Benchmarking for Breakthroughs

How to incentivize AI for national priorities through a strategic challenge and evaluations program

Séb Krier and Zhengdong Wang

Benchmarking for Breakthroughs

How to incentivize AI for national priorities through a strategic challenge and evaluations program | Séb Krier and Zhengdong Wang

---

This essay is part of The Launch Sequence, a collection of concrete, ambitious ideas to accelerate AI for science and security.

Summary

Grand challenges and evaluations have greatly influenced AI capabilities. Today's challenges and evaluations can benefit from more public incentive alignment, more expertise from the public sector, and the support of public endorsement. We propose the creation of a Targeted Evaluations for Long-term Objectives in Science (TELOS) program office to systematically commission grand challenges and evaluations for the public good. We envision that this agile, efficient program office will direct private research and development towards areas that are critical to American leadership. If successful, this pilot may be expanded to a permanent institute that could accelerate breakthroughs in the nation's most ambitious, urgent, and solvable challenges by decades.

Motivation

Recent advances in AI defy easy summary. New capabilities, such as native multimodality, advanced reasoning, long-context processing, and even nascent agency emerge monthly. AI now outperforms humans on many complex tasks: it can code, answer PhD-level questions, and even pinpoint geolocations from

nondescript images. Yet it still struggles with basic math, hallucinations, and coherence over time.

Some take this "jagged frontier," where AI is superhuman at some tasks but makes basic errors, as a reason to discount progress. We see it instead as highlighting a strategic opportunity for government: challenges and evaluations. As White House OSTP Director Michael Kratsios observed, "What we target is what we measure, and what we measure is what we get more of." In an uneven landscape, what gets measured determines where progress concentrates. Research labs steer the development of AI systems by aiming for high performance on specific benchmarks.

This evaluation ecosystem has been remarkably effective, but also commercially-driven and emergent, with predictable blind spots. Industry favors benchmarks that improve immediate product viability. Academia rewards specialized tasks suited for publication. That leaves a gap: foundational challenges with high scientific value that require coordinated public investment.

The unreasonable effectiveness of challenges and evaluations

The remarkable capabilities of today's AI systems were shaped by an ad hoc, emergent history of evaluating general intelligence as much as they were by data and computational power. Grand challenges focus research in the field. Some considered Google DeepMind's AlphaGo system to have defeated the 18-time Go World Champion Lee Sedol as having hit a major AI research milestone a decade ahead of its time. In protein structure prediction, the Critical Assessment of protein Structure Prediction (CASP) competition made progress on the grand challenge legible. Google DeepMind's AlphaFold system — widely considered to have solved the problem — depended on CASP as the "gold-standard assessment" for the field. Most recently, the privately sponsored Vesuvius Challenge has for the first time identified new text from an ancient carbonized Herculaneum scroll. When AI research converges around grand challenges, it makes progress at pace.

Day to day, benchmarks like MMLU (Massive Multitask Language Understanding), GPQA (Google-Proof Q&A), and BIG-bench (Beyond the Imitation Game)

significantly influenced early language model capabilities across abstract algebra, logical fallacies, American history, and more. Labs also began reporting scores on familiar human exams like the bar, LSAT, GRE, SAT, and the American Invitational Mathematics Examination to compare model intelligence to human intelligence. Meanwhile, newer platforms like Chatbot Arena, a head-to-head leaderboard where users rank AI systems, have emerged as a de facto standard for judging the "best" model, cited by Fortune 500 executives and influencing millions in trading volume. These evaluations serve as highly visible yardsticks for perceived progress and competitive positioning. Leading labs including Google DeepMind, OpenAI, and Anthropic prominently report their scores on them. Whenever an evaluation becomes well-used, the capability it measures improves. When FrontierMath was launched in 2024, no model scored better than 2%. A few months later, the highest scores now reach 25%.

