CMS Artificial Intelligence Playbook

Version 3

CMS
CENTERS FOR MEDICARE & MEDICAID SERVICES

This document may mention specific businesses, devices, or materials to clearly explain a concept or use case. This does not indicate CMS endorsement nor is it intended to designate them as the only or best choice for your needs.

The Artificial Intelligence Playbook (AI Playbook) by CMS discusses a dynamically evolving subject. This document is planned for continuous review and updates. Feedback is welcome via email to ai@cms.hhs.gov. Additional opportunities to contribute to this document will be posted in the CMS slack #ai-community.

You may access the latest version of this publication without charge at: https://ai.cms.gov/

Acknowledgments

CMS would like to thank Keith Otis (DSAC), Remy DeCausemaker (DSAC), Leo Meister (OEDA), Benjamin Simcock (OAGM), Andres Colon (OIT), Rick Lee (OIT), Matt Artz (OIT), and Marina Levy (OIT) who provided invaluable feedback during the development of the AI Playbook v3. Their insights and input have been instrumental in shaping the final version of the Playbook, and we are deeply grateful for their support.

Version Information

Updates are organized through a versioning system where major changes will increment the first number (e.g., from 1.0 to 2.0), while minor alterations will be indicated by adding a decimal (e.g., from 1.0 to 1.1). All updates are recorded in the Version Control Table below, showing the version, change date, and change details.

Table 1. Version Control Table

Version	Change Date	Editor Control	Details
1.0	2021-09-15	X. Wu	Initial Publication via AI Pilot
2.0	2022-10-18	X. Wu, T. Ahmad	Reorganization, Use Case, Expansion into RAI
3.0	2024-05-15	X. Wu, C. Rutherford	Complete rewrite focused on CMS application of AI/ML and Human-Centered AI Development

Table of Contents

Expanded Table of Contents

List of Figures

List of Tables

1. Introduction

As the steward of health coverage for over 160 million Americans, the Centers for Medicare & Medicaid Services (CMS) is embracing artificial intelligence (AI) to improve health care administration and delivery. This AI Playbook outlines practical frameworks and actionable insights to harness AI effectively within CMS's operations, aiming to enhance service delivery, optimize efficiency, and uphold the highest standards of care and ethical responsibility.

1.1. Purpose and Use

Following the National Artificial Intelligence Initiative Act (NAIIA) of 2020 to accelerate AI research and application across the federal government, the Office of Information Technology (OIT) in CMS launched the AI Explorers (AIE) Program. AIE strives to broaden the understanding of AI and its uses in CMS. Built upon experience from across CMS and contributing sources across the United States federal government, this Playbook is a culminating effort to fit the world of AI into a smaller, CMS-specific "box" giving readers a one-stop shop to dive responsibly into AI development.

1.2. Contents

This document's primary content begins with a general overview in Overview of AI Technologies Section 2 as a baseline introduction into AI. Narrowing the scope, Section 3: Foundations for AI at CMS discusses the state of AI in CMS and shares recommended principles to guide your AI development in the agency. Section 4: Implementation and Operation of AI at CMS bridges these guiding principles with an example approach to AI projects and models within CMS. Finally, the appendices contain several helpful use cases and reference materials to support the reader both in further understanding this document and expanding your knowledge elsewhere.

1.3. Audience and Objectives

Those shaping policies and compliance measures will find this Playbook valuable for aligning AI initiatives with current and future guidance in CMS and elsewhere. While the topics of privacy and security are not the primary focus of this document, it does include principles for creating and deploying AI tools that respect privacy, security, and ethical norms. It guides readers through understanding the foundations of AI and provides resources to develop trustworthy AI-based solutions. Program managers and administrators overseeing projects will find insights on incorporating AI to enhance impact without compromising on operational costs and trustworthiness of their products. This playbook also offers example use cases for their awareness with the appropriate contact details for connection. Practitioners tasked with embedding AI into systems will find resources that guide the implementation of appropriate tools and methodologies. They will also find best practices for system lifecycle management. Training and development personnel tasked with AI education across CMS will find useful resources to build effective training programs. This playbook may serve as a foundation for ensuring the workforce is equipped to use AI responsibly and effectively. Table 2 provides key inquiries for each audience type and the corresponding section in the Playbook.

Table 2. Audience Type and Relevant Sections

Audience Type	Key Inquiry Areas	Corresponding Sections
Program Managers and Administrators	• What are guidelines for ethical AI use that align with CMS's ethical norms? • How can we ensure that an AI-based solution is the right solution for our challenge? • Who and what should be considered when governing AI at the organizational level?	• Section 3.2. Example Guiding Principles for AI at CMS provide a foundation in guiding the implementation and operation of AI within the agency. • Section 4.1. Gathering Requirements and Conducting User Research provides best practices on conducting discovery research and how to determine the requirements of the proposed solution. • Section 4.5 Governing AI introduces an AI governance framework that advocates for collaboration across roles, responsible AI principles, and governing processes.
Design, Development, and Data Practitioners	• What are the recommended steps in pursuing the implementation and operation of an AI-based solution? • Are there examples of AI implementations within CMS that can serve as benchmarks for new projects?	• Section 4. Implementation and Operation of AI at CMS provides detailed steps for technical practitioners and AI product teams pursuing AI implementation. • Appendix A Case Studies provides overviews of two AI case studies that have been implemented at CMS.
Training and Development	• What are the essential AI technologies, challenges, and considerations that beginners should be familiar with when starting their AI education? • What are the guiding principles for AI at CMS?	• Section 2. Overview of AI Technologies contextualizes AI capabilities, how AI is being used at CMS, and challenges and considerations. • Section 3. Foundations for AI at CMS provides examples for AI implementation use cases at CMS, guiding principles, and future AI trends.

This document does not include guidance for consumer AI tools geared towards individual use (such as ChatGPT, Gemini, Claude, etc.), but the principles and considerations are just as applicable to such scenarios.

2. Overview of AI Technologies

As a high-level definition, the term "artificial intelligence" refers to the scientific field where computers can execute diverse tasks typically requiring the equivalent of human intellect. With origins back to Alan Turing's 1950 paper "Computing Machinery and Intelligence," this concept has been around for decades but only in relatively recent times started to be realized (Turing 1950). The growth of AI can mainly be attributed to the availability of cost-effective, high-speed computational power and rapid growth of accessible data.

The following sub-sections will define core capabilities, break down AI usage and trends, discuss challenges and considerations related to using AI, and provide additional resources.

2.1. AI Capabilities

Included below are the five core capabilities associated with AI. To help establish a baseline understanding, each capability contains a brief definition, a description of key concepts and terminology (Noblis 2023), and an example of use within an existing CMS system. Additional CMS AI Use Cases and Pilot Projects examples are provided in Appendix A. These capabilities are fundamental to the guiding principle of well-grounded and data-driven AI, further discussed in Section 3.2.2.

