GETN

Global EdTech Trialing Network

United States

Building a Responsible AI Testbed Infrastructure for U.S. Education in 2025

Version 1.0

EXECUTIVE SUMMARY

The United States is at a defining crossroads for artificial intelligence (AI) in education. Without rigorous systems for evaluating AI tools, our schools risk being flooded with unproven products and missed opportunities. A national AI testbed infrastructure can ensure safe, effective, and educator-led innovation, while leveraging AI's transformational potential to accelerate impact for all learners in every context—urban, rural, and virtual.

Given the rapid evolution of generative AI tools, existing educational research and procurement cycles are too slow to keep up. New tools emerge faster than researchers can evaluate them and faster than district leaders can process their implications. Without a coordinated testbed infrastructure, insights will arrive too late, and decisions will continue to be made based on outdated or incomplete information. **Now is the moment to align federal guidance, philanthropic resources, private investment, and public need into a single, trusted infrastructure for real-world evaluation.**

INTRODUCTION

MEETING THE URGENCY OF AI IN EDUCATION

As AI becomes increasingly woven into the fabric of our economy, infrastructure, and daily lives, ensuring its safety, reliability, and innovative potential is paramount. The solution is not to slow down the development of the technology or the tools, but to accelerate it responsibly.

In July 2025, the U.S. Department of Education issued guidance and a proposed supplemental priority to promote the responsible use of AI in education, following the release of the White House's *America's AI Action Plan*. These coordinated federal strategies signal a once-in-a-generation opportunity to harness AI not just to modernize education, but to transform learning for all through safe, inclusive, and educator-led innovation.

However, just as we rely on the Food and Drug Administration (FDA) to conduct clinical trials before approving a medicine for public use, AI products require a rigorous testing and vetting process before being fully deployed in schools.

In the education context, testbeds serve as the equivalent of a "Phase III clinical trial"—ensuring real-world performance, safety, and fairness before widespread adoption.

To meet this moment, we propose **a national AI testbed infrastructure for education—a coordinated network of real-world and synthetic learning environments, built on shared standards and designed to test what works, for whom, and under what conditions.** These testbeds would operationalize the shared American values outlined in the recent federal guidance: they would be educator-led, innovative, transparent, privacy-protective, and accessible to all learners.

This vision builds on policy momentum, including the *Executive Order on Promoting Trustworthy AI in Government*, the development of the AI Action Plan, and the call for the Research and Development Plan for AI, and is informed by years of fieldwork from the Global EdTech Trialing Network (GETN). The AI Action Plan, in particular, calls for establishing sandboxes or "AI Centers of Excellence" where researchers and companies can deploy and test their products, while openly sharing their data and findings.

In addition, a network of specialized AI testbeds for education builds upon the framework and resources of the National Artificial Intelligence Research Resource (NAIRR). As NAIRR aims to democratize access to crucial AI research assets, it provides a foundational infrastructure that educational AI testbeds can directly utilize and extend. NAIRR, currently in its pilot phase, is a U.S. government initiative designed to provide researchers and educators with access to computational power, diverse datasets, pre-trained models, AI software, and vital training materials. Its core mission is to broaden participation in AI research and development, fostering innovation and a skilled AI workforce.

At their core, AI testbeds are sophisticated sandboxes. They provide researchers, developers, and policymakers with the essential infrastructure to test, evaluate, and refine AI systems in a contained, real-world setting. This capability is not just a matter of academic curiosity; it is a critical component for national security, economic competitiveness, and building public trust in a technology poised to redefine our world.

WHY TESTBEDS?

Bureaucracy and red tape prevent technologists and researchers from gaining timely access to real-world classroom environments to test emerging technologies. At the same time, the state of the education technology solutions marketplace is fragmented, with an abundance of untested products, inconsistent evidence, and few trusted systems for verifying impact in real-world classrooms. Current research processes are unable to keep up; the lack of speed to insights leaves our education leaders in the dark when faced with making critical decisions about which instructional tools to implement, where, and for whom. Without a national

infrastructure for evaluating AI tools, educators, school systems, families, and taxpayers are left to navigate a confusing marketplace filled with unproven promises.