The role and advantage of government

Today's ecosystem of challenges and evaluations is well-suited to academic and commercial research. But it holds untapped potential to align AI development with areas critical to American leadership. The case for a strategic government program rests on three coordination failures:

1. **Public incentives.** Current benchmarks tend to steer AI research towards commercial use cases which are not always long-term, high-impact societal challenges. For example, AI research may underinvest in evaluations for escalation risk in high-stakes geopolitical situations, or breakthroughs in energy resilience. These problems do not promise easy returns. Even when firms are incentivized to evaluate these cases, they may not be incentivized to share them to the national commons. Publicly commissioned benchmarks would direct private research and development towards national priorities, and strengthen shared scientific infrastructure for cheap.
2. **Public expertise.** Publicly commissioned benchmarks will benefit from government expertise in problems on the scale of the entire nation. Even if a firm or nonprofit wanted to evaluate its AI systems in its capabilities in solving societal issues such as energy grid resilience, supply chain weaknesses, or state-level cybersecurity, it would not immediately have

access to the wealth of expertise honed in the civil service. In turn, the public sector would stay abreast of the frontier. Researchers face pressure to publish positive results, which obscures true progress by fueling survivorship bias and comparisons to outdated baselines. This is untenable for policymaking. Public expertise can strengthen and be strengthened by commissioning benchmarks in a virtuous cycle.

3. **Public credibility.** Evaluations are low-cost and high-leverage, a prime candidate to benefit from the high social returns to public research and development. While frontier model training costs billions, only a few committed individuals are needed to launch influential benchmarks. In addition to the aforementioned FrontierMath, the Center for AI Safety and Scale AI developed Humanity's Last Exam. The ARC Prize Foundation offers a million-dollar prize for an evaluation it believes defines general intelligence. Grassroots evaluations, once they gain traction, drive improved performance in what they measure. The public sector can replicate and vastly magnify the visibility and prestige of the benchmarks it commissions. If small teams can bring challenges and evaluations into prominence, one with the full support of the American government and people is sure to make an impact.

Solution

We propose that the US government should establish a Targeted Evaluations for Long-term Objectives in Science (TELOS) program office. This office would commission AI grand challenges and evaluations that focus on areas critical to the national interest but underserved by current incentives.

TELOS activities and strategy

Activities: What should TELOS do?

TELOS would function as a strategic commissioning body, not an internal evaluator. Rather than building benchmarks in-house, TELOS would identify, fund,

and endorse challenges and evaluations, delegating development and maintenance to expert third parties. The office should also welcome proposals from the wider research community. This model lets the government leverage the distributed expertise of academia, industry, and nonprofits without duplicating their efforts. The office may fund one of two instruments:

1. A **grand challenge** is a long-term scientific goal, often intractable without a step change.
2. An **evaluation** is a precise, measurable benchmark used to track progress toward that goal or assess a specific AI capability.

TELOS would operate through a three-pronged approach:

1. **Proactive commissioning**. A small, technically fluent team of program managers would identify high-priority needs and issue public tenders. Embedded in research communities, they would track emerging bottlenecks and opportunity areas.
2. **Open review of outside proposals.** To remain flexible and innovative, TELOS would accept unsolicited proposals from researchers and institutions. This bottom-up channel helps ensure relevance and builds long-term buy-in for challenges and evaluations as community standards.
3. **Support existing benchmarks.** TELOS could also improve and legitimize existing challenges and evaluations — offering stability, funding, and reputational lift to promising efforts that would otherwise lack institutional support.

Once a challenge or evaluation is built and run by the grantee, TELOS would endorse and host a public leaderboard. This final step confers visibility, legitimacy, and status, turning benchmarks into powerful signals that attract talent and resources. Research groups would understand that these benchmarks measure national priorities. Progress would confer prestige on those who make it happen. In this way, TELOS is less a developer than a catalyst and standard-bearer, steering attention towards what matters.

Targeting strategy: What kind of challenges and evaluations should TELOS commission?

TELOS will focus on identifying and funding critical, underserved areas where AI could drive breakthroughs for national priorities. Commissioned benchmarks must be clearly important, technically demanding, address a real gap, and offer the potential to transform a field. Demis Hassabis, the CEO of Google DeepMind, has outlined three traits of problems well-suited to AI: a vast search space, a clear objective function, and abundant data or a reliable simulator. TELOS will look for these conditions when selecting targets.