Learning Patterns

Learning patterns provides the bases for a subfield of AI referred to as "Machine Learning" (ML) in which algorithms are utilized to learn from data and discover patterns to drive further capabilities, such as prediction and decision making, with limited human interaction. Table 3 has key concepts and examples of Learning Patterns.

Table 3. AI Capability Basics: Machine Learning

Key Concepts	Example
• **Supervised Learning** – A time-consuming but effective technique that uses manually labelled data to directly train algorithms to recognize patterns. • **Unsupervised Learning** – Machine learning technique that requires algorithms to discover data patterns without use of labeled data. • **Reinforcement Learning** – Trial and error approach to improve algorithm performance against desired objectives through incentivization (reward maximization or penalty minimization). • **AI/ML Model** – A component that utilizes computational, statistical, or ML techniques to generate outputs based on provided parameters (Department of Homeland Security 2024).	**Medicaid And CHIP Financial (MACFin) Anomaly Detection Model for Disproportionate Share Hospital (DSH) Audit** (United States Government 2023): MACFin utilizes an ML model to predict anomalies within DSH audit data. The model flags outliers in the submitted DSH data to facilitate targeted investigations and support the audit process by minimizing overpayment and underpayment.

Understanding Language

Understanding language relates to the ability of AI to extract linguistic content from a wide variety of media sources (written text, acoustic speech, image with text embeddings), represent the data in a machine-readable format for query or analysis, and generate natural language outputs. This is often referred to as Natural Language Processing (NLP). Table 4 has key concepts and examples of Understanding Language.

Table 4. AI Capability Basics: Understanding Language

Key Concepts	Example
• **Natural Language Understanding (NLU)** – The ability to comprehend meaning from language in its usual representation. • **Natural Language Generation (NLG)** – The ability to produce language to convey meaning in a manner comprehensible to ordinary users.	**Feedback Analysis Solution (FAS)** (United States Government 2023): FAS uses CMS and other publicly available data (such as Regulations.Gov) to review public comments and/or analyze other information from internal and external stakeholders. It applies NLP tools to aggregate, sort, and identify duplicates to create efficiencies in the comment review process and uses ML tools help identify topics, themes, and sentiment outputs for the targeted dataset.

Gathering and Using Knowledge

Gathering and using knowledge represents the realization of the key advantage of NLP by providing mechanisms to derive knowledge (insights and facts of patterns and behaviors) from analysis of processed data. Table 5 has key concepts and examples of Gathering and Using Knowledge.

Table 5. AI Capability Basics: Gathering and Using Knowledge

Key Concepts	Example
• **Labeling and structuring data** – Identification and selection of features and elements on which to develop context. • **Taxonomy / Ontology development** – Creating a structure for how concepts within words, phrases, or documents are categorized and relate to each other. • **Data synthesis and augmentation** – Accounting for unavailable or missing data through creation of new datapoints derived from existing data elements. • **Knowledge management** – Selection of appropriate methods (e.g., for organizing, processes, or retrieving data) and models (e.g., frameworks or algorithms) to manage enterprise data.	**Knowledge Management Platform (KMP)**: KMP is a CMS internal cloud-based AI software solution that uses NLP, ontology development, and ML capabilities that are purpose-built for converting information contained in documents and files into structured data for analysis.

Making Decisions

AI's strength is its ability to make decisions and recommend solutions based on historical data. This capability also allows for automated decision-making (when appropriate) and allows for increased speed, accuracy, scalability, and consistency of the processes determining the output. Table 6 has key concepts and examples of AI for Making Decisions.

Table 6. AI Capability Basics: Making Decisions

Key Concepts / Terminology	Example
• **Predictive** – Machine learning technique which uses historical analysis to forecast future outcomes/results. • **Prescriptive** – Machine learning technique which provides recommendations based on modeling of historical performance from similar decision environments.	**Predictive Intelligence (PI) - Incident Assignment for Quality Service Center (QSC)** (United States Government 2023): Predictive Intelligence analyzes short incident descriptions provided by the end user within the QSC to identify key words that also appear in previously submitted incidents. Then, tickets are assigned to the appropriate assignment group based upon the analysis. The model is re-trained on updated incident data every 3-6 months.

Enabling Creativity and Generating Content

Enabling creativity and generating content refer to a combination of multiple techniques that allow AI to produce new and innovative outputs in creative, functional, and engaging ways. Table 7 has key concepts and examples of Enabling Creativity and Generating Content.

Table 7. AI Capability Basics: Enabling Creativity and Generating Content

Key Concepts / Terminology	Example
• **Generation and composition** – Models that utilize algorithms to produce creative output based upon existing data. • **Manipulation** – Models that manipulate existing content (such as image or text) and alter it based upon parameters (changing tone of text). • **Optimization** – Models that optimize existing creative processes and identify new implementations (better layouts, color schemes).	**AI Explorers Use Case – Using Generative AI Solutions to Create a Multi-Model Interface for Enhanced User Accessibility**: The generative AI use case led to the prototype / proof-of-concept development of a web-based chatbot interface that utilized both voice-to-text and text-to-voice user interaction to ask questions of preprocessed documents and output the generated response. The proof of concept aimed to test open-source generative AI models to create an interactive platform that improved accessibility by enhancing user engagement (CMS AI Explorers 2024).

2.2. AI Reported Usage and Trends

To demonstrate the prevalence of AI use in the government sector, an analysis of the annual Consolidated AI Use Case Public Inventory (last updated September 2023) showed a total of 710 reported instances of AI use to improve services provided across 21 different departments. The top 10 departments reporting the highest number of AI-related use account for 87.9% (624 out of 710) of all instances and represent a wide variety of governmental services from the Department of Energy (DOE) to the Department of Labor (DOL).

Shifting to the relevant department-level focus specifically on the Department of Health and Human Services (HHS) (comprising 22.1% of the total government-wide instances), a total of 157 use cases were reported across eight HHS-related agencies. As depicted in the chart below, the top three agencies include the National Institutes of Health (NIH) (30%), the Food and Drug Administration (FDA) (28%), and CMS (15%).

