Association for Long Term Existence and Resilience (ALTER)
20/5 HeBesht, Rehovot, Israel, 7621032
Tel: +972-51-246-9363
https://alter.org.il

Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

Response to NTIA Request for Comments

Note: We have responded to a subset of questions, and have not addressed all sub-questions. Our non-response does not indicate that we think questions are not critical, but stem from our limited ability to make useful comments. We hope and expect that other commenters will provide useful ideas or suggestions in those areas.

Question 1 - How should NTIA define "open" or "widely available" when thinking about foundation models and model weights?

First, we congratulate NTIA for differentiating between "Widely Available Model Weights" and "open." we strongly agree that the word "open" should be reserved for something more specific than "public" or "widely available" model weights.

Historically, open source has referred to a collaborative method of software development and an ability of users to re-compile the software. Neither is implied by public model weights, which are being labeled "open." Further, the word open is intended to imply a level of transparency that many "open-source" models fundamentally lack, since they are potentially using non-public training data owned by the developer, or training data that is copyrighted by others and has dubious provenance.

To be called open, as a parallel, the training data must be public, and available free of restrictive licenses. In addition, details like the model hyperparameter search method, final hyperparameters, and training code for the base model, as well as all RLHF examples and the training regime used, as well as any changes introduced in other ways, should all be available. This allows another party to re-create the model. This would be necessary, for example, in order for someone to independently verify that the model they have does not have backdoors put in place by developers or other third parties[1]. Ideally, these models would then use open licenses which apply the requirements to any derivative models.

All of that said, widely available models, whether open source or simply those with public model weights or other forms of access, do have a variety of issues which the RF correctly notes should be addressed. The later questions properly asks whether the additional availability of training data and source code could increase risk of misuse, and other questions about access.

[1] Yang, Haomiao, et al. "A comprehensive overview of backdoor attacks in large language models within communication networks." *arXiv preprint arXiv:2308.14367* (2023). https://arxiv.org/abs/2308.14367

a. 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?

There is a wealth of information and potential sources for inferring this, which we hope others will provide. However, one often ignored aspect of the question is unintentional leakage or theft. Given the historical record of firms like Microsoft failing to secure their source code repositories, and the theft by 4chan of Meta's LLaMA model. Clearly, the cybersecurity of the labs is certainly not sufficient to ensure that sophisticated attackers could not steal the model weights, and in some cases, publicize them.

b. 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?

No response.

c. 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?”

No response.

d. 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?

Access to the model at scale allows model-stealing attacks, though it is unclear how much these preserve capabilities of the base model. API access which allows fine-tuning is a potential enabler of reversing post-training safety features. Local hosting and edge deployment, on the other hand, create additional risks of model theft or leakage.

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

No response.

Question 2 - How do the risks associated with making model weights widely available compare to the risks associated with non-public model weights?

a. 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?

Training data and source code seem unlikely to materially change the risk from open models, but it is plausible that fine-tuning and RLHF data could make reversing safety measures somewhat easier. It seems worthwhile to investigate whether safety measures can be reversed effectively without that data, however, as it seems very unclear that the data would make a material difference.
If model weights are available anyways, barriers to releasing such data seem likely to be negative, on net, due to the impacts on safety and other research.

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

It is possible that they reduce equity, but is seems far more likely that open models they increase equity. The places where use of models to reinforce inequity, especially where there is profit in (intentionally or accidentally) misusing the models, seem unlikely to be accelerated by more public or cheaper models, whereas efforts to ensure models are equitable, or efforts to evaluate whether a given model is inequitable, can be greatly assisted by having and being able to modify the model. Additionally, use of closed models which are biased can more easily be legally challenged, or regulated, when other more open and/or more fully audited models are available.

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

First, it seems overwhelmingly likely that models training on non-public data would be able to be used to access at least parts of that data. If a firm is willing to release model weights based on data they are not

willing to release, it is unclear what the longer term privacy of the training data is.. Therefore, widely available model weights based on using non-public data is rolling the dice on whether future work will allow others to extract that data. The presumption should be that such releases violate the privacy of any persons whose data was used.

Second, misuse of these systems will create a wide variety of issues, but wide availability seems likely to avoid risks similar to how search results are stored and reduce user privacy, by allowing many groups to run the models.

d. Are there novel ways that state or non-state actors could use widely available model weights to create or exacerbate security risks, including but not limited to threats to infrastructure, public health, human and civil rights, democracy, defense, and the economy?

i. How do these risks compare to those associated with closed models?

First, it seems very important to address a fundamental assumption that many unthinkingly include in their reasoning about open versus closed models, which is that AI firms will successfully keep model weights from state actors or other sophisticated adversaries. This seems incredibly unlikely. State actors sophisticated enough to run such models are likely to at least attempt to infiltrate not just premier AI labs organizations, but the external servers on which they train, with supply chain and social engineering attacks which will not necessarily even be visible to the organizations.

