DHS Playbook for Public Sector Generative Artificial Intelligence Deployment

January 2025

Homeland Security

Contents

Plays

Appendices

Letter from the Secretary and the Chief AI Officer

The rapid advancement and commercialization of Generative Artificial Intelligence (GenAI) presents both significant opportunities and challenges for the Department of Homeland Security (DHS) and public sector organizations at all levels. In alignment with Executive Order 14110, which emphasizes the responsible deployment of AI technologies, DHS has proactively initiated GenAI pilot programs to explore their potential in enhancing our mission capabilities.

These pilot initiatives have provided valuable insights into the practical applications of GenAI within our operations. They have underscored the importance of a measured and thoughtful approach to AI integration, ensuring that our deployments are responsible, trustworthy, and effective while protecting privacy, civil rights, and civil liberties.

The DHS GenAI Playbook encapsulates the lessons learned from our pilot programs and offers a structured framework for the responsible adoption of GenAI technologies. It is intended for and designed to guide sector organizations in implementing GenAI solutions aligned with their unique missions and operational contexts.

The playbook also discusses how organizations should consider building holistic, modern technology delivery capabilities including GenAI. Successful GenAI deployments depend on a strong foundation of AI-enabling skills and technologies, including human-centered design, agile software delivery, and data management. While GenAI is powerful and exciting, it is not the only type of AI. DHS continues to leverage Predictive AI across our operations, and organizations must consider which type of AI is appropriate for each task.

We recognize that the successful integration of GenAI requires collaboration across government, industry, academia, and civil society. We invite you to better understand the insights presented in this Playbook and to join us in fostering a culture of innovation that is both responsible and mission-focused.

Together, we can harness the transformative potential of GenAI to enhance our capabilities and better serve the American people.

Alejandro N. Mayorkas
U.S. Department of Homeland Security
Secretary

Eric Hysen
U.S. Department of Homeland Security
Chief Artificial Intelligence Officer

Introduction: The DHS GenAI Pilots

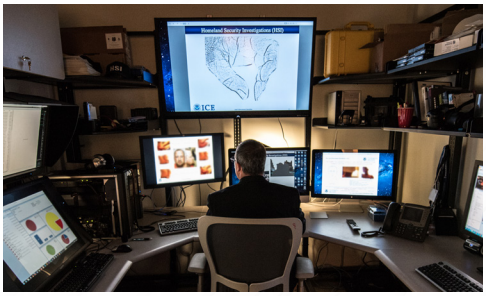
DHS's generative artificial intelligence (GenAI) pilots advance the Department's goal of using AI in service of the homeland security mission. DHS has implemented more traditional forms of AI and machine learning (ML) for over 15 years and was early among government agencies to invest in GenAI. DHS managed this effort due to two primary factors: first, a focus on and enthusiasm for leveraging AI in the Department's mission from senior leaders, including the Secretary, which has been key for uniting the Department and leveraging its resources; and second, strong interest and initiative from the Department's components[1] to find ways to responsibly leverage AI to support their operational work.

On March 18, 2024, DHS released a first-of-its-kind Artificial Intelligence Roadmap to describe the Department's plans to guide its responsible use of AI while ensuring individuals' privacy rights, civil rights, and civil liberties are protected. The roadmap laid out DHS's AI initiatives, described the potential of AI technologies across the Department, and provided visibility into DHS's approach to AI, while underscoring the Department's commitment to AI's responsible and trustworthy use.

As part of this roadmap, DHS publicly announced its goal to complete three GenAI pilots over the course of 2024, described below.[2] In selecting these pilots, DHS prioritized GenAI applications that would help employees do their work, not replace them. DHS also scoped pilots so that they could be broadly applicable in other parts of the Department and the federal government if successful. These pilots targeted mission-enhancing processes that workers could incorporate easily into their workstreams.

Pilot 1: Strengthening Investigative Leads with LLM-Enhanced Search and Summarization

Homeland Security Investigations (HSI) tested AI's ability to strengthen their investigative processes. The pilot had HSI investigators use a large language model (LLM)-based system to enhance the efficiency and accuracy of the summaries they rely upon to identify investigatory leads. The system was designed to leverage open-source technologies to allow investigators to more quickly summarize and search for contextually relevant information within investigative reports. When deployed, the pilot could, among other uses, lead to increased detection of fentanyl-related networks, aid in the identification of perpetrators and victims of child exploitation crimes, and surface key patterns and trends to advance HSI's vital work.



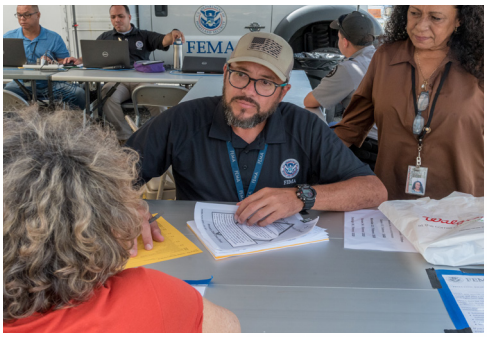
An HSI invesigator assesses evidence.

1 Offices and agencies contained with DHS are referred to as "components" within this document. These include, for example, U.S. Customs and Border Protection, the Federal Emergency Management Agency, United States Citizenship and Immigration Services, and the Cybersecurity and Infrastructure Security Agency.