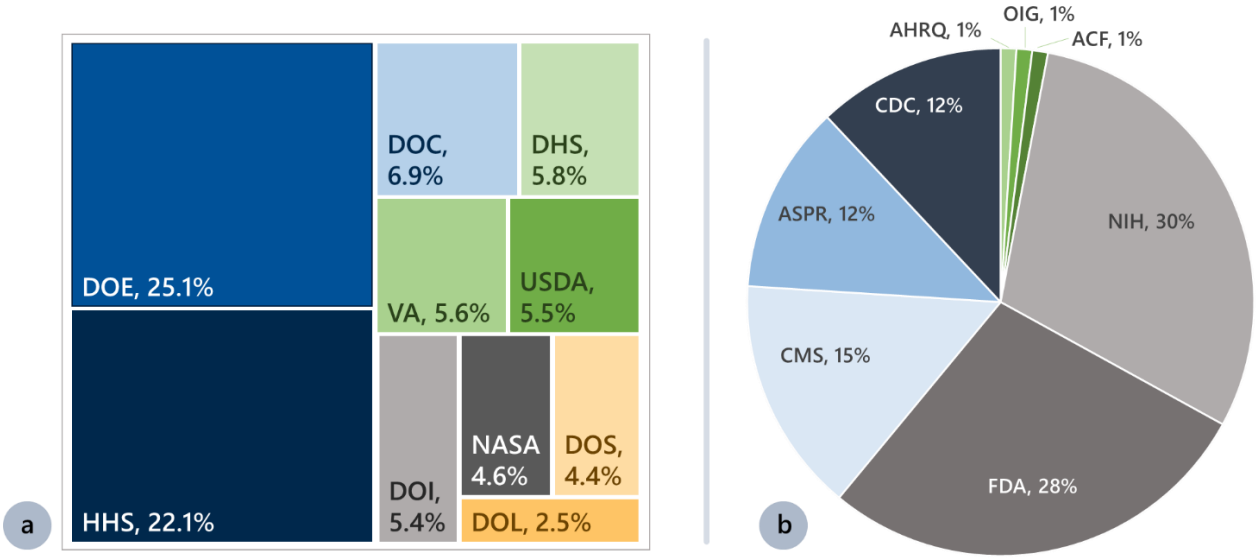


Figure 1. a. Top 10 Departments by Number of Reported Public AI Use Cases (September 2023); b. AI Use Case Breakdown for HHS Agencies (September 2023)

While the government-wide AI public inventory focuses broadly on use cases involving AI, the CMS System Census[1] provides an annual reporting mechanism for components within the agency to identify systems that includes tracking of AI usage (CMS 2024). Since 2021, when the census first tracked AI usage, the number of systems reporting current or planned usage has steadily grown (as depicted in the chart below).

As of 2023, the 56 CMS Systems reporting AI current usage or plans for usage represent 28.6% of all 196 official systems tracked by the agency. Although the total number of systems reporting AI usage between 2022 and 2023 went up by ten, the percentage of total system use remained at 28.6%. To highlight the breakdown of AI capabilities currently in use by CMS systems, the chart below shows reported categories for the 17 systems. More than half of these CMS systems reported usage of the following AI capabilities, listed in descending order: Classification, Process Efficiency Improvement, Natural Language Processing, and Anomaly Detection and Correction.

[1] The CMS System Census is conducted and reported on an annual basis, with results accessible to all users with access to CMS SharePoint.

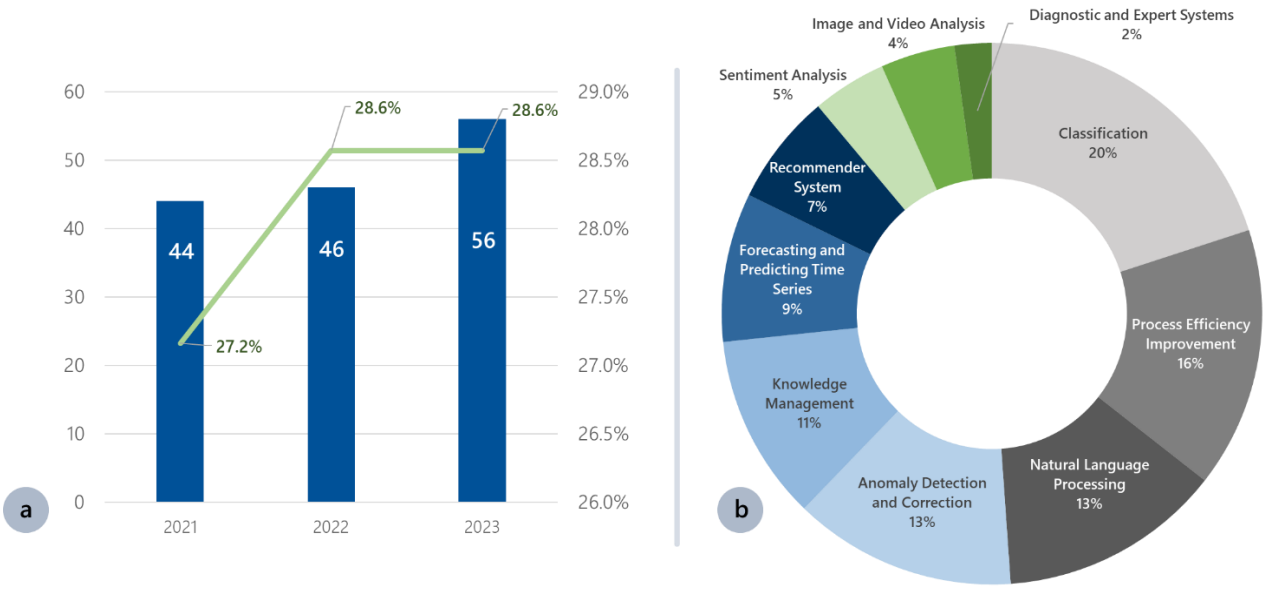


Figure 2. a. CMS Systems Reporting AI Usage (CMS System Census 2021-2023); b. AI Capabilities Identified in CMS System Census (2023)

As shown above in the high-level analysis, AI currently has a broad impact across the breadth of services provided at both the government-wide and agency-wide scale. Since AI is a rapidly developing and maturing field, it's likely the total reported usage will increase and the scope of application will extend to more and more departments and agencies. For instance, within CMS the interest and discussion of incorporating Generative AI has gone up with the release of tools such as OpenAI's ChatGPT, which will likely result in increased usage in future analysis.

2.3. Challenges and Considerations

Agencies determining whether to use AI tools or capabilities to help accomplish a task or project should consider the potential barriers, which fall into three high-level categories: technical, societal, and organizational (Endra 2023). The following sub-sections define and describe each of these categories to provide a baseline of topics to be aware of when implementing AI and adhering to the example guiding principles discussed in Section 3.2. Section 4.6.1 offers additional information on addressing challenges, continuous improvement, and best practices.

