Cohere Comments on the National Telecommunications & Information Administration's Request for Comments on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights

Docket Number NTIA 240216-0052
March 27, 2024

Cohere appreciates the opportunity to submit comments in response to National Telecommunications & Information Administration (NTIA)'s Request for Comments (RFC) on Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights (open foundation models).

Introduction

Cohere is one of the leading enterprise-focused foundation model developers worldwide. Cohere empowers every developer and enterprise to build amazing products and capture true business value with language AI.

Cohere is the only foundation model developer to have signed each of the White House's *Updated Voluntary Commitments*[1] and the *Canadian Federal Government's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems*[2] and to also endorse the G7's *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems*.

Cohere For AI (C4AI) is Cohere's non-profit research laboratory[3] and is committed to the development of safe and responsible AI, with a focus on cutting-edge research and approaches to safety, and creating more points of entry into machine learning research. To-date, C4AI has published 39 papers with 40+ cross-institutional collaborators, and ran open science projects and research communities involving 3000+ independent researchers across 119 countries. C4AI has also helped bridge access gaps through a scholars program to support researchers around the world[4], a compute grant program to bridge access to resources[5], and research artifact releases to share technical expertise with the wider machine learning ecosystem.[6]

Both Cohere and C4AI are members of the *US AI Safety Institute Consortium*.[7]

[1] Cohere Signs White House's Updated Voluntary AI Commitments: <www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/>.

[2] Cohere Signs Canada's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems: <ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems> .

[3] For more information see:<cohere.com/research/>.

[4] For information about C4AI's scholars programme, see: <txt.cohere.com/c4ai-scholars-program>

[5] For more information about C4AI's Research Grant Program, see: <txt.cohere.com/c4ai-research-grants/>

[6] An up-to-date list of C4AI's research publications is available at: <cohere.for.ai/#spotlight-papers>.

[7] The U.S. Artificial Intelligence Safety Institute Consortium (AISIC). <www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute>.

Summary

Our response to this RFC on open foundation models is informed by our experience in both building a for-profit enterprise centric AI business as well as investing in a world-recognized not-for-profit research lab dedicated to contributing to sharing leading edge innovation in the field. Having experience in building models that we release as both fully open weights and proprietary models via flexible deployment mechanisms (including via API), we fundamentally believe in the benefits of these different approaches. We strongly encourage the US government to explore a range of regulatory options to support this flexibility going forward.

In February, C4AI released Aya[8], a state-of-art multilingual model that doubles the number of languages served by an open source LLM while establishing a new precedent for technology that serves under resourced languages. Along with Aya, C4AI released the largest multilingual instruction fine tuning dataset to-date, covering 513 million instances covering 101 languages. Both the Aya model and dataset were released under a fully permissive Apache 2.0 license to further access to communities historically underserved by language coverage of commercially available LLMs.[9]

In March 2024, C4AI released the weights for C4AI Command-R[10] publicly so that it can be used for research purposes.[11] C4AI Command-R is a highly performant state-of-art, weights-only release licensed for non-commercial uses, which is intended to allow researchers access to a productionized model from Cohere for research and safety auditing. This is part of our commitment to explore ways to enable third party cross-institutional efforts around fundamental research in safety. This model weight release is part of our wider support across both Cohere and C4AI for the machine learning ecosystem alongside research compute grants[12], publishing research at top tier venues[13] and regularly releasing code and libraries under a variety of licenses[14].

[8] Aya is a global initiative involving over 3000 independent researchers across 119 countries who contributed to datasets, testing, refining and more. Information about Aya is available at: <cohere.com/research/aya>; papers relating to the Aya dataset and model are available at: <cohere.com/research/papers/aya-dataset-paper-2024-02-13> and <cohere.com/research/papers/aya-model-paper-2024-02-13>.

[9] The Aya model and dataset are available via their repositories on Hugging Face. Model weights are available at: <huggingface.co/CohereForAI/aya-101>; and datasets are available at: <huggingface.co/datasets/CohereForAI/aya_collection>.

[10] The C4AI Command-R model is: (a) a version of Cohere's 'scalable' category Command-R LLM; and (b) released under a *Creative Commons Attribution Non-Commercial Public License* with an addendum requiring users to adhere to C4AI's acceptable use policy.