2 As of time of publication, the pilots have been completed and the respective teams are conducting analysis to determine whether next-stage deployment is appropriate. This playbook will be complemented by additional technical documentation on the pilots.

Pilot 2: Helping Local Governments Create Hazard Mitigation Plans

The Federal Emergency Management Agency (FEMA) tested LLM capabilities to help local governments develop hazard mitigation plans for communities across the country. Hazard mitigation plans are essential to building resilient communities, but they are also lengthy, complex documents that can be difficult for many communities to produce. The pilot was designed to enable state and local governments to efficiently identify and understand their communities' risks and corresponding mitigation strategies. The pilot created draft planning elements from publicly available, well-researched sources that could be customized to meet each community's unique needs. The goal is to enable more communities to submit grant applications for hazard mitigation funding, become more resilient, and reduce disaster risks.



A FEMA employee assists a disaster survivor.

Pilot 3: Creating Novel Training Opportunities for Immigration Officers

United States Citizenship and Immigration Services (USCIS) tested AI's ability to improve refugee and asylum immigration officer training. USCIS used GenAI to provide dynamic, personalized interview training that adapted to officers' specific needs and ensured the best possible knowledge and training on a wide range of scenarios, current policies, and laws. The enhanced training materials gave each officer the opportunity to practice and prepare for interviews with refugees and asylum applicants on their own, as often as needed, enabling continuous learning and reducing their reliance on limited instructor-led training sessions. This dynamic, adaptable training resource aimed to enhance trainees' interview skills, increase comprehension and retention of mission-critical information, improve the accuracy of their decision-making process, and limit the need for retraining over time.

A USCIS employee assists applicants at a contact center.

Across each of these pilots, and other GenAI integrations, DHS carefully considered the risks and benefits of each of these technologies. Using GenAI and other new technology comes with risks: risks posed by adversarial use of GenAI, risks posed to GenAI systems, and risks associated with the implementation of GenAI.[3] These may include operational, security, data, privacy, and inappropriate bias risks. While some of these dangers are context-specific – for example, the risks associated with deploying GenAI for officer training programs look different than for document searches – forward planning can help incorporate risk mitigations by design. The Department proactively incorporated risk mitigations through robust, cross-functional governance processes that included ongoing monitoring and testing.

3 See DHS's Safety and Security Guidelines for Critical Infrastructure Owners and Operators, April 2024, https://www.dhs.gov/sites/default/files/2024-04/24_0426_dhs_ai-ci-safety-security-guidelines-508c.pdf

Start Now, Wherever You Are

DHS designed this playbook to meet organizations wherever they are in their journey to understand and incorporate AI technology in their work. Any public sector organization can start today to assess and gather resources, cultivate internal buy-in, and lay the groundwork for effective deployments of GenAI. Laying this foundation will pay dividends in future developments and create organizational momentum.

Each of the following steps describe actions that organizations can take to advance their own uses of GenAI. These steps address technical, policy, and administrative considerations; therefore, some recommendations may be more relevant to specific experts within an organization, such as technical or policy personnel.

- **Mission-Enhancing GenAI Use Cases:** Public sector organizations must ensure that GenAI deployments align with their mission. Narrowly scoped, mission-enhancing pilots are useful tools for exploring how an organization can use GenAI.
- **Coalition Building and Effective Governance:** Organizations should cultivate support for GenAI applications from top leadership and across functional teams to give GenAI the greatest chance for successful deployment and effective oversight.
- **Tools and Infrastructure:** Organizations should evaluate the technical tools and infrastructure they already possess and consider what technical capabilities they require to deploy GenAI applications.
- **Responsible Use and Trustworthiness Considerations:** From the very beginning, organizations should consider how to make sure GenAI use is responsible and trustworthy and how to address potential risks like privacy, security, bias, and safety.
- **Measurement and Monitoring:** Teams that are developing GenAI applications should measure progress with appropriate metrics and report on that progress to leadership and other stakeholders.
- **Training and Talent Acquisition:** Organizations should train their staff on responsible and effective GenAI use and hire skilled employees who can support GenAI development.
- **Usability Testing and Other Feedback Mechanisms:** Organizations should incorporate iterative feedback from users and other stakeholders to develop and improve GenAI applications.

Mission-Enhancing GenAI Use Cases

Develop GenAI Pilots that address mission need and target supporting mission-enhancing processes.

Organizations should carefully determine the scope of any potential GenAI deployment and consider designing a GenAI pilot before deploying a functionality broadly. Well-designed GenAI pilots provide a low-stakes environment to take risks and learn about resource and governance needs. Teams should design pilots that solve a **specific problem** in a way that **serves the organization's mission**. Importantly, pilots should first focus on **supporting mission-enhancing processes** before adding GenAI to mission-critical operations, where failure could have more serious consequences. Organizations should also consider the long-term potential of a pilot product, including whether other departments or processes could make use of the tool. Finally, organizations should **enlist an executive sponsor** early to support and endorse pilot effort(s).

Pilot plans should **assess resource needs**—including funding, staff, data, and technology—and **criteria for evaluating the pilot**, whether objective (e.g., performance, accuracy, and adoption) or subjective (e.g., user experience and sentiment). Data may be a particular challenge, and teams should identify data needs, datasets, potential privacy issues, and responsible data use before launching a pilot.

