*QUESTION 1. 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?*

Yes, there is evidence suggesting that weights of models similar to currently-closed AI systems will become widely available. Most recently, this has happened in two cases:

1. LLaMA 1[[1]](#footnote-2). This open-weight model is superior to the unreleased 1.5B version of GPT-2[[2]](#footnote-3), which was the most powerful publicly known language model in the world before the development of GPT-3.

2. Mixtral of experts[[3]](#footnote-4). This model is about on par with GPT-3.5 Turbo, a superior version of GPT-3.5 released in March of 2023[[4]](#footnote-5).

Something that should be noted is that, so far, no open model is better than the best closed models. It remains to be seen if this pattern continues, however.

*QUESTION 1. 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?*

From all currently available evidence, it is impossible to predict the timeframe between the deployment of a closed model and the deployment of an open foundation model of similar performance on relevant tasks, to any reasonable degree of accuracy. However, preliminary evidence suggests that the timeframe is under 1 and a half years.

As mentioned in the response to Question 1. a., in the past, such an effort has taken over 4 years – the gap between GPT-2[[5]](#footnote-6) and the official release of LLaMA 1[[6]](#footnote-7) – but this has recently shrunken to ~9 months – the gap between GPT-3.5 Turbo[[7]](#footnote-8) and Mixtral[[8]](#footnote-9). If LLaMA 3 truly comes out in July 2024 as *The Information* reports[[9]](#footnote-10), and is truly competitive with GPT-4 as is rumored[[10]](#footnote-11), then the gap will grow to 1 year and 4 months. This severe variation suggests that the field is in a constant state of flux, and due to both that and a lack of data, predictions to even within a month are likely to be unreliable as of this time. However, as mentioned above, the timelines indicate that a closed model can generally expect to be followed by an open model of equivalent power within 1 and a half years.

One important thing to note is that this is reliant on the closed models weights not being leaked, like what happened with LLaMA 1[[11]](#footnote-12). In that case, timelines are effectively impossible to estimate.

*QUESTION 1. 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?’’*

Wide availability of model weights should be defined by level of distribution, but it should also be defined by where and how the weights are distributed.

For the first criteria, level of distribution, wide availability should be set to 1,000 entities. This is around the time when it becomes impossible to effectively monitor all usage of the weights, and thus when the risk of unauthorized duplication begins to rise.

Note that this prevents distribution of a model’s weights to the entire population of many AGI labs – for example, DeepMind alone has ~2000 employees[[12]](#footnote-13) - but in our opinion, this is a plus, not a negative. Not everyone in an AGI lab should have access to the weights of a model; instead, only those for which it is critical should have access to the weights (e.g. an engineer in charge of monitoring pretraining of the model.)

For the second criteria, weight distribution, wide availability is determined by how the weights are distributed, and what security measures are put in place to prevent unauthorized duplication. For example, a server accessible only from a company’s intranet, that requires a username, one time use password, and two factor authentication in order to access the weights is inherently much safer than a torrent link, which can be easily leaked and is impossible to cut off access to[[13]](#footnote-14).

*QUESTION 1. 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?*

Inherently, there are some forms of access to an open foundation model that provide both more benefit and more risk than others. These are, in order of least benefit and least risk to most benefit to most risk:

1. Web applications. This category includes tools like OpenAI’s ChatGPT or xAI’s Grok. It is only possible to interact with a web application in authorized ways, with authorized forms of input. This is both the safest and least beneficial form of access; while this greatly reduces the risk of harm, it is still possible for a determined attacker to perform hostile activities[[14]](#footnote-15).

2. APIs. This category is slightly less safe than web applications as it usually allows both custom system prompts and using old versions of a certain model – both of which could allow an attacker to circumvent normal safety measures. For example, exploiting the OpenAI API was how Carlini et al. were able to extract the embedding projection layer from two of OpenAI’s closed models[[15]](#footnote-16). Additionally, certain features such as finetuning are only available inside the API[[16]](#footnote-17), increasing both benefits and risk.

3. Edge deployment. Depending on how this is set up, this can work as either a web application or an API, so the same risks apply in their respective case. However, note there is a greater risk of the models being stolen ‘in-transit’ by a hostile actor who is in control of the edge computer.

