accessed via API is as much capable of misuse in specific situations as an open one which is operated by a third-party. A closed approach simply puts an unreasonable and unmaintainable burden on the central operator to foresee such cases and misplaces responsibility for improper use on the tool maker rather than on the operator who is misusing the tool.

2c. What, if any, risks related to privacy could result from the wide availability of model weights?

Making model weights widely available comes with the risk of leaked information from the dataset used to train the model. This is known as a membership inference attack, which is a common concern cited in the literature on differential privacy. However, in order for a member of the public to have privileged information leaked from a membership inference attack on the model weights, some third party would have already had access to their privileged information in order to put it into the training set. Since the individual's personal information would have already been accessible on the internet prior to the model's production, the responsibility for an information leak would be on other third parties rather than the foundation model developer.

These risks are mostly proprietary – the contents of training datasets are often considered intellectual property in the form of trade secrets. Leaking information about data mix is more harmful to the producer of the foundation model in a commercial sense than it is for the public or government at large. A possible scenario arises in which a model owner *and operator* continuously retrains their foundation model using user input they have collected as an operator. If this information was provided under an expectation of privacy from the user, and if the operator later releases a model trained using that information, then there could be a concern that the model leaks the information at a later point. Model operators must be forthright with their users about how their inputs may be used in the future for model training and should be informed if model weights are released to the public using their data.

7a. What security, legal, or other measures can reasonably be employed to reliably prevent wide availability of access to a foundation model's weights, or limit their end use?

There are no security, legal, or other measures that can reasonably be employed to reliably prevent wide availability of access to a foundation model's weights or limit their end use once they become publicly available. Please refer to the response in 1c for clarification.

7b. How might the wide availability of open foundation model weights facilitate, or else frustrate, government action in AI regulation?

The wide availability of open foundation model weights would frustrate government action in AI regulation for several reasons. Once foundation model weights are openly available, they can be downloaded and used by anyone with the necessary computational resources, regardless of their intent. This makes it difficult for governments to control who is using these models and for what purposes, potentially leading to misuse. Additionally, foundational models can be developed and hosted in one country but used globally, making it challenging for individual governments to enforce regulations since different countries have varying levels of regulatory frameworks and enforcement capabilities. This can lead to a patchwork of regulation that is hard to navigate and enforce effectively. Finally, the field of AI is evolving at an unprecedented rate, and government regulatory processes (which often involve extensive deliberations, public consultations, and legal frameworks) can struggle to keep pace with the speed of innovation, making it difficult to create timely and relevant regulations.

7d. What role, if any, should the U.S. government take in setting metrics for risk, creating standards for best practices, and/or supporting or restricting the availability of foundation model weights?

AI systems should be regulated in a sector-specific manner with a risk-based framework dependent on the model's context of use. Rather than the U.S. government implementing overarching standards for best practices of all AI systems, a sector-specific approach would address the unique challenges and risks each industry presents, ensuring that safety and innovation are balanced effectively.

8b. Noting that E.O. 14110 grants the Secretary of Commerce the capacity to adapt the threshold, is the amount of computational resources required to build a model, such as the cutoff of 1026 integer or floating-point operations used in the Executive order, a useful metric for thresholds to mitigate risk in the long-term, particularly for risks associated with wide availability of model weights?

Metrics related to raw computing power, like floating-point operations (FLOPs), present challenges for accurate and enforceable regulation. They aim to gauge risks associated with AI capabilities, with FLOPs serving as a measure of compute used in model development. However, the efficiency of compute use varies—inefficient use may not accurately reflect actual risk, while highly efficient methods can signify strong capabilities with less compute. Additionally, emerging hardware for model training might lack a direct equivalent to FLOPs for comparing training scales.

Auditing and validating FLOP reports pose significant challenges for regulators, given that estimation methods vary by organization. This variance could lead to underreporting or inaccurate claims about computational resources used, complicating regulatory efforts and potentially undermining the effectiveness of such regulations.

Moreover, imposing a computational threshold might inadvertently encourage the development of more efficient AI technologies to avoid regulatory scrutiny, potentially benefiting environmentally but also enabling less resourced entities to develop advanced AI capabilities. This could counteract government efforts to manage AI risks by diminishing any competitive advantage established through regulatory measures.

Thank you for taking the time to review our comment.

Best regards,

Jess Ross

Ross

Senior Government Affairs Lead, Unlearn