ACTIONABLE STEPS

- ☐ Align any GenAI deployment's potential mission and value with the organization's priorities.
- ☐ Scope a GenAI pilot that improves a specific mission-enhancing process and is potentially scalable to similar processes or useful in other parts of the organization.
- ☐ Enlist an executive sponsor whose organization will benefit from the GenAI pilot.
- ☐ Assess resources, including staff, funding, data, and technology.
- ☐ Define the pilot's minimum viable product (i.e., when it will be ready for evaluation) and metrics for evaluating success.

THE DHS STORY

In April 2023, the Secretary of Homeland Security stood up an AI Task Force (AITF) and charged it with identifying ways AI could enhance the DHS mission while protecting privacy, civil rights, and civil liberties across the Department. The AITF solicited and received nearly three dozen proposals for pilot projects from across the Department to be considered for short-term funding from DHS Headquarters. The AITF chose pilots to fund based on the projects' alignment with the DHS and component mission, technical and funding feasibility, potential cross-organizational applications, access to data and infrastructure, and reasonable project duration. The AITF also strategically chose pilots that represented different use cases and types of technology applications, so that DHS could collectively learn as much as possible about GenAI applications. A sample pilot proposal template can be found in Appendix C.

Scope was an important consideration for the AITF in selecting pilots. To facilitate measurement and tracking, the AITF funded pilot projects that focused on discrete, tangible outcomes, such as reducing hours officers spent working on specific cases. Also, because DHS policy prohibits using GenAI as the sole basis for decision-making, pilots instead targeted process improvements, such as enhancing training or assisting research. Each pilot team identified its criteria for a minimum viable product and benchmarks that would indicate when pilot results would be ready for evaluation.

To finance the pilots, DHS drew from a fund for IT modernization that was readily available for one-time use but could not be used on a recurring basis. These funds gave DHS the flexibility to develop and experiment with GenAI without impacting the Department's budget.

Executives at each component running a GenAI pilot provided integral support for the projects. While DHS did not require pilot teams to identify component executives as sponsors at the onset, DHS recommends doing so in the future. Component (or sub-agency head) executives can help marshal resources more directly and respond to issues as they come up.

Coalition Building and Effective Governance

> **Seek sponsorship from the most senior executive, then build cross-organizational coalitions to oversee GenAI deployments.**

ACTIONABLE STEPS

- ☐ Prioritize early buy-in from the most senior leadership possible and gain input from key stakeholders across all parts of the organization.
- ☐ Evaluate current governance structures and policies to identify potential gaps.
- ☐ Designate an existing governance body that includes cross-functional representation or stand one up for management and oversight. Include technical experts, cybersecurity professionals, privacy, civil rights and civil liberties professionals, and dedicated responsible use and trustworthiness experts in teams developing and developing the GenAI tool.

Organizations should solicit buy-in for GenAI projects from key internal stakeholders, from the **most senior leadership possible** (up to and including the head of the organization). It is especially important to include risk management, compliance, and oversight leads (such as civil rights and civil liberties experts), gain their support, and incorporate their insights early in the planning process. Identifying the right stakeholders may take some work in larger bureaucracies, but having a **cross-organizational coalition** working on GenAI can lead to better decision-making, faster development, and stronger responsible use practices.

Organizations should **assess their current governance structures** and consider how they could be leveraged, updated, or added upon to address AI governance. Sometimes an organization can simply add responsibilities to an existing body; at other times, it will need to create a new body to oversee AI and any GenAI pilots. These structures may change over time as policy and the technology evolves, but AI governance structures should similarly represent a **wide range of stakeholders and teams** across the organization, such as privacy professionals, civil rights and civil liberties experts, legal counsel, and cybersecurity and data experts. This approach serves to reinforce coalitions and ensure these bodies are well-positioned to identify the full range of potential risks, support impact assessments, design mitigation strategies, and implement remediations in real time.

THE DHS STORY

DHS benefitted from the Secretary of Homeland Security's direct sponsorship of GenAI application adoption, which unlocked resources and buy-in for GenAI across the Department. In addition, many DHS components already had long-standing experience incorporating AI into their missions. This combination of top-level leadership and component-level experience and support helped DHS build coalitions across the Department to support AI governance, development, and deployment.

Secretary Mayorkas participates in a fireside chat about AI.

DHS reinforced this approach on the ground with integrated project teams (IPTs) that included cybersecurity experts and legal, privacy, and civil rights and civil liberties professionals. The IPTs oversaw the pilots' development and deployment, helping the GenAI projects establish legitimacy and earn buy-in across the Department.

Cross-functional collaboration was key for DHS's AI governance efforts, both for the GenAI pilots and other AI deployments. When DHS began its GenAI pilots, the AI Task Force—a new, limited-term entity—was initially the Department's only governing body focused on AI. The AITF was led by senior representatives from the Office of the Chief

Information Officer/Chief AI Officer, DHS Science and Technology Directorate, and the DHS Office for Civil Rights and Civil Liberties. Membership included representatives from key organizations throughout the Department, including the DHS Office of Strategy, Policy, and Plans and the DHS Privacy Office. The multidisciplinary IPTs also provided critical on-the-ground governance for AI risks, as team members could address issues like cybersecurity concerns and privacy considerations as they arose. The pilot teams could then elevate problems and lessons learned from DHS component-level AI working groups to the AI Task Force, informing DHS's broader AI governance efforts.

