ANTHROP\C

Strengthening U.S. AI Innovation Through an Ambitious Investment in NIST

Progress in artificial intelligence (AI) has rapidly advanced such that contemporary AI systems are poised to transform a number of areas of society. Large language models (LLMs) in particular have become increasingly powerful, and we stand at an inflection point for widespread integration of these models into everyday life. America's ability to harness their potential depends on whether we can systematically evaluate the capabilities, limitations, and potential harms of LLMs. Effective measurement tools are also a necessary precursor for sensible AI policy. Given this, we need to significantly enhance the government's ability to build and use these means of quantitative assessment.

The National Institute of Standards and Technology (NIST), the nation's leader in measurement science and technical standards, is uniquely positioned to lead this critical work that is central to promoting U.S. innovation. **We argue that a strongly resourced AI program at NIST — on the order of 22 additional positions and an increase of $15 million over FY 2023 levels — can advance research and development (R&D) through the creation of a robust ecosystem of AI assurance.** This would involve doubling the staff at NIST working on AI programs and would create a robust improvement in policy making capacity.

NIST's work on standards development,[1] the AI Risk Management Framework,[2] and the Face Recognition Vendor Test (FRVT)[3] has set

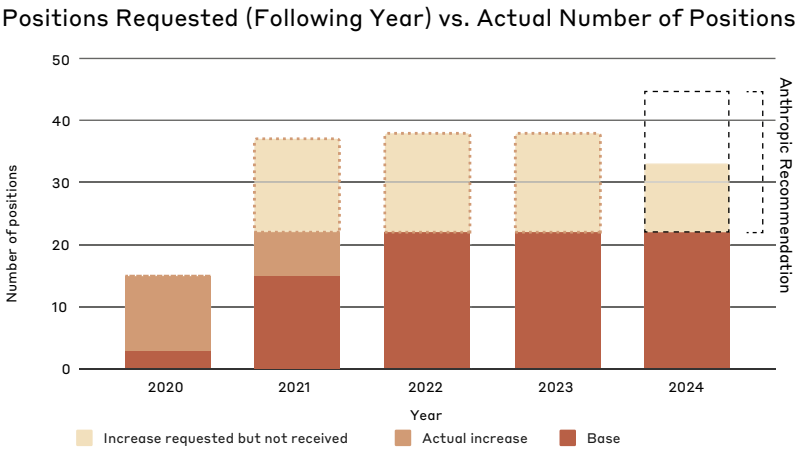


Fig 1. Number of staff positions working on AI-related programs at NIST from 2020-2024. Red bars represent the base number of positions each year, orange bars *combined* with yellow bars represent requested increase in AI program positions, and yellow bars represent the actual increase in number of positions over the previous year. Base positions for 2024 TBD and are presented as constant from 2023 staffing levels. Dashed black lines represent Anthropic recommendation of 22 additional positions over FY 2023 (inclusive of the estimated 11 positions already requested by NIST).

1 National Institute of Standards and Technology. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence
2 National Institute of Standards and Technology. (2023). AI Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework
3 National Institute of Standards and Technology. (2010). Face Recognition Vendor Test (FRVT). https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

the foundation for an AI assurance ecosystem. Congress can further increase the impact of this work by making an ambitious investment in NIST in 2024 and beyond. **In doing so, Congress will enhance public trust in AI, provide the government with confidence in the safety of AI systems, and encourage industrial competitiveness.**

During the time NIST has been active in AI measurement efforts, we have witnessed rapid and tremendous technological progress. AI systems now score in the 90th percentile of a simulated bar exam,[4] accelerate biology research by predicting the 3D models of protein structures,[5] and generate photo-realistic images from plain text descriptions.[6] However, the government's ability to meaningfully assess these systems for potential capabilities and risks hasn't kept pace.[7] In our review of past budget submissions and omnibus spending bills, we found a general under-resourcing and concerning stagnation in funding for AI programs at NIST.