Testbeds play a crucial role in accelerating the development, validation, and scale of emerging educational technologies; they reduce barriers to entry, accelerate feedback cycles to improve product performance, and evaluate product effectiveness. When coordinated in networks, they can be leveraged to increase adoption and scale, creating the dynamism necessary to reshape American classrooms in the age of AI with cutting-edge technology that accelerates student learning.

WHAT ARE TESTBEDS?

The GETN white paper, "Tenets and Principles of Trialing Environments" breaks down the terminology of testbeds as: "the systematic ways in which we test 'if,' 'to what extent,' 'how,' 'why,' 'why not,' 'in what conditions,' and 'for whom' EdTech interventions work for their intended users within certain settings and places—the beds. In other words, testbeds exist as lab environments to understand 'what works, for whom, and under what conditions,' as the U.S. Department of Education's Institute for Education Sciences often asks" (p.5).

A coordinated testbed network may be composed of school districts and charter schools, as well as virtual, micro, and home school communities from diverse regions across the country. Testbeds can be hyper-local, regional, and then part of a coordinated network, allowing for local control and national learning.

By providing technologists access to a network of real-world learning environments, testbeds de-fragment the market and make the ongoing testing of emerging technologies feasible, in-turn expediting the development, validation, and scale of these emerging innovations to the benefit of the market.

The tenets and principles paper anchors on four tenets (inclusivity, infrastructure, innovation, impact) and outlines contextual and implementation principles supportive of building testbeds, informed by organizations leading this work domestically and abroad.

Within the context of generative AI, testbeds can also play a crucial role in the development of ethical, safe, and responsible models that meld foundational learning science principles and practical applications for teachers and learners.

What Must Be True for an AI Education Testbed Ecosystem

Given the generative nature of AI tools, we recommend a five-pillar approach to advance the development of an AI Testbed Ecosystem.

1. Dual-Layer Benchmarking: Accuracy and Real-World Fit

AI tools must meet two key thresholds for efficacy of use:

- **Model Accuracy and Safety Benchmarks:** These evaluate the technical performance of AI tools—including whether they hallucinate, make inappropriate recommendations, or operate outside their design scope. They also assess transparency, explainability, and alignment to core academic standards and privacy expectations. They must be aligned to recognized sector-specific trust frameworks like the EDSAFE AI SAFE benchmarks framework, addressing core issues of privacy, security, accountability, and transparency.

- **Safety and Reliability:** In high-stakes applications like education, failure is not an option. Testbeds enable the simulation of countless edge cases and real-world scenarios to ensure that AI systems perform as intended. One example of this is security: AI systems can be vulnerable to new forms of attack. Testbeds allow for the simulation of adversarial attacks to identify and patch vulnerabilities before they can be exploited in the wild.
- **Accountability:** Testbeds serve as practical laboratories for policymakers and regulatory bodies. They can be used to develop and refine standards for AI safety and performance. By observing how AI systems behave in these controlled environments, governments and industry groups can create more effective and evidence-based regulations. Initiatives like the U.S. government's TEST AI Act aim to use federal testbeds to create these necessary safeguards and standards.
- **Fairness and Transparency:** AI systems built on flawed data can repeat and even worsen existing problems in society. Testbeds give experts a controlled environment for auditing algorithms for fairness and developing techniques to mitigate harmful biases. To earn public trust, people need to understand why an AI system made a certain decision. Testbeds help create and test new ways to make AI more open and easier to explain while also assuring models are ideologically neutral.

  - **Efficacy:** Real-world conditions are often unpredictable. Testbeds allow for "red teaming," where AI systems are intentionally challenged with adversarial data and unexpected scenarios. This stress testing reveals how robust a model is and under what conditions its performance degrades, which is essential for determining its reliability in critical applications. A key tenet of scientific and engineering rigor is the ability to reproduce results. Testbeds provide a stable environment where different teams can repeat experiments to verify the original findings. This builds confidence in the claimed capabilities of an AI system.

- **Practical Readiness in Diverse Learning Environments:** Tools should be tested in the kinds of schools, diverse educational models, and learning environments that make up the American public education system: rural, urban, large, small, well-resourced, and under-resourced. Products that only function well in elite, high-tech settings are unsuitable for national deployment. We need tools that work where students are—from inner-city classrooms to rural communities with limited bandwidth. Access to AI, and the opportunity it presents, must not be a function of your zip code and download speed.