2.3.1. Technical Barriers

Technical barriers to consider when evaluating the use of AI include data limitations, infrastructure challenges, security/privacy concerns, and model development complexity. AI works best with large volumes of high-quality and relevant input data to produce reliable outputs, which presents challenges in existing data collection, cleansing, and labeling processes. Existing collection processes may not even exist or be inadequate to provide the data necessary to support the AI development efforts. Data quality issues may require additional effort focused on cleansing and preparation before use with AI. The labor-intensive process of data labeling poses a final challenge that can greatly impact the output of AI.

Infrastructure challenges pose another significant barrier. Advanced hardware and computational power are often needed to achieve results efficiently. These resources can come with a significant cost and present additional challenges when integrating with existing systems.

Security and privacy concerns are additional technical barriers to consider and add to the overall importance of existing IT policies. Since AI uses a large amount of data, systems containing sensitive data and using AI-based tools require robust encryption and protection mechanisms to ensure the safety of the data. These mechanisms must also adapt as AI use grows and evolves. AI-enabled systems also require security measures to protect against data breaches and potential misuse.

Finally, model development complexity refers to the steep learning curve associated with AI-related development, including expertise with algorithms, mathematics, programming, and domain-specific knowledge. With this complexity comes the additional challenge of model development too closely aligned with training data (known as "overfitting") that affects performance against real-world scenarios, and propagation of human bias resulting in unfair outcomes.

2.3.2. Societal Barriers

Societal barriers to consider when evaluating the use of AI include ethical use, job displacement, job replacement, and overall trust and acceptance. Ethical use of AI means avoiding the potential for biased outcomes and establishing of trust via transparency and accountability, as discussed in the technical barriers above. Use of AI can also raise concerns over the possibility of technological advancements causing people to lose jobs (job displacement). Another concern is the shift of job skill requirements from unskilled to more skilled, or consolidation of duties handled by automation (job replacement).

Establishing trust and acceptance of a models' output is critically important to the success of its use and requires transparency of results to ensure compliance with any potential regulatory standards. Consideration for responsible AI (RAI) and its building block of explainable AI (XAI) can be used to ensure the fairness, robustness, privacy, security, and transparency of AI usage (Baker and Xiang 2023). This approach helps reduce the occurrence of "black box" models, where only the input and output values are known to the user, and not the logic behind the result.

2.3.3. Organizational Barriers

Organizational barriers to consider when evaluating the use of AI include cultural resistance, skill gaps, and cost implications. From a cultural standpoint, AI may pose challenges to existing, well-established processes and operational goals. While this may lead to improvements, it can also create resistance among employees who are concerned about the role of AI in the organization. Where there are skill gaps in the workforce, hiring managers may need to place additional emphasis on finding and retaining the right people to do the job. On a positive note, the identification of a skill gap can result in opportunities for staff to receive additional training. Finally, organizations must account for the financial implications of AI, such as the high initial and recurring costs associated with new technologies and the overall return on investment.

2.4. Resources

AI is a rapidly changing and evolving landscape, with new technologies and advancements occurring at a fast pace. For staff to stay current with emerging technologies, it is good practice to utilize trusted resources at various levels of focus to maintain skills and knowledge of concepts, tools, and methodologies. Internal initiatives like the CMS AI Explorers Program provide forums to exchange learning resources, engage in discussions with colleagues on relevant topics (such as the #ai_community Slack Channel), and review an inventory of in-progress and completed AI-related pilot projects. Public-facing sites like the Artificial Intelligence at CMS Site provide AI governance-related information on existing AI programs and initiatives within CMS.

At a higher level, another valuable resource is the HHS Office of the Chief Artificial Intelligence Officer (OCAIO). This office drives HHS AI implementation strategy and governance, facilitates collaboration across all HHS entities, and is the source of key publications such as the HHS Trustworthy AI Playbook.

Lastly, standards organizations such as the National Institute of Standards and Technology (NIST) are a reliable source for AI research, and AI.gov provides the latest information on policy and administration actions. From an industry perspective, institutions such as the Stanford University Human-Centered Artificial Intelligence (HAI) provide well-vetted research and reports to demystify AI. Other groups such as Gartner regularly evaluate and publish reports regarding trends and guidance for incorporation of AI, such as the Gartner Hype Cycle for AI.

3. Foundations for AI at CMS

3.1. CMS and AI: Setting the Context

3.1.1. A Hope for AI in CMS

CMS has access to extensive data resources, providing the agency with the opportunity to use AI to improve healthcare quality and equity for millions of beneficiaries. The agency has the potential to improve in areas such as data insights, decision making, management efficiency, beneficiary-centric services, and cost management through a deeper understanding of AI implementation.

To encourage implementing AI responsibly, this document focuses on workforce education, AI exploration, best practice sharing, and developing proof-of-concept initiatives, all within the framework of the Example Guiding Principles of AI at CMS, covered in Section 3.2. The CMS AI Playbook documents these efforts and serves as a reference for AI project teams, making guidance and best practices accessible to federal staff and the public.

3.1.2. Examples of AI Implementation at CMS

AI experimentation, workforce upskilling, and community building are essential for creating a positive and responsible AI culture within CMS. AI experimentation, conducted with the goal of providing solutions to business use cases, has been encouraged through the AI Explorers Pilot Program (CMS AI Explorers 2024) and the AI Health Outcomes Challenge (CMS n.d.). CMS has hosted innovation spaces for project teams to propose use cases for AI and has funded research and development of proof-of-concept projects. CMS has also invested in upskilling its workforce by hosting courses and collaboration spaces covering AI and data science concepts (CMS n.d.). Lastly, CMS has provided an internal, online AI community space via Slack for employees to share AI-related questions, best practices, and examples of use cases.

3.1.3. AI Case Studies

In the last few years, CMS teams have begun to experiment with designing and developing AI-based tools and software. The following are examples of projects that are currently using AI. Additional case studies are provided in Appendix A.

CPI Fraud, Waste, and Abuse Prevention

The Center for Program Integrity (CPI), a division within CMS focused on combatting fraud within Medicare and Medicaid programs, is using AI to analyze large amounts of data and look for patterns and anomalies such as unusual billing behaviors that could indicate fraudulent activities. Once a fraudulent scheme is detected, it is managed through vulnerability analyses, referrals to law enforcement, administrative measures, or by implementing adjustments to current policies and regulations (Whitfield 2023).

OIT Knowledge Management Platform (KMP)

The Office of Information Technology (OIT), responsible for managing and overseeing the agency's essential IT services, is currently implementing the KMP. This AI-based tool gathers

unstructured data from various CMS repositories, including text documents, spreadsheets, and video, and then structures this information. This allows business owners and data scientists to easily analyze the data, identify patterns, and make well-informed decisions promptly. Prior to the implementation of KMP, the Information Security and Privacy Group (ISPG) team manually extracted data from document sources like System of Record Notices (SORNs) and Privacy Impact Assessments (PIAs) for Computer Matching Agreements (CMAs). This process previously took fifty-one days, consuming significant time and resources. However, by implementing KMP, this previously manual process now takes just a few minutes (Tompkins 2024).

