ANTHROP\C

548 Market St., PMB 90375
San Francisco, CA 94104-5401

412-837-9797
anthropic.com

February 2, 2024

National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

RE: Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11)

Submitted via: Regulations.gov

Thank you for the opportunity to respond to the National Institute of Standards and Technology's (NIST) Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence. We understand that the purpose of this request is to inform "an initiative for evaluating and auditing capabilities relating to Artificial Intelligence (AI) technologies and to develop a variety of guidelines, including for conducting AI red-teaming tests to enable deployment of safe, secure, and trustworthy systems."[1]

At Anthropic, we believe that a core tenet of AI safety is the ability to accurately describe and measure the capabilities and safety characteristics of AI systems. This ability is both an enabler and a prerequisite to effective regulation, as measurement tools allow us to objectively describe the capabilities of AI systems and ensure they meet appropriate safety thresholds. To that end, we strongly encourage NIST—as the leading U.S. federal agency on measurement science and standards development—to prioritize building, running, maintaining, verifying, and sharing the results from authoritative benchmarks for AI systems. Such benchmarks will help to establish norms and standardize practices for the evaluation and reporting of AI systems and will in turn drive transparency and trust in AI systems.

About Anthropic

Anthropic is an AI safety and research company working to build reliable, interpretable, and steerable AI systems. Our mission is to develop and deploy advanced AI systems that are helpful to people. Core to our mission is the development of cutting-edge, or "frontier" large language models—we use these to conduct empirical safety research and as the main ingredient in the systems we deploy commercially.

[1] 88 Fed. Reg. 88368 (Dec. 21, 2023).

Prioritize Building an Authoritative Benchmark for Generative AI Systems

Anthropic strongly encourages NIST to direct its limited resources towards building a robust and standardized benchmark for generative AI systems. While individual companies like Anthropic can (and will) continue to conduct internal tests and publish results, benchmarks and evaluations have greater impact when operationalized and distributed by a trusted independent third party like NIST. That is why we believe that, in the near term, NIST will have the greatest impact by constructing authoritative benchmarks and evaluations optimized for generative models across a range of capability and safety issues.

Indeed, NIST is well-positioned to serve as a trusted and neutral third party for the measurement of generative AI models due to the agency's expertise and proven track record developing measurement infrastructure—for example, NIST successfully implemented the Face Recognition Vendor Test (FRVT), which provides for the independent evaluation of commercially available facial recognition systems. By publishing the FRVT performance benchmarks, NIST created transparency amongst products using facial recognition and helped consumers and regulators to better understand the capabilities and limitations of different facial recognition systems. The FRVT also created a race-to-the-top by incentivizing facial recognition developers to perform well on evaluations across different populations. Anthropic believes that NIST should do the same for generative AI systems by building a robust benchmarking infrastructure, making it accessible to external developers and regulators for independent assessments, and publishing results to drive transparency, better systems for users and customers, and enable oversight.

As NIST works to build its own generative AI benchmarking capabilities, we would encourage the agency to consider prioritizing an opinionated and dynamic set of the most important benchmarks for model developers to run.[2] NIST should also reference existing efforts like BIG-bench[3] and HELM,[4] as well as other comprehensive multi-evaluation suites, with a focus on enabling functionality across diverse systems. By emphasizing interoperability and standardization as core design goals, NIST can maximize adoption of its benchmarks by all sectors and thereby encourage a safety race-to-the-top that promotes transparency across models, enables safer AI systems, and responsibly advances innovation.

Finally, we encourage NIST to continuously verify any eventual benchmarks it develops by running evaluations across the latest generative AI models. This in-house measurement by NIST will serve multiple purposes, including enabling seamless integration and out-of-the-box compatibility with commercial generative engines, verifying rigor, simplifying external usage, and providing insights to guide improvement. Additionally, NIST should continue to convene experts from across industry, academia, government, and civil society to collaboratively and continuously refine and improve the benchmarks.

[2] *See, e.g.*, Anthropic, Challenges in Evaluating AI Systems, Oct. 4, 2023, *available at*: https://www.anthropic.com/news/evaluating-ai-systems (accessed Feb. 2, 2024).
[3] *See generally*, BIG-bench, available at: https://github.com/google/BIG-bench (accessed Feb. 2, 2024).
[4] *See generally*, HELM, available at: https://crfm.stanford.edu/helm/lite/latest/ (accessed Feb. 2, 2024).