In parallel with its GenAI pilots, DHS took other steps to develop and mature its policies and practices to guide responsible AI use. Critically, the Chief AI Officer issued a Department-wide policy that provides initial guidance for responsible GenAI use within DHS. The policy detailed what data employees can share with a GenAI tool, required employees who wanted to use GenAI to undergo a training and secure the approval of a supervisor, required human review of GenAI outputs before those outputs could be relied upon, and set forth appropriate use cases. The Chief Privacy Officer published a Privacy Impact Assessment on the Department's use of commercial GenAI tools.[4] The AITF also set up communities of practice, such as a Responsible Use Group, which collaborated on responsible AI best practices and priorities. DHS will continue to assess and update policies to guide responsible GenAI use at DHS as the Department and the technology evolves.

4 See DHS/ALL/PIA-097 Use of Conditionally Approved Commercial Generative Artificial Intelligence Tools, November 2023, https://www.dhs.gov/publication/dhsallpia-097-use-conditionally-approved-commercial-generative-artificial-intelligence

Tools and Infrastructure

Build organizational capacity to develop and deploy GenAI by using existing tools and infrastructure for pilots.

Organizations should **assess existing tools and infrastructure** available to them for GenAI and plan around their strengths and limitations. Organizations can adapt their existing processes for security compliance, scalability, and integration with existing IT infrastructure and data management practices.

Organizations should consult with their technical experts and consider whether **commercial[5], open-source[6], or open-weight[7] AI models** are more appropriate for their needs, especially if they face resource constraints or handle sensitive data. Open-source models offer organizations more control and visibility into the model's functioning and security, often with a lower total cost of ownership, while commercial and open-weight models may have fewer long-term maintenance costs. An organization's cloud environment could affect its model choice, as some large models are less expensive to use through existing cloud contract APIs, while open-source models require a service provider or organizational infrastructure. The purpose of an organization's GenAI pilot can also influence tooling and infrastructure decisions. This is especially true for law enforcement applications, which may require offline models that can run locally. Open-weight AI models can support such offline uses.

Exploring novel and/or additional AI tools should come *after* building **strong software design, development, and cybersecurity fundamentals**. Without sufficient baseline expertise or tooling, a GenAI pilot's resource needs could quickly overwhelm an agency's budget and infrastructure.

ACTIONABLE STEPS

- ☐ Assess existing tools, infrastructure, and technical capabilities against the needs and goals of the GenAI pilot or deployment, including securing sensitive data.
- ☐ Consider the use of commercial, open-source, or open-weight models, and whether using more than one type of model would be appropriate.
- ☐ Determine if any additional tooling or configurations are needed, keeping in mind your team's technical capacity.

THE DHS STORY

DHS's GenAI pilot teams began their work by assessing their available tools, infrastructure, and technical skills, and then planned accordingly. The teams found that using available tools and licenses simplified pilot development.

The GenAI pilot teams had access to commercial GenAI models through approved cloud services that were already available to federal employees. They also explored various open-source and open-weight models. Pilot teams found that using models linked to pre-approved cloud services streamlined compliance with security needs, scalability, and integration processes that otherwise may have proven to be significant obstacles.

The pilot teams also identified and used tools that matched the teams' skillsets. For teams with limited resources, low-code platforms such as commercial AI studios proved useful.

5 'Commercial GenAI tools' are defined as generative AI products or services available for license, subscription, or purchase by the general public or through commercial agreements. This definition does not include software or services developed or customized specifically for the government through IT acquisition process.

6 An open-source AI model is an artificial intelligence model whose source code, model weights, and training data are publically available for anyone to use, modify, study, and share.

7 An open-weight AI model is an artificial intelligence model whose model weights and information about the training data are publicly available and that has a permissive use of the capability. In contrast to open-source models, open-weight models may require licenses for use and generally do not provide ful access to the model's training data or source code.

Responsible Use and Trustworthiness Considerations

Develop responsible use and trustworthiness principles and include oversight partners in GenAI pilots to minimize potential harm.

Organizations need to prioritize responsible and trustworthy AI use early when deploying GenAI, and it is essential to establish a common organizational understanding of "responsible"[8] and "trustworthy.[9]" Leaders should clearly communicate the limitations and risks of GenAI and charge teams to **identify potential risks**, including inaccuracies, discrimination, privacy impact, and data bias. Consider, for example, that data shared with commercial GenAI tools may be used in their training data, posing broader information security and privacy risks to public organizations.

To manage these risks, organizations should **evaluate existing policies and regulatory frameworks for their application to Gen AI use** and, as needed, **develop clear organizational guidance and principles** for responsible and trustworthy use. Organizations should also **scope the application of GenAI tools appropriately** given their limitations. GenAI outputs should not be the sole basis for any critical decisions, and organizations should make clear where GenAI tools and their outputs need human review before use. All GenAI tools should undergo **testing**, with the expectation that it may be necessary to roll back development efforts if a tool does not meet specified testing criteria.[10]

ACTIONABLE STEPS

- ☐ Identify areas of potential risk, including confabulations/hallucinations, privacy violations, discrimination, data bias, threats to civil rights and civil liberties, physical safety, and data security.
- ☐ Scope the application of GenAI tools appropriately, accounting for their limitations and risks.
- ☐ Develop clear organizational guidance, principles, and best practices for responsible and trustworthy GenAI use.
- ☐ Develop approaches for risk management, such as regular testing.
- ☐ Ensure that lessons learned from risk identification, mitigation, and remediation are regularly used to improve policies and keep pace with technology developments.