On personnel, the FY 2020 budget submission appears to have been the last time NIST's AI-related headcount request was met (12 new positions added at the start of 2021).[8]

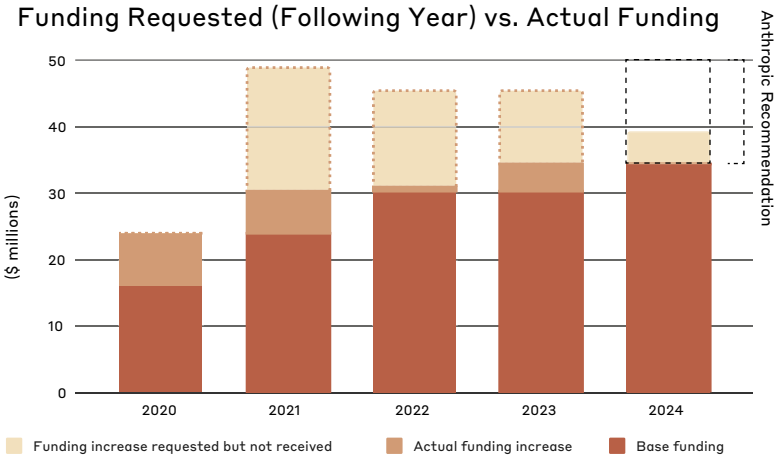


Fig 2. Amount of funding (in millions) for AI-related programs at NIST from 2020-2024. Red bars represent the base budget each year, orange bars combined with yellow bars represent funding increase requested by NIST for AI programs over previous year, and yellow bars represent actual increase in funding awarded. 2024 amounts reflect the current FY 2024 budget request. Dashed black lines represent Anthropic recommendation of a $15 million increase over FY 2023 (inclusive of NIST's $5 million request).

Subsequent years exhibit unmet staffing requests and the data show a concerning plateau in staffing levels from the start of 2022 onward (Figure 1). Funding levels have been similarly modest (Figure 2). The FY 2020 budget submission was the last time NIST's funding request was met in full, and financial support stagnated from FY 2022 to FY 2023. The FY 2024 budget request represents only a modest increase of $5 million dollars for AI-related programs[9] (down from a requested $25 million for FY 2021).[10] This reduced financial support has happened against the backdrop of publicly deployed, increasingly capable AI systems.

While ambitious in scope, our recommendation is not unprecedented — it represents approximately the *same* number of positions as the request for FY 2021 and the (partially met) funding requests for FY 2022 and FY 2023. We encourage the Subcommittee on Commerce, Justice, Science, and Related Agencies, of both the U.S. House and Senate Appropriations Committees, to consider this an investment in the development of safe and innovative AI systems that can benefit all Americans.

4 OpenAI. (2023). GPT-4. https://openai.com/research/gpt-4
5 DeepMind. (2021). AlphaFold. https://www.deepmind.com/research/highlighted-research/alphafold
6 Google Research, Brain Team. (2022). Imagen. https://imagen.research.google/
7 Whittlestone, J., & Clark, J. (2021). Why and How Governments Should Monitor AI Development. arXiv. https://arxiv.org/abs/2108.12427
8 National Institute of Standards and Technology. Budget Submission to Congress for Fiscal Years 2020, 2021, 2022, 2023, and 2024. Note that in some years, AI-related programs are combined with other programs, so personnel and funding requests are approximate measures.
9 National Institute of Standards and Technology. (2023). Fiscal Year 2024 Budget Submission to Congress. https://www.commerce.gov/sites/default/files/2023-03/NIST-NTIS-FY2024-Congressional-Budget-Submission.pdf
10 National Institute of Standards and Technology. (2020). Fiscal Year 2021 Budget Submission to Congress. https://www.commerce.gov/sites/default/files/2020-02/fy2021_nist_ntis_congressional_budget_justification.pdf