AI testbeds should integrate both benchmark types into their evaluation protocols. This ensures that taxpayer-funded tools are accurate, durable, and effective in the full range of classrooms they are intended to serve.

2. Fund the "Missing Middle" in R&D

The U.S. education technology pipeline suffers from a well-documented gap: early-stage development and late-stage efficacy trials receive investment, but the middle stages—where real-world validation and iteration happen—are often neglected. As documented by GETN, this "missing middle" prevents promising tools from maturing and scaling responsibly. Testbeds serve as critical infrastructure that bridge significant gaps in the early stages of technology development, transforming promising ideas into viable, market-ready innovations. They address fundamental challenges that often cause novel technologies to fail before they ever reach their potential.

AI testbeds can fill this gap by:

- Supporting **codesign research** with teachers, students, and families.
- Facilitating **feasibility and implementation studies** in a range of contexts.
- Generating **practical evidence** about how AI performs in the day-to-day complexity of teaching and learning.
- Building a national network anchored in diverse contexts to understand investor-driven metrics to fuel **capital investment**, such as product market fit and capacity for growth and scale.

The AGILE Network, launched by GETN-US in 2024, shows what is possible. In its first year, AGILE reduced study recruitment time by 23%, saved an average of $6K–$18K per study, and supported 11 AI/edtech product trials across various school systems. This targeted, inclusive network approach proves we can make AI research faster, more affordable, and more relevant.

3. Clear Market Signals that Support Smart Adoption

Even the best tools need clear signals to support responsible adoption. Educators and school leaders often lack the time and technical expertise to evaluate AI products thoroughly. Public systems need trustworthy indicators of readiness and impact. Testbeds send clear, crucial market signals by transforming abstract technological potential into tangible, evidence-based

results. They function as a bridge between the laboratory and the marketplace, providing objective information that helps investors, customers, and the wider industry make informed decisions. In an environment often filled with hype and speculation, testbeds provide a much-needed dose of reality.

AI testbeds can help by supporting:

- **Evidence Badges and Ratings:** Clearly defined markers of performance, safety, and implementation readiness.
- **Public Indexes and Registries:** Resources like the ISTE EdTech Index to help school leaders navigate options.
- **Outcome-Based Procurement Tools:** Contracting mechanisms that reward tools proven to deliver results in the classroom.

For investors, the primary market signal is whether a technology is a sound investment. Testbeds provide concrete data to answer this question.

- **De-Risking Investment:** A technology that has successfully passed through a rigorous testbed environment is significantly de-risked. By demonstrating that a product works under controlled, realistic conditions, testbeds provide the proof of concept that venture capitalists and other investors need before committing significant capital.
- **Validating Business Models:** Testbeds allow companies to not only test their technology but also to experiment with and validate their business models. This demonstrates a clear path to commercialization and profitability, which is a powerful signal to investors.
- **Objective Performance Data:** Instead of relying on a startup's projections, investors can look to objective, third-party data from a testbed to gauge a technology's performance. This independent validation builds confidence and can lead to more favorable funding terms.

For potential customers, from individual consumers to large enterprises, the key market signal is whether a new technology is reliable, safe, and worth adopting.

- **Demonstrating Real-World Performance:** A successful trial in a testbed within a representative real-world environment, such as a middle school classroom or school, demonstrates to potential customers that a product can effectively handle the complexities of their operations. This is far more persuasive than a simple demonstration.
- **Certifying Interoperability Compliance:** Testbeds are often utilized to ensure that new products conform to industry standards and integrate seamlessly with existing systems. This is a critical signal for customers who need to integrate new technologies without disrupting their current infrastructure.
- **Building Trust Through Safety and Security Validation:** When a product has been rigorously tested for security vulnerabilities and safety protocols in a controlled environment, it sends a strong signal to the market that the manufacturer is committed to producing a trustworthy product. This is especially important in high-stakes education applications like healthcare, finance, and critical infrastructure.

For the broader industry, including competitors and potential partners, testbeds signal where the market is heading and which technologies are gaining traction.