GenAI pilot teams should collaborate with legal, privacy, civil rights and civil liberties, and cybersecurity experts to address responsible use considerations throughout development. Pilot teams should **use the lessons learned from their pilots to improve policies and keep pace with technology developments**, which helps address emerging risks and fosters public trust.

THE DHS STORY

Traditionally, DHS oversight offices did not play an active role in day-to-day IT management, instead reviewing compliance at the end of deployment or as secondary stakeholders. The Department changed that model with GenAI due to the evolving privacy, civil rights, and civil liberties considerations, embedding its oversight offices throughout the deployment process. DHS placed privacy, civil rights, and civil liberties experts on each pilot's IPT to ensure the GenAI pilots aligned with and adhered to the Department's principles for responsible AI use. These principles include a requirement that DHS will only acquire and use AI in a manner that is consistent with the Constitution and all other applicable laws and policies, including those addressing privacy, legal considerations, civil rights, and civil liberties.

8 "Responsible AI" is an AI system that aligns development and behavior to goals and values. This includes developing and fielding AI technology in a manner that is consistent with democratic values. (Source: National Security Commission on Artificial Intelligence: The Final Report)

9 Characteristics of trustworthy AI systems include: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. (Source: NIST AI Risk Management Framework)

10 Federal agencies can refer to OMB guidance M-24-10 for minimum risk management standards for safety/rights-impacting AI.

Working closely with the IPTs, the Department's privacy, civil rights and civil liberties experts identified and documented lessons learned, provided project-level responsible use governance, and provided input on broader responsible use considerations.

While the IPTs worked to ensure AI was used responsibly in their respective pilots, DHS also required employees to understand GenAI risks and mitigations before using GenAI in their work. DHS employees who wish to use GenAI are required to complete training on the Department's responsible GenAI use policies, principles, and best practices before they are given access to GenAI tools.

These responsible use policies and associated implementations are part of an ongoing effort to create a culture of awareness about GenAI risks across the Department that ensures all AI use at DHS—including the use of GenAI—aligns with applicable statutes, regulations, guidance, and rights-respecting best practices. The Department's best practices include a commitment to transparency. For example, DHS discloses non-classified and non-sensitive uses of AI, including the Gen AI pilots, through its AI Use Case Inventory, to the extent practicable and in accordance with applicable law and policy.

DHS leadership convenes the first AI Safety and Security Board meeting in May 2024.

In support of its broader mission, DHS also developed guidance to advance responsible and trustworthy AI use in critical infrastructure. For example, DHS released Safety and Security Guidelines for Critical Infrastructure Owners and Operators to aid organizations in the evaluation and mitigation of risks to and from GenAI system integrations in mission settings. DHS also consulted closely with its AI Safety and Security Board to write a Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure to outline responsible and trustworthy best practices for the safe and secure development and deployment of AI in critical infrastructure. The process of consulting experts to develop these guidance frameworks helps DHS learn from and contribute to the ongoing global efforts to use AI safely and responsibly.

Measurement and Monitoring

Determine appropriate key performance indicators for the pilots and monitor them throughout the pilot lifecycle.

Organizations should **identify or develop qualitative standards and quantitative metrics** that reflect the goals of their GenAI pilots. These standards and metrics should allow development teams, leadership, and other stakeholders to regularly assess the effectiveness of GenAI pilots. The most effective metrics should cover mission impact, value, and oversight objectives such as demonstrating responsible use and trustworthiness. Organizations should also ensure the **necessary infrastructure** is in place (e.g., a dashboard) to monitor these metrics, capture relevant data, and communicate their pilots' progress to internal stakeholders.

ACTIONABLE STEPS

- ☐ Finalize and validate metrics for assessing the GenAI pilot's effectiveness.
- ☐ Create infrastructure to measure and monitor results over the entire pilot.
- ☐ Determine the appropriate format and cadence for sharing metrics with stakeholders.
- ☐ Develop a process to make iterative improvements to the pilot product based on performance.

Pilot teams should work with organizational leadership to determine the **appropriate format, cadence, and audience** for sharing metrics. At minimum, metrics should be shared monthly, depending on the length of time allotted for a given pilot. These updates should inform a **process of making iterative improvements** or updates based on a pilot's performance. Organizations should closely track the release of new metrics and benchmarks developed by standards organizations or local authorities that may apply or support the monitoring of their use case, such as the National Institute of Standards and Technology's AI Risk Management Framework or the NIST-AI-600-1 Generative AI Risk Profile.[11]

THE DHS STORY

DHS GenAI pilot teams set performance indicators for mid-term, long-term, and ongoing milestones that they shared with senior leadership in monthly progress reports. In these reports, teams shared a Pilot Dashboard that highlighted metrics, obstacles, and next steps. See Appendix D for a sample Pilot Dashboard.