Second, given the computational costs of running SoTA models, it seems unlikely that the risk from technologically unsophisticated actors is the critical issue, and their inability to get the latest models seems like it doesn't materially change the risk. This is especially true because closed models which allow API access to LoRA and similar fine tuning seem likely to be approximately as vulnerable as open models to having safety features reversed, and unless AI system providers are implementing oversight and security that far exceeds the norm in the technology sector, and beyond that, is significantly more secure than the frequently-circumvented types of AML/KYC/Fraud prevention used in banking and similar, it seems very unlikely that non-state actors will not simply use shell companies, stolen credentials, or stolen credit cards to sign up for the access and use the systems.

ii. How do these risks compare to those associated with other types of software systems and information resources?

The risks are difficult to ascertain, especially for future models, given the explicit attempts to build human level intelligence and previous lack of success by firms in correctly predicting the capabilities of their own models in advance. For this reason, it seems many currently-comparable systems are a poor reference

class - AI is not simply software.

However, other high-security software systems and resources are tested in ways that are not currently used or feasible for LLMs, via extensive specifications prior to development, architecture maps, unit tests, and finally, verification and validation of the systems. Not only that, but at least in domains where security is taken seriously, even the final secured systems are fully air-gapped, or at least remain entirely inside of well secured networks. We are unaware of, for example, any military intelligence systems which are open to public API access, and if they exist, would question the wisdom and basic competence of those deploying the systems.

e. What, if any, risks could result from differences in access to widely available models across different jurisdictions?

The competitive impacts are plausibly significant, but given the current concentration of talent and resources, reasonable standards which are enforced in the US seem likely to be enforced across most model developers. This is because it seems to create a commercial disincentive from developing the most advanced models in ways that the US does not allow even elsewhere, because of the prominence of the US market. Even if not, the lack of availability within the US does not seem to change the risk from foreign use of the systems; security researchers will still be able to test and research unsafe models, even if they cannot be developed or used commercially or otherwise within the United States.

f. Which are the most severe, and which the most likely risks described in answering the questions above? How do these set of risks relate to each other, if at all?

No response.

Question 5 - What are the safety-related or broader technical issues involved in managing risks and amplifying benefits of dual-use foundation models with widely available model weights?

First, we agree with the implicit assertion that these are “dual-use” foundation models, by nature of being open. Narrow technologies can often be non-dual-use - though even in that case, they can be misused. General technologies can in some cases be restricted in ways that make their by-default dual-use nature into something less dangerous, if control and oversight is exercised
However, general technologies which can be adapted freely are inevitably able to be used in ways that are unconstrained. The very usefulness of language models is the way in which they can be used for any task.

AI is a tool which has many uses. There's no such thing as non-military, or non-medical, or non-disinformation, or non-job-displacing general AI. If capable enough, there’s no such thing as a safe general AI. There are only constraints on the general AI that (hopefully) prevent some uses. A capable model can be used to plan a party, or a kidnapping, and can orchestrate a research project, or a terrorist

attack. Safety measures can make these systems partly or largely incapable of contemplated misuse, but these safety measures can be reversed by those with access to model weights. Of course, similar risks may exist for those who have API access for fine tuning, or even those who spend time finding the latest jailbreaks, if there are not proactive measures to counter these uses.

1. What model evaluations, if any, can help determine the risks or benefits associated with making weights of a foundation model widely available?

For open models, a fundamental question for evaluations is the capability of the model, rather than the specific evaluation of whether a model can be misused in a given form. Because publicly available weights make safety features reversible, any evaluations that are intended to evaluate the risk should be based on a pre-RLHF model.

The benefits of open model weights, however, are significant, both in terms of allowing research to determine safety, and in terms of increasing our understanding of large models. That said, most of those benefits are accrued by allowing research groups access. The unfortunate result of any such policy is that access is exclusionary, and the administrative burdens of controlled access are significant.

Question 9 - What other issues, topics, or adjacent technological advancements should we consider when analyzing risks and benefits of dual-use foundation models with widely available model weights?

The use of foundation models in combination with other technologies is a critical enabler of misuse. One area which has been highlighted are biological design tools[2], which are far more concerning for misuse by sophisticated actors than language models alone - but language models can greatly lower the barrier for their misuse. These tools are often public but difficult to use; experts can presumably already design and build dangerous proteins, though these are far from actually weaponizable designs.

That said, there have been claims that bioterrorists would be able to commit attacks that they otherwise would not. The current class of models, however, are feasibly used only by experts, rather than unsophisticated users, and it seems that we are at least a generation away from the most concerning applications in any case. Given that, now is the time to build robust evaluations, structured access models,

[2] Sandbrink, Jonas B. "Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools." *arXiv preprint arXiv:2306.13952* (2023).https://arxiv.org/abs/2306.13952

and oversight capabilities, and investigate ways to enable future dangerous models, whether open or not, to not be run in environments without safeguards.