- **Identifying Emerging Technologies:** The technologies being actively tested in prominent national or industry-led testbeds often represent the next wave of innovation. This signals to other companies where they should be focusing their research and development efforts to stay competitive.
- **Accelerating Market Adoption:** When multiple companies use a common testbed, it can accelerate the development and adoption of an entire class of technology. This creates a positive feedback loop where successful tests encourage more companies to enter the market, leading to more innovation and lower prices for consumers.
- **Revealing Gaps in the Market:** The challenges and failures encountered in a testbed can be just as significant as the successes they reveal. These can signal unmet needs or technological gaps, creating opportunities for other innovators to develop new solutions and complementary products.

In essence, testbeds cut through the marketing noise by providing a platform for generating objective evidence. A technology that emerges successfully from a testbed sends an undeniable signal that it is not just a promising idea, but a viable, reliable, investable, and market-ready solution.

These market supports ensure that strong products rise to the top, not just the ones with the biggest marketing budgets. These structures protect public dollars from being wasted or subjected to fraud and abuse.

The New York City Public Schools Model

New York City Public Schools (NYCPS) has pioneered an innovative model for educational technology development by establishing a dynamic testbed network that integrates central strategic initiatives, district-specific contexts, research-based pedagogy, and trusted evaluation frameworks. This approach ensures that new tools are not only practical and secure but also grounded in science-based, high-quality instructional materials and aligned to evidence-based teaching practices that improve student outcomes.

At the heart of this model are NYCPS's strategic initiatives—comprehensive blueprints that articulate the district's curriculum standards, instructional priorities and shifts, and technical specifications. These priorities reflect the district's commitment to rigorous, research-informed instruction and provide a "north star" for developers, public and private, guiding them to design tools that address the authentic needs of all New York City's diverse learner community from the outset.

Safe, promising, and NYCPS-compliant learning tools are introduced into classroom environments led by instructional goals, and resulting in a testbed of participating schools. To guide evaluation, the district draws on established frameworks—such as the EDSAFE AI Alliance framework and others—to assess products for safety, efficacy, equity, interoperability, and instructional alignment. These evaluation processes emphasize alignment with high-quality, standards-based curricula and the real-world application of effective pedagogical strategies.

This model fosters a collaborative ecosystem in which developers receive meaningful, real-time insights; educators act as co-designers of the tools they will ultimately use; and the district is able to scale solutions with greater confidence and reduced risk. By bridging the gap digital skills, strengthening foundational math and literacy, and innovation that transfers to classroom adoption, NYCPS is helping to shape an edtech landscape that is more responsive, equitable, and impactful—supporting deeper learning and improved outcomes for all students.

4. A National Infrastructure: Federated Testbeds, Data, and Ecosystem Supports

Federated Trialing Environments

A national AI testbed network should be built as a "network-of-networks," combining a shared backbone of research protocols with localized testing environments that reflect the true diversity of U.S. schools and family choice in education. This includes:

- Traditional public school districts
- Rural and remote schools
- Charter and microschool networks
- Community-based and alternative learning environments
- Out-of-school time environments

Each site contributes data and insights to the national network, helping refine tools and share what works across contexts. A deliberate outcome of this work would be to both understand what context supports a tool's successful implementation and what outcomes can be expected. This type of evidence-based research network would support state and local infrastructure in making evidence-based decisions about the technology students and educators use.

Essential for robust participation in the network are incentive structures that make engagement attractive, rather than extractive, for testing environments. As detailed in GETN's tenets and principles, this includes direct compensation for participating educators and systems, funding for coordination and infrastructure needs, and policy frameworks that leverage testbed participants' insights to inform decisions about adoption.

Data Infrastructure

One of the most persistent obstacles to timely research is access to secure and usable data from schools and districts; the current system of individualized data use agreements does not scale efficiently. A modern R&D ecosystem will require both traditional and synthetic testbed environments. While real-world classrooms are essential for generating implementation insights and assessing fidelity in live conditions, synthetic data environments offer a complementary opportunity to accelerate early-stage experimentation.

Some state agencies are now exploring synthetic data models, which create anonymized but structurally identical copies of real student data. These systems show promise in enabling secure, ethical, shareable, and scalable research, but limitations remain, particularly when working with private companies that limit access to efficacy data. A national testbed infrastructure should explore new governance frameworks and technical approaches to make research-grade data accessible without compromising student privacy.