4. Local hosting. This is when the weights of a model are given to a user, and it is by far both the most beneficial form of access, as well as the most risky. This is for a number of reasons: firstly, it is incredibly easy to get around any form of safety built into the model. In many cases injecting an “Ok,” response into the model of the output is enough to completely bypass the model’s constraints, and new techniques for bypassing safety are being discovered every day[[17]](#footnote-18) and will likely only get more powerful over time. On the flip side, finetuning allows you to modify a model to such a degree that it can even outperform much larger closed models at a certain task, which lets you save significantly on both cost and complexity[[18]](#footnote-19).

*QUESTION 2. 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?*

There are a multitude of risks associated with widely available model weights, and from the publication of their training data and/or source code.

Firstly, the primary risk from open-weight models is the possibility of fine tuning a model in order to generate biological weapons. The reason for biology being of more concern than e.g. cyberweapons is twofold: firstly, unlike software, you cannot “hot-patch” the human body in the same way you can a computer system. While a patch for a software exploit can be written and deployed to millions of computers within a day, making a human immune to a virus requires a long and expensive process of development before a vaccine can be used, including clinical trials[[19]](#footnote-20), which is simply not a viable option in a world where all you need to do to create a pandemic-capable biological agent is ask your open-weight model for the instructions.

Secondly, unlike most other sciences (such as physics, math, etc.), biology does not require large breakthroughs in order to create dangerous organisms. The way chemicals in the body interact with each other on a molecular level is a mostly solved problem; what is hard is the sheer scale of the task: the human body contains a countless number of things that interact with other things, and it’s effectively impossible for any human to keep track of all those things when designing a bioweapon. This is much less of a concern for a model that is trained on most of the internet, especially when paired with a verifier[[20]](#footnote-21), which would guide a model towards producing a bioweapon. Open-weight model based automated cyberweapons are also a concern, but are of secondary importance to bioweapons.

A distant tertiary risk of open-weight models is the possibility of using the model to generate CSAM/spearfishing emails/spam, etc. While these risks are absolutely worth consideration, especially as current models are already capable of doing these things, they are not as relevant or as urgent as biological and cyber risks.

Lastly is the risk associated with open sourcing a model’s code/training data. Unlike open-weight models, bioweapons are not a primary concern for one main reason: it is very, very expensive to take the code/training data and create an actual model from it. According to Meta, LLaMA 2’s training took 3.3M GPU hours on A100 GPUs, each of which consumed at least 350W[[21]](#footnote-22). *(Continued on next page.)*

Even assuming unreasonably low electricity prices (roughly 0.025 USD/kWh in Iran[[22]](#footnote-23)), the electricity costs alone would total around 3.8M USD, and this is not considering the cost of acquiring enough A100 GPUs to train the model, the datacenter required to house the A100s, the networking equipment required to link all the GPUs together, etc. Though cloud computing is a cheaper option, the total cost to go from training data and source code to model weights is still in the millions of USD.

This is well outside the reach of the kind of low resource terrorist who is the main producer of low-skill AI-generated bioweapons, especially considering that the future models capable of generating bioweapons are likely to require several times more compute than what was used to train LLaMA 2.

The main concern with releasing the training data/source code of the model is, perhaps expectedly, nation-state actors. They are the primary beneficiaries of this information, having both the compute necessary to train the model, and lacking the knowledge to develop an equivalent model on their own.

Additionally, while it is possible that a nation-state may attempt to use open models in order to create bioweapons (either via training the model on the open sourced code, or simply by using available weights) and deploy them against their adversaries, given evidence from the 2020 pandemic it has become abundantly clear that pandemic agents are not controllable, and are liable to infect both your citizens and your enemies’. The greatest risk from nation-state actors in the near term is thus AI cyberweapons, and in the mid to late term, AGI (though this relies on a number of assumptions that are impossible to estimate before the development of an AGI – will the AGI have a fast or a slow takeoff, how intelligent will an AGI be when run with a certain amount of hardware, etc.)

*QUESTION 2. 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?*

Our response to this is mostly the same as the response to Question 2. a.: the most obvious novel threats are AI created bioweapons, AI created and/or controlled cyberweapons, and CSAM/spearfishing emails/spam. Any one of these could have catastrophic effects on infrastructure, public health, human and civil rights, etc., and in the latter case, current models are already being used in this way[[23]](#footnote-24).

*QUESTION 8. b. 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 10^26 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?*

Using amount of computational resources as a measure of mitigating risks associated with the wide availability of model weights is not very useful. The amount of computational resources used to train a model tells you little about the size of the resulting model (which is the main bottleneck of running models on local consumer-grade systems), and tells you nothing about methods of distribution, the security of the organization who made the model, etc.