[11] C4AI Command-R is available at: <huggingface.co/CohereForAI/c4ai-command-r-v01>.

[12] For more on C4AI's Research Grant Program, see: <https://txt.cohere.com/c4ai-research-grants/?_gl=1*1fprkmg*_ga*MTA2NjMwMjA0NC4xNzA5MjI3MjM1*_ga_CRGS116RZS*MTcxMDk4MjA0OC4yMS4xLjE3MTA5ODIwNTMuNTUuMC4w>.

[13] Cohere for AI research papers are available at: <https://cohere.com/research/papers>.

[14] For more information about Cohere and Cohere For AI's sandbox, see: <txt.cohere.com/introducing-sandbox-coheres-experimental-open-source-initiative/>.

As it relates to open foundation models, we believe that:

1. Concepts such as ‘openness’ and ‘widely available’ should be defined contextually on the basis that openness is not a binary concept but rather a spectrum.
2. There should not be a monolithic approach to openness, and the level of accessibility and release format (e.g., open weights, API access only, private deployment) should depend on a holistic case-by-case evaluation of the risks and benefits.
3. It is important that licenses and access restrictions reflect perceived risks.
4. When considering approaches to model ‘openness’, considerations must go beyond the degree of weight release to also consider questions such as access to compute and expertise.
5. The regulatory approach to open foundation models needs to be dynamic, proportionate and contextual, taking a risk-and-principled approach.

Our detailed submission below outlines the five issues highlighted above. As the capabilities of our generative models advance and as the policy and governance debate around open foundation models evolves, Cohere and C4AI are committed to iterating and developing our approach to openness, and we will continue to pioneer best practices and share our learnings.

Our Detailed Submissions

(i) Openness is not binary but rather a spectrum

There is currently no consensus on what is considered ‘open’ or ‘widely available’. Existing research describes different ‘gradients’ of openness that can apply to foundation models[15] with different considerations for risks associated at each gradient from fully closed, to fully open.[16] Cohere and C4AI have experimented with different levels of access, from ‘closed’ API access to proprietary models (Cohere Command, Rerank and Embed), research releases of model weights (C4AI Command-R), and fully open source releases (C4AI Aya).

Accordingly, when defining where a model sits in relation to the gradients of openness, the purpose of release and permitted uses are crucial aspects to consider in addition to any technical considerations. Factoring the purpose of release, intended uses, and consequently the licenses applied are as follows:

- Aya’s components are available under a fully permissive Apache 2.0 license and intended for other researchers and developers to adapt, use and build, to advance the capability and safety of LLMs for underserved languages.[17]

[15] Solaiman, I., (2023). The Gradient of Generative AI Release: Methods and Considerations. <doi.org/10.48550/arXiv.2302.04844>.

[16] World Economic Forum (2024) Presidio AI Framework: Towards Safe Generative AI Models. <https://www3.weforum.org/docs/WEF_Presidio_AI%20Framework_2024.pdf>.

[17] To view Aya’s components on Hugging Face see <huggingface.co/CohereForAI/aya-101>.

- C4AI Command-R is a weights-only release under a CC-BY-NC 4.0 License with an Acceptable Use Addendum which establish red-line use cases that are not appropriate for downstream use, intended to allow researchers access to a state-of-art model for non-commercial purposes such as fundamental research and safety auditing.[18]
- Cohere's proprietary models available via an API - Command, Rerank and Embed - are released under a commercial license[19]. To ensure our customers leverage our product responsibility, we've also established and linked *Usage Guidelines*[20] to the commercial licenses that specify domains where model use requires extra scrutiny and prohibit high-risk use-cases that aren't appropriate. We have systems and infrastructure in place to enforce usage guidelines. This includes rate limits, content filtering, monitoring for anomalous activity, and revoking or suspending access when necessary. We also encourage builders using Cohere's models to monitor model behavior for their use-case and incorporate redress mechanisms where Cohere is included in outputs at scale. This allows us to detect any emerging biases or harms through usage.[21]

(ii) There should not be a monolithic approach to openness, and the level of accessibility to model weights and release format should depend on case-by-case estimation of risks and benefits.