3.1.4. Aligning AI Goals to CMS

As AI maturity increases across the organization, aligning AI with goals tailored to CMS means integrating AI technologies in ways that advance healthcare quality, accessibility, efficiency, and affordability for Medicare and Medicaid beneficiaries (CMS n.d.). Implementing AI at CMS can support CMS's goals by:

- Enhancing operational efficiency, such as automating administrative tasks, streamlining claims processing, and optimizing resource allocation.
- Managing health care data more effectively, enabling CMS to make data-driven decisions that improve program management and policy formulation.
- Supporting equitable access to healthcare services, addressing disparities, and ensuring beneficiaries receive high-quality care.
- Identifying cost-saving opportunities, reducing waste, and making healthcare more affordable for beneficiaries and the government.
- Assuring compliance and quality by monitoring healthcare providers' compliance with CMS standards and regulations.

Aligning AI with CMS goals requires careful consideration. The guiding principles covered in the next section will provide guidance on implementation and emphasize a holistic approach to the design and development of AI-based projects, tools, and software.

3.2. Example Guiding Principles for AI at CMS

CMS is a significant contributor to our nation's healthcare sector. The agency's actions set an example and have many downstream impacts. These effects don't happen in a vacuum, especially when it comes to AI, which sits at the crossroads of people, tools, data, agency integration, and ethics. This section introduces four principles to serve as an exemplar foundation throughout the Playbook in guiding the implementation and operation of AI within the agency.

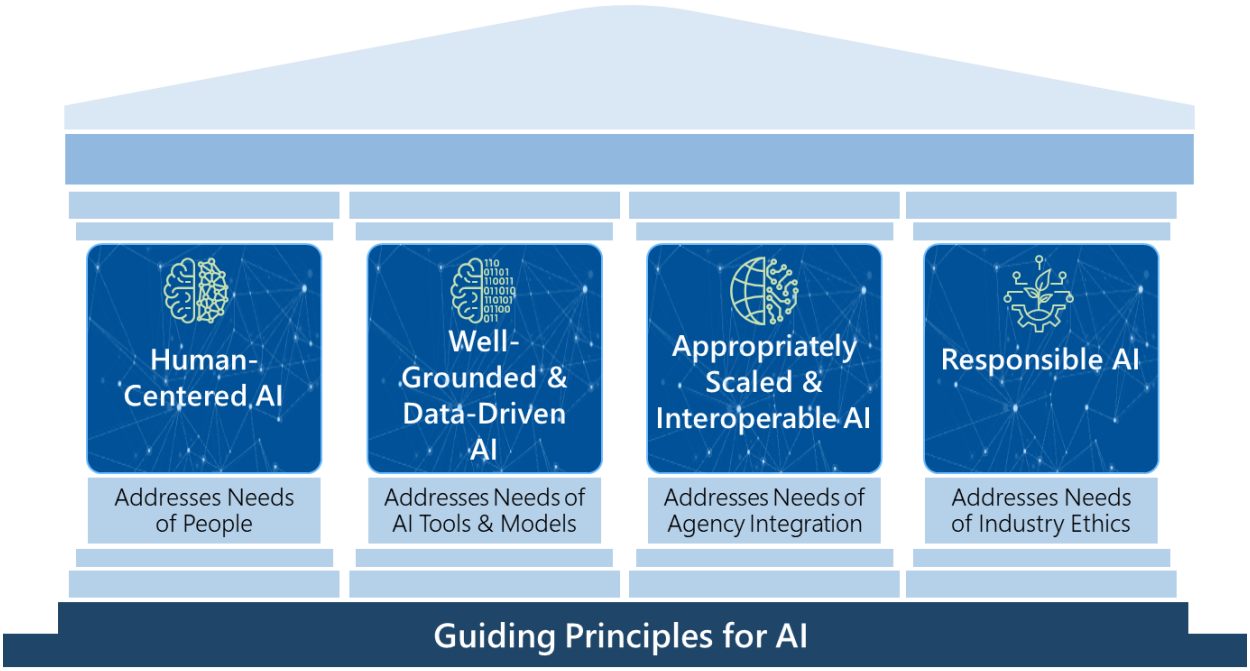


Figure 3. Example Guiding Principles for AI

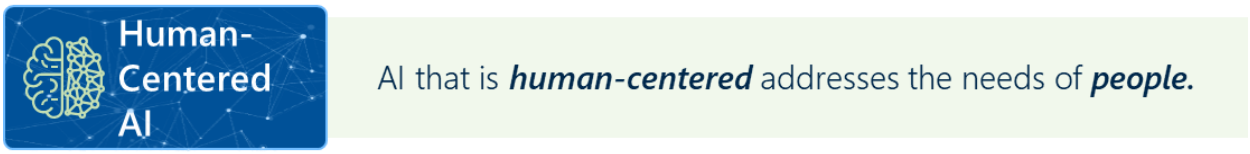
Human-centered AI (HCAI) addresses the needs and prioritizes the well-being of end users and society, emphasizing the human-centered practices throughout the design and development process.

Well-grounded and data-driven AI addresses the needs of AI tools and models, ensuring that they are based on the right information for the right task.

Appropriately scaled and interoperable AI addresses the needs of agency integration, ensuring AI adoption within CMS is incremental and can adapt to the agency's needs.

Responsible AI (RAI) addresses the needs of industry ethics, applying trustworthy best practices and federal guidance to balance the complexity of moral principles and system requirements. Responsible AI domains include fairness and impartiality, transparency and explainability, accountability and compliance, safety and security, privacy, and reliability and robustness.

3.2.1. Human-Centered AI



Principle in Brief: HCAI emphasizes the impact of AI technologies on individuals and society. HCAI is guided by human values, needs, and goals. By building on user experience design methods, resultant AI systems and products will be designed to "amplify, augment, empower, and enhance human performance" while emphasizing human control (Shneiderman 2022).

AI brings new capabilities and sophistication to decision automation that require specialized care. As AI proliferates in our everyday lives, it is important to build positive relationships between people and technology so that there is trust, acceptance, and human involvement where needed throughout the process. By aligning AI with human values, needs, and goals, AI can enhance rather than replace human capabilities, reinforcing human control and positively impacting individuals and society. To do this, HCAI emphasizes the impacts of AI on individuals and society through process-oriented approaches, navigation of challenges and risks, and human evaluation.