Each GenAI pilot team based its metrics on the function and mission impact of its pilot:

- **USCIS** measured success through reduced training interview times and improved officer training exam scores, with a goal of enhancing understanding, retention and decision accuracy, while minimizing the need for retraining.
- **HSI** scored its pilot's ability to increase data recovery speed and create accurate summaries, aiming to reduce research time and improve decision-making.
- **FEMA** evaluated the ability of its pilot's AI-generated content to meet the agency's requirements for hazard mitigation plans and support the needs of participating communities. Their goal was to reduce the time needed to develop hazard mitigation plans and increase the number of communities with current plans to improve overall resiliency and grant eligibility.

GenAI pilot teams also developed technical metrics to evaluate the accuracy and usefulness of generated results. HSI and FEMA used human and automated testing frameworks to evaluate retrieval augmented generation (RAG) results for accuracy, relevancy, reliability, and other measures. The USCIS pilot team developed other measures to assess technical performance and to compare different models. These metrics helped the pilot teams and stakeholders track progress, assess mission alignment, and identify potential obstacles.

11 https://doi.org/10.6028/NIST.AI.600-1

Training and Talent Acquisition

Offer targeted training in support of GenAI pilots and invest in hiring technical talent in the long-term.

Employees who use any GenAI application should understand its capabilities, limitations, risks, and opportunities. Managers who may oversee GenAI use cases should have a baseline familiarity with the technology to ensure they take an informed approach to their respective roles and responsibilities. Organizations should, therefore, **generally train their staff** on the use of GenAI tools and help them develop fundamental technical skills.

When initiating a GenAI pilot, organizations should first **identify necessary technical skills** and determine if they already exist on their teams. Where possible, organizations should **adapt training** to provide current employees targeted opportunities for up-skilling or cross-training. In parallel, organizations should consider how to **hire skilled employees** to support GenAI development. **External organizations**—such as temporary public sector technology fellows—may be a helpful resource.

ACTIONABLE STEPS

- ☐ Offer GenAI literacy training across technical and non-technical business lines to create a common understanding and lexicon across the organization.
- ☐ Identify needed technical skills for the GenAI pilot and assess if these skills exist in the organization.
- ☐ Offer upskilling and cross-training opportunities for current employees.
- ☐ Partner with external organizations to leverage their specialized expertise.
- ☐ Hire technically skilled employees to support AI development efforts.

THE DHS STORY

For the last decade, DHS has invested in growing more technology talent across the organization. With the advent of GenAI, the Department's need for such skills only accelerated. DHS offered its technical staff opportunities to learn about developing and using GenAI and other machine learning technologies. The Department also provided several trainings across components for senior, non-technical staff to understand the basics of GenAI and other types of AI. Through these efforts, DHS raised the overall level of GenAI literacy in the Department and drove more buy-in for managers who could now understand and use the tools themselves. These training efforts also resulted in an explorative development process where non-technical staff were able to help test and evaluate solutions alongside technical staff.

The GenAI pilot teams drew from both internal and external skill sets for their projects. Each team assessed its internal technical skills and needs as part of its pilot proposal and scoped its pilot's technical needs accordingly. The GenAI pilot teams also hosted technical detailees from the U.S. Digital Service who supported the pilots and provided an outside perspective.

Later in pilot development, DHS also expanded its in-house technical expertise through the AI Corps hiring sprint. DHS worked with the Office of Personnel Management (OPM) to design a targeted, streamlined, and expedited hiring process to rapidly onboard over 40 experienced technical experts from across sectors. DHS focused on hiring personnel with product management experience who could help drive the full lifecycle of AI projects across the organization. These centrally hired AI Corps experts joined projects across the Department to support AI development, including the GenAI pilots.

Usability Testing and Other Feedback Mechanisms

Throughout the GenAI pilot lifecycle, seek and incorporate user feedback and share updates with stakeholders.

At the outset of any GenAI pilot, the organization running the pilot should identify relevant users for the pilot product—both inside and outside of the organization—and **engage them in regular usability testing** throughout the development process. Organizations can also build trust and gain valuable feedback by communicating proactively with other stakeholders (non-users), which may include executive leadership, as well as oversight, public engagement, and policy professionals. Organizations should share pilot updates with these stakeholders so they remain engaged and are able to provide **informed feedback** throughout the process.

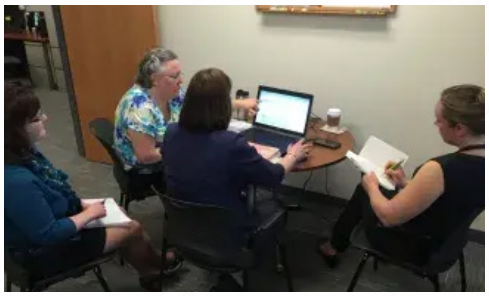
This two-way communication helps pilot teams improve their products, while also bolstering stakeholders' confidence in the projects. Internal stakeholders can incorporate lessons learned elsewhere in the organization, while external stakeholders—including users—can provide feedback about the tool's performance, usefulness, and governance.

ACTIONABLE STEPS

- ☐ Identify relevant internal and external users and iteratively test the product with them throughout the development lifecycle.
- ☐ Communicate regularly with users and other stakeholders throughout the pilot planning and execution to share progress, challenges, and lessons learned.
  - ▶ For external stakeholders, determine if an additional document (e.g., a memorandum of agreement, a data license, or a disclaimer) is required before external stakeholders can engage the product.
- ☐ Provide opportunities for stakeholders to share feedback and incorporate this feedback as appropriate.