Where a model sits on a spectrum from 'fully open' to 'fully closed' brings greater or lesser degrees of certain benefits and risks.[22]

The benefits of more open foundation models include:

- *Increased access to resources for research and development that can serve communities and needs that may not be primarily met by commercial incentives*. For example, releasing Aya made a state-of-art multilingual large language model and dataset available for 101 languages, 50 of which were previously unserved by existing language models - representing over 1.25 billion language speakers globally. Similarly, C4AI Command-R offers researchers and developers access to a leading-edge frontier model to experiment, explore, and build on for non-commercial purposes, including to facilitate model evaluation and safety research.

- *Ability to enable independent and third-party model evaluations and testing, and community-led approaches to safety research*. For example, C4AI Command-R is a weights-only release licensed for non-commercial uses, intended to allow researchers to better understand a state-of-art model, evaluate how it works, and to use it in research that improves collective understanding of how to build and deploy language models responsibly and safely. This release helps to

[18] To view C4AI's model weights on Hugging Face see <huggingface.co/CohereForAI/c4ai-command-r-v01>.

[19] To view Cohere's: Terms of Use (see<https://cohere.com/terms-of-use>) and SaaS Agreement (see<https://cohere.com/saas-agreement>).

[20] To view Cohere's Usage Guidelines see <https://docs.cohere.com/docs/usage-guidelines>.

[21] To view our recommendation for evaluation techniques and tooling see< *https://txt.cohere.com/evaluating-llm-outputs/>.*

[22] Seger, E. et al. (2023) 'Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives'. Available at: <https://doi.org/10.2139/ssrn.4596436>.

address gaps in researcher and developer access to state-of-art model weights, the appetite for which is demonstrated by the 24,500 downloads of C4AI Command-R within the first three weeks of release.[23] Additionally, our experience in developing and releasing Aya in collaboration with 3000+ researchers globally demonstrates how open model weights can be adapted, improved, analyzed, safety-mitigated, and reproduced by a wide open-science research community, accelerating progress towards safer and more responsible language models through community-led approaches.

At the same time, release of open weights can amplify certain risks, including reduced traceability of downstream users and uses, and the possibility of misuse - which we define as the intentional use of foundation models to cause harm. For instance, if model weights are fully released without API restrictions, a primary concern is that visibility and control over downstream use is relinquished. In these cases, efforts to limit safety for sensitive downstream use can be undermined. For example, threat models of concern include downstream finetuning of open weights that undoes safeguards[24] or adversarial approaches to circumventing or inferring safeguards.[25]

The benefits and risks across the different gradients of openness are summarized by the below table which shows that as 'openness' increases, certain risks and benefits increase and vice versa.

	Increased openness	**Decreased openness**
Benefits	● Greater access to resources for research and development that can serve communities and needs that aren't met by commercial incentives. ● Enables independent model evaluations and testing, and community-led approaches to safety research.	● Greater control over access and downstream uses, reducing risk of misuse. ● Greater traceability of downstream users and uses.
Risks	● Less control over downstream uses, which could increase risk of misuse. ● Less traceability of downstream use, making model proliferation hard to monitor.	● Less ability for independent and third-party research and evaluation. ● Limited access to resources for research and development that can support underserved communities.

[23] C4AI Command-R was released on 11 March 2024. As of Tues 26 March there were 24,558 downloads reported on the C4AI Command-R-v01 Hugging Face repository: <huggingface.co/CohereForAI/c4ai-command-r-v01>.

[24] Chan, A. et al. (2023) 'Hazards from Increasingly Accessible Fine-Tuning of Downloadable Foundation Models'. arXiv. < http://arxiv.org/abs/2312.14751>.

[25] Rando, J. et al. (2022) 'Red-Teaming the Stable Diffusion Safety Filter'. arXiv. <https://doi.org/10.48550/arXiv.2210.04610>.

Because of the non-binary and sliding-scale nature of benefit and risk associated with open foundation models, more work is needed to understand what release formats best balance risk and benefits in different contexts. Amplified risks do not necessarily mitigate the benefits of openness, or vice versa, but it must be acknowledged in terms of understanding the trade-offs with openness. As such, approaches to open foundation model release that can account for the different levels of risk while enabling certain benefits, such as controlled API access and partial access models, should continue to be explored as formats which allow some traceability while providing benefits in terms of access.

(iii) It is important that licenses and access or use restrictions reflect perceived risks.