Similarly to person-centered care in CMS's Innovation Center (CMS n.d.), AI services should be "delivered in a setting and manner that is responsive to individuals and their goals, values, and preferences" through systems which empower developers and stakeholders to communicate and make effective AI-based tools and software together. HCAI emphasizes a process-oriented approach where humans are directly involved in the design, development, implementation, and evaluation of AI systems (Shneiderman 2022). "HCAI builds on user experience design methods of user observation, stakeholder engagement, usability texting, iterative refinement, and continuing evaluation of human performance in use of system that employ AI and machine learning" (Shneiderman 2022). The human-centered approach prioritizes the end users' needs and considers the context of their experiences to determine how the AI-based solution could support their goals. User experience design methods incorporate needs and challenges into methodology to help teams navigate through what is often a "value maze" of missing data, dependencies, stakeholder concerns, and more.

The challenges and risks identified through HCAI approaches demonstrate some of the potential negative impacts of AI on human well-being. In addition to accounting for negative impacts on individuals (e.g., privacy), society (e.g., civil rights), or even the planet (e.g., environmental), the healthcare domain will have its own set of risks, such as health inequity and erosion of beneficiary trust. HCAI requires efforts to identify and minimize these negative impacts.

Further enabling the effectiveness of AI requires collaboration and trust between humans and AI-based technology. To build this trust, humans should be involved in the evaluation and oversight of AI systems, both in its performance and its ethical implications. HCAI will equip humans with the necessary tools, information, and opportunities to access, measure, and control the effects of AI to build confidence in AI outcomes. Even then, HCAI practices should be sustained, as they could reveal changes to human-centered values or needs over time. This oversight will guide invaluable AI maintenance or identification of new AI use cases.

Explore Human-Centric AI in the Playbook

- **HCAI Processes:** 4.1, 4.3.1
- **Challenges and Risk Mitigation**: 2.3, 4.1.4, 4.6.1
- **Human Evaluation and Oversight:** 4.1.4, 4.2.3, 4.4.1, 4.5

3.2.2. Well-Grounded and Data-Driven AI

AI that is ***well-grounded data-driven*** addresses the needs of ***AI tools and models.***

Principle in Brief: Data is the foundational element that AI needs to learn, be assessed, and improve. The quality of an AI model is directly correlated with the quality of the data used to train it. Therefore, for any organization to work toward understanding and applying AI, it must first understand and effectively use its data. This will require thorough review into datasets themselves, as well as overarching data governance to guide organization-wide data policy.

AI needs data. Models are trained on data, evaluated against data, and generate data. Sometimes, their outputs are even used to train other models, so the cycle restarts. With AI, the colloquialism "garbage in, garbage out" rings true. AI-based tools and models need to be trained on reliable data – that is, data that accurately reflects whatever aspect of the world the user intends to model – to be considered well-grounded. AI that is well-grounded and data-driven will be led by projects that take the necessary steps to evaluate data quality and impacts while upholding data governance.

AI models aim to "understand" the world by recognizing patterns and relationships in data to guide some kind of output or decision. To provide the best foundational inputs for models, exploratory data analysis (EDA) and responsible data evaluation make up the early and often complex phase of preparing data to be used with AI technology. EDA is the process of investigating and cleaning data to mitigate common data challenges such as data availability, quality, quantity, and feature relevance (Noblis 2023). Beyond surface-level statistics, responsible use of data also requires evaluating the contexts within the data and the impacts they could have for components of the real world, such as related to representativeness, bias, privacy, and security. In cases where synthetic data is used, it must be evaluated against real and reliable data with a high degree of confidence that the process to generate the data accurately reflects the real data it is attempting to imitate (Gonzales, Guruswamy and Smith 2023). Use of foundational models similarly warrants extra care if they were trained on external and unknown data. Data users and domain experts will need to work together to ensure the thoroughness of their exploration to understand the full picture of data needs and potential impacts.

Specific techniques for exploring and evaluating data for use in AI systems should be guided by organization-wide policies known as "data governance." These policies can more broadly describe how data is managed within the organization, from roles, authorities, and IT resources to what criteria must be met for specific data to be ingested by an AI model (Federal Data Strategy Data Governance Playbook 2020). Concerns that are included in governance guidelines may include, but are not limited to, security implications, data availability, accessibility, traceability, and sufficient data infrastructure. CMS's data principles and operating norms developed by the Data to Drive Decision Making Cross-Cutting Initiative (Data CCI) set a great starting point for data governance development (Centers for Medicare and Medicaid Services n.d.). Both internal and external entities, from CMS roles (e.g., ISSO, CRA) to patrons and consumers (e.g., providers, beneficiaries), play a role in defining governance needs and implications.

Explore Well-Grounded and Data-Driven AI in the Playbook

- **Data Evaluation**: 2.3, 4.1.3, 4.2.1, 4.3.1, 4.4.1, 4.6.2
- **Governance: 4.4.1**, 4.5, 4.6.2

3.2.3. Appropriately Scaled and Interoperable AI

AI that is ***appropriately scaled and interoperable*** addresses the needs of ***agency integration***

Principle in Brief: Emphasizing complete scalability and interoperability during an agency's early stages of AI adoption puts the cart well before the horse. AI adoption should take an incremental approach which can adapt to maximize the value of AI efforts based on the agency's maturity level and current needs.

Appropriately scaled and interoperable AI refers to the organization's tactical planning and delivery of AI efforts to best suit its current needs. Organically building CMS's scalability and interoperability will require assessment of the agency's current AI maturity levels, adopting a "think big, start small" approach, and continuously measuring and improving upon past processes.

To effectively design for appropriately scaled and interoperable AI, agencies must first acknowledge that different levels of AI maturity exist, and that an agency's current maturity level will indicate different priorities and effective guidance. AI maturity levels span *small-scale* (pilots and proof of concepts), *medium-scale* (operational products), and *large-scale* (larger operational solutions) AI capabilities within an agency. CMS has completed numerous successful small-scale pilots and proofs of concept projects. Stakeholders across the agency have also begun researching, building, and acquiring the necessary tools, frameworks, infrastructure, and culture that demonstrate the early transition into medium-scale.

Successful programs at CMS have been guided by holding two ideas in tension with each other — *think big* and *start small*.

To ***think big*** means to establish an ambitious program vision to effect a specific change over multiple years, and to articulate and re-articulate it to key stakeholders with each progress update. To ***start small*** means to identify the smallest unit or minimum viable product of the program that can be de-risked and demonstrate proven value to begin a virtuous cycle of outcomes and investment with minimal wasted effort or rework.

In contrast, it's key to avoid *thinking small* and *starting big*. To ***think small*** is to propose a hard-to-justify allocation of management attention and financial resources. To ***start big*** is to engage in a program so large, it fails to identify and resolve major risks before a vicious cycle of no outcome, flagging investment, and substantial wasted effort and rework takes hold.