THE DHS STORY

User feedback, collected through usability testing and product demos, proved critical to the success of DHS's GenAI pilots. FEMA recruited over a dozen state and local disaster planners to test a beta version of its hazard mitigation planning tool, gaining insights on usability and feature improvements. HSI and USCIS worked with internal users to assess and refine their GenAI pilots, documenting the results of each interaction.



A team conducts a usability test at a USCIS service center.

Beyond usability testing, DHS found immense value in two-way communication with both internal and external stakeholders during the GenAI pilots. Internally, pilot teams came together to provide monthly updates to senior AI leaders, receiving feedback and guidance in return. These joint check-ins with leadership also allowed pilot teams to learn from each other's experiences, thereby fostering a supportive ecosystem where teams could share information and helpful resources such as testing frameworks.

Externally, DHS engaged oversight partners and the public to build trust and demonstrate transparency. By publicly announcing the GenAI pilots, DHS held itself accountable for responsible implementation. The GenAI pilot teams also briefed colleagues in other departments and Congress on their projects.

Appendix A: Resources

Artificial Intelligence at the Department of Homeland Security, https://www.dhs.gov/ai

Artificial Intelligence Use Case Inventory, https://www.dhs.gov/data/AI_inventory

Cybersecurity and Infrastructure Security Agency (CISA) Roadmap for Artificial Intelligence, November 2023, https://www.cisa.gov/resources-tools/resources/roadmap-ai

Department of Homeland Security (DHS) Artificial Intelligence Roadmap, March 2024, https://www.dhs.gov/publication/ai-roadmap

Department of Homeland Security Commercial Generative Artificial Intelligence Acceptable Use Training, https://www.dhs.gov/sites/default/files/2024-02/24_0226_cio_training-dhs-personnel-genai-use.pdf

Department of Homeland Security Compliance Plan for OMB Memoranda M-24-10, September 2024, https://www.dhs.gov/sites/default/files/2024-09/2024_0923_cio_dhs_compliance_plan_omb_memoranda.pdf

Department of Homeland Security Report on Reducing the Risks at the Intersection of Artificial Intelligence and Chemical, Biological, Radiological, and Nuclear Threats, April 2024, https://www.dhs.gov/publication/fact-sheet-and-report-dhs-advances-efforts-reduce-risks-intersection-artificial

Directive 026-11 Use of Face Recognition and Face Capture Technologies, September 2024, https://www.dhs.gov/sites/default/files/2023-09/23_0913_mgmt_026-11-use-face-recognition-face-capture-technologies.pdf

Policy Statement 139-06 Acquisition and Use of Artificial Intelligence and Machine Learning Technologies by DHS Components, September 2023, https://www.dhs.gov/sites/default/files/2023-09/23_0913_mgmt_139-06-acquistion-use-ai-technologies-dhs-components.pdf

Policy Statement 139-07 Use of Commercial Generative Artificial Intelligence Tools, October 2023, https://www.dhs.gov/sites/default/files/2023-11/23_1114_cio_use_generative_ai_tools.pdf

Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure, November 2024, https://www.dhs.gov/sites/default/files/2024-11/24_1114_dhs_ai-roles-and-responsibilities-framework-508.pdf

Safety and Security Guidelines for Critical Infrastructure Owners and Operators, May 2024, https://www.dhs.gov/publication/safety-and-security-guidelines-critical-infrastructure-owners-and-operators

Appendix B: Glossary

Accuracy (in the context of AI). Accuracy in this playbook refers to the extent to which the output of the generative AI model correctly reflects factual information or adheres to the intended task. It measures how often the model's responses are correct and free from errors or falsehoods.

Artificial Intelligence (AI). 'Artificial intelligence'[12] includes the following: any artificial system that performs tasks under varying and unpredictable circumstances without significant human oversight, or that can learn from experience and improve performance when exposed to data sets; an artificial system developed in computer software, physical hardware, or other context that solves tasks requiring human-like perception, cognition, planning, learning, communication, or physical action; an artificial system designed to think or act like a human, including cognitive architectures and neural networks; a set of techniques, including machine learning, that is designed to approximate a cognitive task; or an artificial system designed to act rationally, including an intelligent software agent or embodied robot that achieves goals using perception, planning, reasoning, learning, communicating, decision- making, and acting. (Source: OMB M-24-10)

AI model. The term "AI model" means a component of an information system that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs. (Source: Executive Order 14110)

AI systems. The term "AI systems" means any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI. (Source: Executive Order 14110)

Commercial AI/GenAI. 'Commercial Gen AI tools' are defined as generative AI products or services available for license, subscription, or purchase by the general public or through commercial agreements. This definition does not include software or services developed or customized specifically for the government through an IT acquisition process. (Source: Adapted from Policy Statement 139-07 Use of Commercial Generative Artificial Intelligence (AI) Tools).

Confabulation or hallucination. The production of confidently stated but erroneous or false results or outcomes by which users may be misled or deceived. (Source: adapted from Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST, 2024).

Generative AI. The term "generative AI" (or, GenAI) means the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content. (Source: Executive Order 14110)

Infrastructure (in the context of AI). In the context of developing an AI tool, infrastructure refers to the foundational systems and services required to support the development, deployment, and maintenance of the AI application. This includes hardware, cloud services, data storage, networking, development tools, and security measures that ensure the AI tool operates efficiently and reliably.