Develop Robust Guidelines and Areas of Concern for AI Red Teaming

Red teaming, or adversarial testing, is a recognized technique to measure and increase the safety and security of systems. As AI systems become more capable, expanded red teaming efforts are integral to identifying and mitigating emerging risks. While we do not believe that the systems available today pose an imminent concern, robust red teaming will help to proactively ensure the safety of future, more powerful systems.[5]

To this end, we encourage NIST to implement guidelines advising AI developers, especially of dual-use foundation models, to institute robust red teaming programs in collaboration with domain experts. Red teaming guidelines should encourage generative AI developers to deeply probe system capabilities, build automated risk evaluations, partner with trusted third parties to handle sensitive findings with stringent information security, and urge ongoing vigilance, responsibility and collaboration among developers, regulators and experts. Additionally, effective red teaming should, at a minimum, yield quantifiable, automated tests, and expert elicitation guidance. In particular, we believe that AI developers would benefit from publicly shareable automated safety tests and protocols for controlled trial-like red teaming exercises.

Recognizing the challenges of red teaming AI systems,[6] we further encourage NIST to work across the federal government to identify core areas of risk where companies should focus their red teaming efforts. NIST should also consider issuing an RFI to better understand the availability and maturity of third-party red teaming evaluations[7]. This RFI will assist the federal government and industry identify gaps and barriers currently limiting broader adoption of third-party red teaming services and could shed light on the level of expertise, time requirements, and costs associated with robust third party testing.

Convene Stakeholders to Define Goals and Attributes for Watermarking Standards

Watermarking shows promise as a tool to discern human- from AI-generated text. However, watermarking has potential limitations, including the ability of bad actors to override schemes, false positives undermining public trust, and privacy concerns if keys are compromised. Given the complications of watermarking, we encourage NIST to convene a multi-stakeholder group—comprising government, industry, academia, and civil society—to define goals and attributes for watermarking standards that meet society's needs. We believe that with thoughtful policies on its appropriate uses and a greater ability to assess its technological maturity, watermarking could support discernment of human- and AI-generated text as part of a comprehensive approach; however, human oversight, institutions, and critical thinking will remain essential.

[5] *See, e.g.,* Anthropic, Frontier Threats Red Teaming for AI Safety, July 26, 2023, *available at*: https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety (accessed Feb. 2, 2024); Ganguli, Lovitt, et. al., Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned, Nov. 22, 2022, available at: https://arxiv.org/abs/2209.07858 (accessed Feb. 2, 2024).

[6] *Id*.

[7] Red teaming is a nascent and fast-developing part of the AI ecosystem. Conducting a public RFI could help to make this part of the AI ecosystem more legible to both policymakers and AI developers who might want to use red teaming services and products.

Empowering NIST Through Robust Funding

Developing benchmarks and evaluations is challenging, expensive, and resource intensive, and NIST must be robustly funded to support this work.[8] Benchmarking remains a pragmatic, meaningful approach for the near term, and improved measurement capabilities can beneficially steer AI progress. In addition, NIST will require significant new resources to stand up the new AI Safety Institute, which will be critical for supporting U.S. competitiveness, innovation, and leadership on AI regulation. Anthropic strongly supports increased funding for NIST to support this important work and position the agency as a global AI safety leader.[9]

Conclusion

We applaud NIST's leadership in charting a course ahead for the responsible development and deployment of safe AI systems. In light of resource constraints, we encourage NIST to prioritize the development of an authoritative benchmark for AI systems to complement development of a companion to the AI risk management framework. Fundamentally, compliance with any framework depends on the quality of that framework's underlying benchmarks. Directly creating and curating reliable benchmarks will enable NIST to have the greatest impact on the safe development and deployment of AI systems.

[8] *See,* Anthropic, Challenges in Evaluating AI Systems, Oct. 4, 2023, *available at:* https://www.anthropic.com/news/evaluating-ai-systems (accessed Feb. 2, 2024).

[9] *See,* Anthropic, An AI Policy Tool for Today: Ambitiously Invest in NIST, Apr. 20, 2023, available at: https://www.anthropic.com/news/an-ai-policy-tool-for-today-ambitiously-invest-in-nist (accessed Feb. 2, 2024).