The largest impact from a new model, application, or data product often happens at its earliest releases for users. Thus, an organization should break down the most complex and impactful human-centered targets in their early AI initiatives. These targets will often align to the mission and purpose of the CMS component, division, or group tackling these initiatives. Addressing associated risks will help ensure that technology platforms, regulatory frameworks, staff skills, and project management disciplines can scale as the organization adopts AI more widely, without needing significant redesign. Consistent maintenance should be made on data and algorithms so that the AI evolves in parallel to the real dynamic environment that reflects human-centric needs and values.

It is important to clarify to practitioners and other business stakeholders is that large-scale and fully interoperable AI doesn't happen immediately and must be planned for. The organization's progress and maturity level should be measured and improved in a continuous process through regular assessments of best practices. Best practices will include efforts to teach the AI mindset, skills, collaboration, open-source, and governance approaches discussed throughout this playbook. Measurements can include monitoring the continued transition from siloed datasets to appropriately shared data stores, increasing the number of data literacy programs being offered to employees, and tracking results from R&D teams. Interoperability criteria should be assessed and assured in both AI development and procurement.

Explore Appropriately Scaled and Interoperable AI in the Playbook

- **AI maturity at CMS:** 2.2, 3.1, 4.2.2
- **Think big, start small:** 4.1.1, 4.3, 4.5.4
- **Measure and improve:** 3.4, 4.3.1, 4.5.4, 4.6

3.2.4. Responsible AI

AI that is ***responsible*** addresses the needs of ***industry ethics.***

Principle in Brief: RAI is the well-intentioned practice of designing and implementing AI projects to uphold society's unambiguous moral values (i.e., social ethics), system safety, and system security. Ethics serve as a guide for social behavior against risks of undesirable outcomes, and this notion is preserved in RAI practices. The playbook focuses on six RAI domains to navigate the competing risks and benefits of AI: fairness and impartiality, transparency and explainability, accountability and compliance, safety and security, privacy, and reliability and robustness. These domains aim to guide AI endeavors toward their intended impacts through the most effective and trusted means.

As additional federal guidance and legislation related to AI are released, it becomes clearer and clearer that there are expectations for how CMS and other agencies incorporate RAI. AI is a growing field with opportunities for high impact, high visibility, and varying levels of trust. This is especially true in the federal space. Yet the legal landscape surrounding AI is in early development. To guide RAI at CMS, this playbook provides a set of RAI domains that align with existing federal guidance and moral principles associated with CMS's work to further instill human-centric values and best practices for trustworthy AI.

Practicing AI responsibly calls for establishing domains aligned with ethical principles that can be defined, operationalized, measured, and evaluated. However, due to the subjective nature of ethics and individual morals, there is no singular criterion that applies RAI universally to equal effect. Many organizations have published their own versions of RAI principles and guidelines in the past few years, such as the IEEE's Ethics of Autonomous and Intelligent Systems (2017) (IEEE Global Initative 2017), the Defense Innovation Unit's RAI Guidelines (2020) (Dunnmon, et al. 2021), HHS's Trustworthy AI Playbook (2021) (HHS Trustworthy AI Playbook 2021), and the Department of Defense's (DoD) Responsible AI Strategy and Implementation Pathway (2022) (DoD Responsible AI Working Council 2022). All of these documents share the same purpose – to ensure their AI systems function fairly, as intended, and are accountable to their results in accordance with the principles and norms defined within their own organizations and domains.

The six principal domains of RAI presented in this section are predominately aligned with HHS's trustworthy AI principles, and include:

- fairness and impartiality,
- transparency and explainability,
- accountability and compliance,
- safety and security,
- privacy, and
- reliability and robustness.

These domains prioritize the needs of CMS as a federal agency serving its patrons (i.e., providers, beneficiaries, payers, and the research community) which may differ from the needs of other organizations or agencies. High-quality AI programs gain trust and increase effectiveness by implementing RAI throughout design, build, and testing phases. The ethical considerations highlighted under each domain provide a glance into the complex considerations of RAI. There will be instances when teams find themselves with several RAI principles in conflict or contention. These considerations should all be given standards and metrics to determine their impacts from the beginning, with preparation for those measures to change as the AI program evolves. Subject matter experts (SMEs) and data owners are good sources for determining these metrics. At the organization level, RAI is the primary element within a larger AI governance framework that uses it to guide the same complexities in the agency's AI policies and procedures.

1. Fairness and Impartiality

Fairness in AI requires addressing bias and discrimination, and ensuring equitable access, which can occur at different stages of the AI application lifecycle. Bias in AI is the influence of prejudiced assumptions in data, and from data users, which carry into the algorithm's output. A biased dataset might not be representative of the real-world or is unbalanced such that is fails to provide sufficient representation of various classes. Consider: Does the training data include any sensitive variables? Does the output perpetuate racial, gender, or other stereotypes? Ideally, an AI tool would not tend towards any human, systemic, or institutional bias that can impact decision making and perpetuate inequities.

Likewise, all parties must be treated fairly both in the learned behavior within the AI model as well as the use case context and application provided to users. Specific questions and resulting assessments of impartiality may vary greatly depending on context, which itself can change over time and with societal and cultural change. Therefore, it is crucial to build in and regularly update checks and balances for evaluating contexts. Impartiality extends to an AI system in whole as a series of products.

From model definitions to user interfaces, all components will need to consider how potential algorithmic and human biases and inequities can lead to unintended consequences. Language and accessibility challenges, for example, are considerations which need to be addressed in AI systems early on.

It is the responsibility of data scientists, analysts, designers, and knowledgeable stakeholders to recognize inherent bias and inequities within an AI-based solution, remove as much as they can, and provide transparency and access to this information to users and stakeholders.

2. Transparency and Explainability

Relevant stakeholders must be made aware of the usage and function of an AI tool within the decision-making process. Furthermore, integrated techniques should offer clear explanations for both direct users and those impacted by the AI-supported decision-making to comprehend the reasoning behind the AI's purpose and behavior. The decision-making process used by an AI model will need to be contextualized and easily explained for less technical individuals. Moving away from the "black box" approach to AI development has been an important area of research, and tools that provide model explainability are constantly evolving. Examples of these tools include SHAP and LIME for identifying feature relevance (what features influenced the outcome), and DICE-ML for counterfactual explanations (alternative scenarios and their outcomes). Explainability in AI models not only fosters trust and confidence in AI systems, but also allows for more effective AI integration and better-informed analysis and feedback from stakeholders. Best practices are to be up front about the capabilities, limitations, and intent of the AI product, and to provide transparency into reliability levels and potential underlying risks and biases in the AI's output.