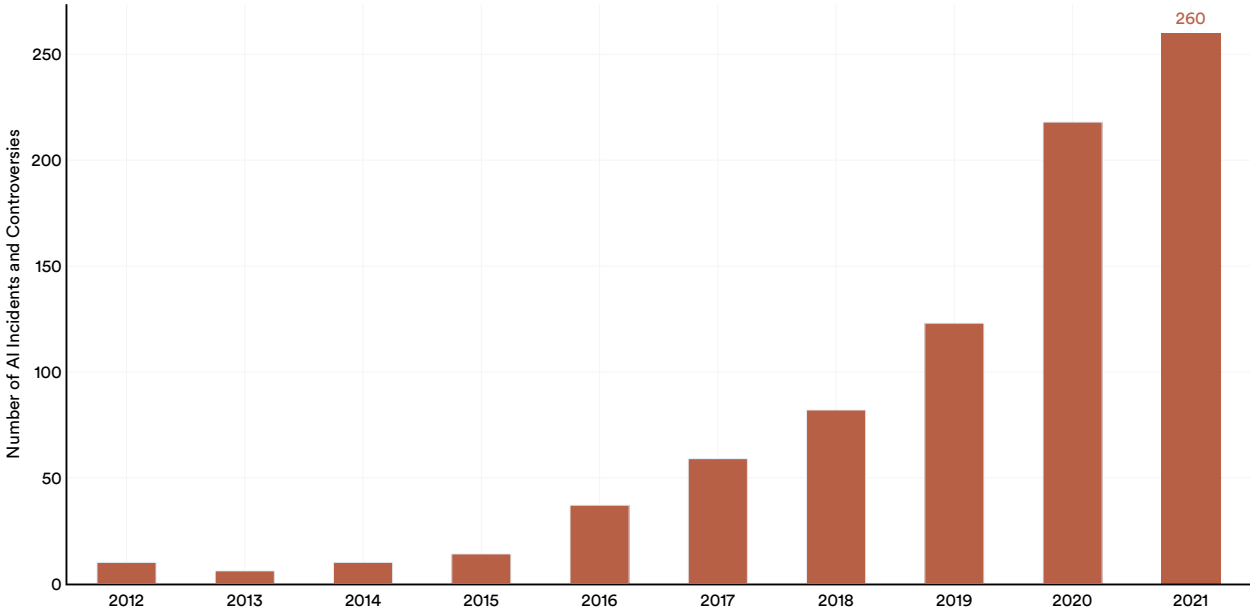
With this additional resourcing, NIST could continue and expand its work on AI assurance efforts like:

- Cataloging existing AI evaluations and benchmarks used in industry and academia
- Investigating the scientific validity of existing evaluations (e.g., adherence to quality control practices, effects of technical implementation choices on evaluation results, etc.)
- Designing novel evaluations that address limitations of existing evaluations
- Developing technical standards for how to identify vulnerabilities in open-ended systems
- Developing disclosure standards to enhance transparency around complex AI systems
- Partnering with allies on international standards to promote multilateral interoperability
- Further developing and updating the AI Risk Management Framework

More resourcing will allow NIST to build out much-needed testing environments for today's generative AI systems, including LLMs that power various productivity tools and chatbots. Recent large-scale AI deployments have highlighted a unique characteristic of these systems, which is that they contain both useful capabilities and harmful behaviors that can not always be anticipated during the development process. The lack of available measurement techniques to identify and measure these behaviors exacerbates this problem — harmful behaviors can emerge unpredictably and may only become apparent once the AI system is deployed in real-world settings.[11]

To address this, NIST could develop a testbed similar to FRVT, but one focused on LLMs instead of facial recognition systems. **FRVT serves as a proof point in how the development of testbeds can both enhance safety and promote innovation as developers compete to build better technology.** With additional personnel and financial support, NIST could create a voluntary, centrally run testbed that helps developers catch risks prior to public deployment and benchmark their AI systems against the current state of the art.

Number of AI Incidents and Controversies, 2012–21
Source: AIAAIC Repository, 2022 | Chart: 2023 AI Index Report



This figure does not consider AI incidents reported in 2022, as the incidents submitted to the AIAAIC database undergo a lengthy vetting process before they are fully added.

11 Ganguli, D., et al. (2022). Predictability and Surprise in Large Generative Models. ACM Digital Library. https://dl.acm.org/doi/abs/10.1145/3531146.3533229

There is an urgent need for this work. AI systems are rapidly moving from controlled lab settings into the world. Alongside this, we've seen a concerning rise in the number of incidents where AI systems don't perform as intended or perform in harmful ways.[12] With increased funding, NIST can build on its work on responsible development standards and design new guardrails to help avoid future incidents.

Ambitiously investing in NIST will be key to building a robust ecosystem for AI assurance. In doing so, Congress can increase public confidence in AI technology, enable a competitive R&D environment, and foster a more productive U.S. economy. Anthropic would be delighted to talk further about the importance of a robust measurement capacity with policymakers both in the United States and abroad.

ABOUT US

Anthropic is a public benefit corporation and AI safety research company that is working to build reliable, interpretable, and steerable AI systems.

12 Maslej, N., et al. (2023) The AI Index 2023 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. https://aiindex.stanford.edu/report/