Integrated project team (IPT). In the context of DHS' GenAI pilots, integrated project teams were the teams that ran the GenAI pilots that incorporated staff from different functional areas across the Department, including cybersecurity experts, privacy experts, and civil rights and civil liberties professionals.

Large language model (LLM). Large language model: a class of language models that use deep-learning algorithms and are trained on extremely large textual datasets that can be multiple terabytes in size. LLMs can be classed into two types: genera-

12 The term "artificial intelligence" has the meaning provided in Section 238(g) of the John S. McCain National Defense Authorization Act for Fiscal Year 2019.

tive or discriminatory. This playbook refers to generative LLMs are models that output text, such as the answer to a question or even writing an essay on a specific topic. They are typically unsupervised or semi-supervised learning models that predict what the response is for a given task. (Source: adapted from AI Assurance: Towards Trustworthy, Explainable, Safe, and Ethical AI. Netherlands, Academic Press, 2022.)

Machine learning. The term "machine learning" means a set of techniques that can be used to train AI algorithms to improve performance at a task based on data. (Source: Executive Order 14110)

Minimum viable product (MVP). A minimum viable product is the stage in development when a product has enough features to be evaluated as a functioning product.

Open-source AI model. An open-source AI model is an artificial intelligence model whose source code, model weights, and training data are publicly available for anyone to use, modify, study, and share.

Open-weight AI model. An open-weight AI model is an artificial intelligence model whose model weights and information about the training data are publicly available and that has a permissive use of the capability. In contrast to open-source models, open-weight models may require licenses for use and generally do not provide full access to the model's training data or source code.

Responsible AI. An AI system that aligns development and behavior to goals and values. This includes developing and fielding AI technology in a manner that is consistent with democratic values. (Source: National Security Commission on Artificial Intelligence: The Final Report)

Retrieval augmented generation (RAG). Retrieval Augmented Generation (RAG) refers to retrieving data from outside an AI model and augmenting prompts by adding relevant retrieved data in context. RAG allows fine-tuning and modification of the internal knowledge of the model in an efficient manner without needing to retrain the entire model. In RAG, a retrieval system first fetches relevant documents or pieces of information from a large dataset based on a given query. Then, a generative model uses this retrieved information to produce a more accurate and contextually relevant response. This approach enhances the quality and reliability of the generated content by grounding it in real, retrieved data. (Source: Adapted from Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, NIST, 2024)

Trustworthy AI. Characteristics of trustworthy AI systems include: valid and reliable, safe, secure, and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. (Source: NIST AI Risk Management Framework)

Appendix C: Pilot Proposal Submission Template

Sample Request for AI Pilot Proposals

Pilot Title: Limited to 10 Words

1. **Objective:** In a few sentences, describe the proposed end-product and mission outcome of your AI pilot proposal.

Limit the Objective Statement to 50 words or 300 characters

2. **Description:** Provide a description of the proposed pilot, including what other [teams/departments/offices] or missions within [YOUR ORGANIZATION] that would benefit from the development of the solution.

Limit the Description to 500 words or 3,000 characters

3. **AI/ML Tool Type:** Provide a description of the proposed AI/ML tool.

Limit the Description to 250 words or 1,500 characters

4. **Estimated cost for the pilot in the first year:** Provide an estimate for the development, deployment, and operational costs of the pilot in its first year.

Cost Estimate

5. **Executive Champion:** Name the senior leader responsible for driving the pilot.[13]

Name	Team/Department/Office	Telephone	Email

6. **Technical Point of Contact:**

Name	Team/Department/Office	Telephone	Email

7. **Relevance to the [YOUR ORGANIZATION] mission:** Identify which [YOUR ORGANIZATION] priorities this pilot would support.

[YOUR ORGANIZATION] priority alignment

8. **Contracts/Acquisitions:** Will this pilot require new contracts or acquisitions to support, or will it rely on existing resources?[14]

Limit the Description to 250 words or 1,500 characters

Additional Details

[Consider providing additional details such as timeline for selecting pilots, points of contact, etc.]

13 Not originally included in DHS Pilot Template, but DHS recommends including moving forward.
14 Not originally included in DHS Pilot Template, but DHS recommends including moving forward.

Appendix D: Sample Metrics Dashboard

DHS | OCIO
Protect. Connect. Perform.

PILOT STATUS

Pilot Scope: [One sentence description]

Pilot Objectives
Objective 1: [one sentence objective]
Objective 2: [one sentence objective]
Objective 3: [one sentence objective]

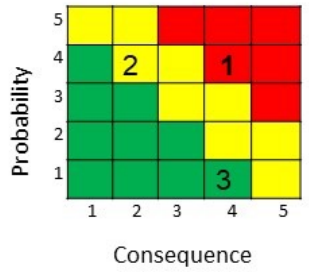
Milestones	Due by	Condition
Phase 1	Month Year	Completed
Phase 2	Month Year	On schedule
Phase 3	Month Year	Started
Phase 4 (MVP)	Month Year	Not Complete

Status Updates

- **Technical**
  - Updates
- **Infrastructure**
  - Updates
- **UX/Design**
  - Updates

Risks

- 1 – Risk #1
  - Description of risk
    - Proposed Mitigation:
- 2 – Risk #2
  - Description of risk
    - Proposed Mitigation:
- 3 – Risk #3
  - Description of risk
    - Proposed Mitigation:



1