Approaches to applying licenses to model weights that prescribe acceptable uses should: (a) reflect the sliding scale nature of model openness; and (b) access to model weights should be tailored to reasonably foreseeable risks.

For example, for projects such as Aya, which serve underserved communities and where the risk of misuse is outweighed by wide societal benefits, we have supported an open license. Aya was released on a more open and permissive license to enable important acceleration of language model capabilities and applications for communities not served by current commercial models.

However, for state-of-art model weight releases where the capabilities of the model raise important concerns about amplification of model risk – like misinformation – a more restrictive license can emphasize red lines for possible misuse. This is the case for C4AI Command-R, which is released under a restrictive non-commercial use license. We believe this appropriately balances between enabling important research access for safety researchers and preventing the worst misuse. We continue to advocate for a nuanced case-by-case approach to decide on the appropriate terms of use to balance the benefits of openness and amplified risk.

(iv) Considerations must go beyond the degree of weight release to also consider questions such as access to compute and expertise.

Releasing weights does not ensure models are accessible because of the current engineering scale: to successfully make use of state-of-art model weight releases requires access to hardware, skills and expertise that is not widely available. This means that researchers and communities may be left unsupported due to a lack of access to state-of-art models. Accordingly, considerations of access to open weights should be accompanied by considerations of national compute programs and access both for technical innovation and for broadening access in participation. C4AI's Research Grants program[26] provides academic researchers with credits to access Cohere's APIs at no cost for benchmarking and experimentation, and offers another example of how a degree of model openness (API access) helps address gaps in access to machine learning resources for independent research.

[26] Cohere For AI Research Grant Program <txt.cohere.com/c4ai-research-grants/>.

(v) The regulatory approach to open foundation models needs to be dynamic, proportionate, and contextual, taking a risk-and-principled approach.

As the NTIA has noted in the RFC consultation paper, there is currently no consensus on what is considered 'open' or 'widely available'. Similarly, the potential risks and societal impacts associated with the release of open foundation model weights versus their potential benefits are a matter of intense debate. Consequently, regulating these models requires a nuanced approach that considers their distinctive properties and the broader implications.

To the extent NTIA considers advancing regulation or other policy approaches as it relates to open foundation models, it should ensure that such an accountability framework is risk-and-principles based and takes a contextual and proportionate approach, including in addressing how 'openness' and 'widely available' are defined. It can do so by taking into consideration the following factors:

- model capabilities (e.g., whether the open foundation model has frontier dual use model capacity versus non-frontier general purpose model capabilities);
- availability and degree of openness of existing models with similar capabilities;
- purpose of release (e.g., is it for non-commercial or research purposes);
- licensing regime and permitted uses under which foundation model weights are released;
- entity releasing the open foundation model and their role in the AI-ecosystem;
- nature and scope of release, including whether other components (e.g., training data, code, compute access) are shared as part of the release;
- impact of such a release on the societal risk analysis (e.g., risks relating to biosecurity, cybersecurity, disinformation, among others), and whether the risks posed outweigh the societal benefits associated with dissemination of such open foundation models (e.g., transparency, promoting research etc.); and
- the state of the technology and research surrounding open foundation models, including ability to control and monitor downstream risks (and deployment), and the understanding of risk and society impacts of foundation models.

It is evident that a one-size-fits-all approach will not adequately address the nuanced and evolving nature of these risks and opportunities and AI model capabilities. Accordingly, the regulatory landscape for open foundation models must account for the dynamic nature of AI development, continually assessing and addressing the emerging risks posed by advances in model capabilities while fostering the opportunities made possible through innovation. Any approach advanced by the US surrounding open foundation models should also take into consideration the need for such regimes to be interoperable with standards and policies advanced at the international level. Such a proactive approach shall ensure that safeguards keep pace with technological advancements.

We recognise and commend the NTIA's efforts to gather evidence to inform their forthcoming report to the President on the potential benefits, risks, and implications of dual-use foundation models for which the model weights are widely available, as well as policy and regulatory recommendations pertaining to those models. We hope our perspective contributes to the task of navigating nuances across the different commercial and non-commercial purposes of foundation models and supporting contributions that both more ‘closed’ and more ‘open’ foundation models make to society and the economy while advancing safe, responsible innovation in machine learning.

Thank you for the opportunity to provide comments on AI accountability. We look forward to serving as a resource as you continue to engage in policy discussions on this issue.