In the service of American leadership, promising target areas abound. They include accelerating clinical trials, discovering novel materials, such as improved carbon-capture technologies or green hydrogen catalysts, building adaptive cyber defense agents for real-time threat detection and response, and even understanding frontier models themselves through mechanistic interpretability. Just this June, the UK government announced an £8 million investment to collect data for drug discovery. No scientific ambition is too great. Commercial and academic incentives do exist to tackle these problems, but research in metascience points to fundamental R&D gaps that these incentives have not yet closed.

The primary function of TELOS is not to prescribe solutions. But here we will go into detail on one representative example to offer a glimpse of what success looks like: accelerating clinical trials. The clinical trial process is often a multi-billion dollar, decade-long gauntlet that stifles medical innovation. Nearly 90% of promising drugs fail, often due to a simple lack of efficacy discovered only after immense investment. Emerging AI models show promise in predicting outcomes by analyzing preclinical data, trial protocols, and biomarkers. But progress is blocked by a public goods problem: no single actor has the incentive to create the high-quality, standardized dataset needed to train reliable models for the entire nation. Proprietary, messy data remains siloed.

This is precisely the kind of grand challenge TELOS could address. It could fund a consortium to build a large, anonymized dataset of historical trials, then launch a benchmark with a clear objective: predict Phase III success from early data. Beyond specific problems, TELOS could fund evaluations probing

meta-capabilities, such as assessing the practical synthesizability of proposed drug compounds, or inverse design, where AI systems must generate candidates with target properties. A public leaderboard would cut costs, derisk development, and accelerate life-saving treatments.

Implementing the TELOS Program Pilot

How should TELOS be structured and financed?

TELOS should begin as an agile program office within an existing federal entity, such as NIST, to leverage established expertise and enable a rapid start, with potential to evolve into an independent institute based on demonstrated success.

The initial pilot would run for 12 months, staffed by around 10 program managers. It would begin with intensive scoping. Within the first three months, an inter-agency working group (e.g., NSF, NIH, DOE, DARPA, NIST, CAISI), supported by an external expert panel, would run workshops across the scientific ecosystem to identify five high-priority challenges. These would then be scoped into promising evaluations and benchmarks.

Ambitious evaluations require substantial resources and broad community engagement. Humanity's Last Exam mobilized a thousand contributors globally with a $500,000 prize pool, with far more actual costs. For example, developing the TheAgentCompany benchmark involved 3,000 person-hours across engineers, researchers, and project managers, representing nearly $1 million in labor alone, excluding prize funds and infrastructure.

We propose budgeting $5 million per evaluation, covering labor, infrastructure, and incentives, totaling $25 million for five evaluations. An additional $25 million would fund TELOS operations: staffing, workshops, and expert contributions. This brings the pilot's total to $50 million.

Success metrics and future scaling

Pilot success would be measured by:

1. Commissioning five high-impact evaluations

2. Uptake by leading AI labs (e.g., DeepMind, OpenAI, Anthropic) as benchmarks for model evaluation
3. Early signs of research progress on the selected metrics

Even if the grand challenges remain unsolved in the short term, evidence of traction and community engagement would validate the TELOS model. If successful, the office could transition into a permanent National Evaluation and Challenge Institute housed within NIST, scaling up modestly in staff and number of challenges and evaluations supported.

Operating principles

Agility is essential. Programs like Operation Warp Speed, Fast Grants, UK ARIA, and the CHIPS Program Office show that rapid, high-impact innovation is possible within government. Given the pace of AI development, TELOS must deploy evaluations quickly to keep up with AI capabilities.

To foster this, TELOS must flexibly engage outside experts from academia and industry, free from rigid hiring constraints. It should be seen as a credible, appealing place for top technologists to work. Building this kind of high-talent, high-trust institution means embracing iterative development, taking calculated risks, and learning from failures, rather than trying to avoid them at all costs.

TELOS should avoid a rigid, top-down agenda. Instead, entrepreneurial managers should drive strategy. These technical leaders would identify "white space" opportunities, where the right evaluation could unlock progress, and shape initiatives through ongoing dialogue with research communities. Their job is to define problems at the "Goldilocks level": clear enough to act on, ambitious enough to matter, and difficult to game.

Should TELOS be successful, the nation may see breakthroughs in its most ambitious, urgent, and solvable challenges decades ahead of schedule.