Ecosystem Supports

A successful testbed system also needs:

- **Policy Alignment:** This should Include guardrails and guidelines developed in partnership with the EDSAFE AI Policy Lab network.
- **Teacher Capacity Building and Professional Development:** Professional development in AI literacy empowers teachers to critically evaluate new educational technologies, providing the sophisticated, real-world feedback necessary for edtech testbeds to validate and refine innovative tools effectively.
- **Data Infrastructure:** This approach requires tools for secure, ethical data collection and performance analysis.
- **Sustained Investment:** To effectively deliver results, public and philanthropic capital must be targeted at the middle stages of R&D.
- **Aligned Purchasing Power:** A federated network can coordinate its purchasing power to purchase emergent technologies collectively, increasing affordability and access to new solutions, and/or incentivize the development of novel technological solutions that solve real-world problems.

Snapshot of EDSAFE AI Policy Lab Network Members

5. An Investor-Friendly Commercialization Pipeline

A national AI testbed network should be leveraged to derisk high-potential early-stage technologies. Technologies deemed promising through piloting activities within testbeds should be systematically connected with private investors, who can deploy the capital necessary to accelerate the growth and adoption of evidence-based solutions. By developing coordinated "hand-offs" between testbed researchers, schools, and investors, we can simultaneously align R&D activities to pressing real-world needs in classrooms, while accelerating market penetration and strengthening the sector's and nation's economic power.

Investment Hand Offs

- Early-stage investors can leverage testbeds to help seed-stage portfolio companies workshop early products into usable and commercially viable solutions, and gain footholds into new school markets across the U.S.
- Growth-oriented investors can leverage evaluation research generated from testbeds to enhance credibility with customers and increase market share.
- Increased coordination across private investors and research activities can quicken the pace of development and adoption by school communities.

CONCLUSION

A PROACTIVE APPROACH TO AI IN PUBLIC EDUCATION

AI is moving fast. The question is whether American public education will lead or lag. A national AI testbed infrastructure—designed with transparency, common sense, and shared goals—can make the U.S. a global leader in education innovation.

By integrating technical and usage benchmarks, addressing the R&D funding gap, and providing clear market signals, we can ensure that AI tools support real learning for all students and equip schools and districts with the information they need to make informed procurement decisions.

This is not about chasing trends. It's about building systems that work—for rural districts, suburban schools, and urban communities alike. That's the promise of a federated approach that considers local context: one standard, many paths to reach it.

Appendix

- Global EdTech Trialing Network (United States), Boody Adorno, K., & Mote, E. (2023). *GETN Tenets & Principles of EdTech Trialing Networks & Environments within the US*. Global EdTech Testbeds Network. https://doi.org/10.53832/opendeved.1023
- The White House. *America's AI Action Plan*. July 2025, https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.
- U.S. Department of Education. *Dear Colleague Letter: Guidance on the Use of Federal Grant Funds to Improve Education Outcomes Using Artificial Intelligence (AI)*. 22 July 2025, https://www.ed.gov/media/document/opepd-ai-dear-colleague-letter-7222025-110427.pdf.

Acknowledgments

Co-authored by

Katie Boody Adorno & **Erin Mote** & **Chris Agnew**
Leanlab Education | InnovateEDU | SCALE Initiative, Stanford Accelerator for Learning

---

Leanlab Education is a nonprofit organization specializing in codesign research between education technology companies and schools. The organization matches parents, learners, and educators with edtech developers to inform, develop, and evaluate the next generation of classroom tools. Leanlab Education is a partner organization of GETN and co-leads the GETN-US work.

InnovateEDU is a national nonprofit focused on catalyzing education transformation by bridging gaps in data, policy, practice, and research to center the needs of the field in accelerating innovation toward an equitable, inclusive, and radically different future for all learners. The organization co-leads the GETN-US work.

The SCALE initiative is part of the Stanford Accelerator for Learning dedicated to transforming educational opportunity by leveraging knowledge for better education decision-making. SCALE conducts rigorous research, identifies, supports and scales promising solutions and engages decision makers to integrate research, policy, and practice across critical issues in K-12 education.