1. Meta Platforms, Inc., “Introducing LLaMA: A Foundational, 65-Billion-Parameter Language Model,” February 24, 2023, https://ai.meta.com/blog/large-language-model-llama-meta-ai/. [↑](#footnote-ref-2)
2. Alec Radford et al., “Better Language Models and Their Implications,” February 14, 2019, https://openai.com/research/better-language-models. [↑](#footnote-ref-3)
3. Mistral AI, “Mixtral of Experts,” December 11, 2023, https://mistral.ai/news/mixtral-of-experts/. [↑](#footnote-ref-4)
4. Greg Brockman et al., “Introducing ChatGPT and Whisper APIs,” March 1, 2023, https://openai.com/blog/introducing-chatgpt-and-whisper-apis. [↑](#footnote-ref-5)
5. Radford et al., “Better Language Models and Their Implications.” [↑](#footnote-ref-6)
6. Meta Platforms, Inc., “Introducing LLaMA.” [↑](#footnote-ref-7)
7. Brockman et al., “Introducing ChatGPT and Whisper APIs.” [↑](#footnote-ref-8)
8. Mistral AI, “Mixtral of Experts.” [↑](#footnote-ref-9)
9. Sylvia Varnham O’Regan and Kalley Huang, “Meta Wants Llama 3 to Handle Contentious Questions as Google Grapples With Gemini Backlash,” The Information, February 28, 2024, https://www.theinformation.com/articles/meta-wants-llama-3-to-handle-contentious-questions-as-google-grapples-with-gemini-backlash. [↑](#footnote-ref-10)
10. OpenAI, “GPT-4,” March 14, 2023, https://openai.com/research/gpt-4. [↑](#footnote-ref-11)
11. ChristopherKing42, “Save Bandwidth by Using a Torrent to Distribute More Efficiently · Pull Request #73 · Meta-Llama/Llama,” GitHub, March 2, 2023, https://github.com/meta-llama/llama/pull/73. [↑](#footnote-ref-12)
12. Amir Efrati, “Briefing: DeepMind Cut 20% of Its Expenses Before Merging with Google,” The Information, October 11, 2023, https://www.theinformation.com/briefings/deepmind-cut-20-of-its-expenses-before-merging-with-google. [↑](#footnote-ref-13)
13. ChristopherKing42, “Save Bandwidth by Using a Torrent to Distribute More Efficiently · Pull Request #73 · Meta-Llama/Llama.” [↑](#footnote-ref-14)
14. OpenAI, “Disrupting Malicious Uses of AI by State-Affiliated Threat Actors,” February 14, 2024, https://openai.com/blog/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors. [↑](#footnote-ref-15)
15. Nicholas Carlini et al., “Stealing Part of a Production Language Model” (arXiv, March 11, 2024), https://doi.org/10.48550/arXiv.2403.06634. [↑](#footnote-ref-16)
16. OpenAI, “OpenAI Platform,” accessed March 16, 2024, https://platform.openai.com. [↑](#footnote-ref-17)
17. Andy Zou et al., “Universal and Transferable Adversarial Attacks on Aligned Language Models,” n.d. [↑](#footnote-ref-18)
18. Boshi Wang et al., “LLMs in the Imaginarium: Tool Learning through Simulated Trial and Error” (arXiv, March 7, 2024), https://doi.org/10.48550/arXiv.2403.04746. [↑](#footnote-ref-19)
19. Public Health Agency of Canada, “COVID-19: How Vaccines Are Developed (Video),” education and awareness, December 14, 2020, https://www.canada.ca/en/public-health/services/video/covid-19-how-vaccines-developed.html. [↑](#footnote-ref-20)
20. Hunter Lightman et al., “Let’s Verify Step by Step” (arXiv, May 31, 2023), https://doi.org/10.48550/arXiv.2305.20050. [↑](#footnote-ref-21)
21. Meta Platforms, Inc., “Llama 2: Open Foundation and Fine-Tuned Chat Models,” Research - AI at Meta, July 18, 2023, https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/. [↑](#footnote-ref-22)
22. “Iran Electricity Market - Irema,” accessed March 10, 2024, https://www.irema.ir/. [↑](#footnote-ref-23)
23. Holly Ramer and Kevin McGill, “Magician Says Political Consultant Hired Him to Create AI Biden Robocall Ahead of New Hampshire Primary,” PBS NewsHour, February 23, 2024, https://www.pbs.org/newshour/politics/magician-says-political-consultant-hired-him-to-create-ai-biden-robocall-ahead-of-new-hampshire-primary. [↑](#footnote-ref-24)