Being able to communicate these RAI efforts effectively typically requires an interface intentionally designed for explainability. User experience (UX) designers must consider how to best display outputs provided by the AI algorithm to an end user and ensure the user trusts the output to an appropriate degree through model explainability and transparency efforts. The user should be given all the information they will need to use the interface easily and intuitively, as well as understand and create value from the results of interacting with the AI. In practice, creating intentional user interface (UI) copies that prompt the user to experiment with the interface can strengthen trust and open opportunities for improvement. Consider utilizing tooltips, explanations, and evaluation measures to convey appropriate details and transparency. Suggesting users take an action as a result of the AI's prediction allows users to better understand why the AI's prediction is useful. Finally, creating channels for feedback promotes continuous improvement of the effectiveness of an AI tool's interface and overall UX.

3. Accountability and Compliance

AI systems must establish governance that ensures accountability and compliance throughout all aspects of an AI-based solution, from initiation to decommissioning. This requires policies with associated roles and responsibilities to specify authority and liability over tasks, risks, and compliance.

At an agency level, leading roles such as the Chief AI Officer (CAIO) and AI Governance Boards must have clearly defined responsibilities and levels of authority to propel the necessary governance. The Memorandum for Advancing Governance, Innovation, and Risk Management for Agency Use of AI defines these roles and their responsibilities for federal agencies (Young 2024).

Delving deeper into the organization, individual offices, components, and especially project teams themselves will require numerous roles with more explicitly designated authority for allocating responsibilities and maintaining compliance. These responsibilities can include approving and monitoring certain aspects and outputs of the system; implementing digital identity management to monitor and regulate impacts of the system; managing the supply chain of AI tools, resources, and data; and ensuring traceability and auditability of the decision-making processes and factors that make up each AI-based solution. The roles across the federal government and within CMS will continue to develop and adjust to address changing needs of the organization and the demands of regulations.

Many governing policies and responsibilities directly address keeping abreast of and adhering to regulations and compliance. Existing legal frameworks, even without direct nods to AI, still apply to most AI programs. Notable examples include laws surrounding data use, privacy, discrimination, and intellectual property. These also include existing standards mandating that organizations monitor and document all attempts to minimize unintentional discrimination, generate good faith justifications for AI use, and uphold and communicate privacy and security relating to user data within the AI. Policies and regulations that may apply to AI usage vary across organizations and are changing as AI becomes more widely used, requiring agencies to develop AI tools with both current and potential future legal landscapes in mind.

4. Safety and Security

To support the implementation of safety in AI systems, there must be protections preventing both physical and digital endangerment of human life, health, property, and the environment. This is achieved through responsible cybersecurity, development, and deployment practices from system owners as well as clear communication of responsible use and risks.

All systems should identify potential security risks and vulnerabilities across the AI-based solution and determine potential impacts and mitigations for these risks. Implementing the proper access controls and authorizations coincide with addressing these concerns, where analyzing, carefully managing, and guiding how users perceive and interact with the AI can mitigate the introduction of security risks. Security measures must also include robust testing against adversarial attacks. The risk of cyberattack applies to AI systems as much as any other digital product. Specific attacks aiming to exploit AI algorithms can include manipulating input data to lead to false predictions, poisoning the dataset provided to the learning algorithm, or intercepting and breaching sensitive system or user information. The risk of users being granted unauthorized access to information from training data or other unnecessary privileges, for example, must be protected against regardless of whether the user has malicious intent. Security measures must identify such vulnerabilities and test systems against simulations of these attacks to strengthen the AI tool's ability to resist them. Adversaries will always be refining their methods or developing new attacks, so these security efforts may change over time and must be reviewed and revised periodically.

5. Privacy

Prioritizing privacy values such as anonymity, confidentiality, and user control are essential in the design, development, and deployment of AI systems. This helps ensure the protection of human autonomy, identity, and dignity, while also safeguarding personal data from intrusion and unauthorized use.

Data within AI-based systems, either through model training, use in deployment, or other API connections with non-CMS systems, are subject to various privacy risks. Security-related concerns carry over to privacy when sensitive data is involved, for example with protection of data in transit or prevention of data exposure in model outputs.

While the United States does not currently have a comprehensive set of data privacy laws for AI-based solutions, many existing privacy laws for certain states, industries, or subcategories of data can still apply. CMS follows applicable law including the HIPAA Privacy Rule (for instances involving PHI) and the Privacy Act (for instances involving personally identifiable information [PII]) and enforces additional CMS policies and compliance documents through its Privacy Office.

In many cases, the collection of personal information typically implicates laws on individual privacy rights, requiring that users are notified of what the applicable practices are and when they change. Communication of user privacy policies should be user-friendly and allow for diverse stakeholders at all levels to easily understand the following points about the use of their information: What data is collected? What is my data being used for? For how long will my data be used? Who is my data being shared with? How is my data being protected? What control do I have over my data usage? How can I access and manage the data collected about me?

6. Reliability and Robustness

As with any software subject to data dependence or with a stake in decision-making, AI-based tools must be reliable and robust. Reliability refers to the ability to perform as required, without failure, under given conditions, providing valid and accurate results from properly functioning models and systems. Robustness refers to the ability to maintain performance and accuracy in the presence of unexpected or adverse conditions, a feature often difficult to achieve under the highly variable conditions in the real world.

Implementing reliability and robustness requires thorough AI testing and evaluation both pre-deployment and ongoing post deployment, as AI algorithms are dynamic and change overtime (as compared to traditional software). The earliest risks against reliable performance can come from poor or mismatched data used to develop the AI. Beyond validating data and data assumptions, the output of an AI system may not be fixed because of how it learns, trains, and adjusts to new information. Likewise, performance can degrade over time due to concept drift, which happens when the models or other system processes are unable to identify and adjust to real world shifts in data or behavior. Consider: What is the scale of the project and how does that affect which tests are needed (e.g., performance, consistency, security) and at what frequency? Do testing methods align with technical standards and is there a process for rollback if the AI tool does not behave as planned? What resources might be needed to retrain or update AI models to ensure continuous learning and robustness over time?

In addition to unit tests on isolated components of the AI system, integration tests are needed to understand how individual ML components interact with other parts of the overall system (across, up, and downstream). System quality checks should go through exhaustive testing with built-in mitigations. User feedback, evolving user needs and inputs, and model performance indicators should be monitored over time, with updates in response to any changes.

Teams will realize that the reliable and robust principle has the most dependencies and tradeoffs with the other RAI domains compared to its counterparts. Needing safe and secure controls to protect the system from threats, for example, is precursory to robustness. Meanwhile, a common tradeoff in some cases applies when the prohibition of certain sensitive information from training data suppresses accuracy and representation. Formal and comprehensive team decisions will determine the balance of RAI considerations appropriate for that AI tool's context and sensitivity.

Explore Responsible AI in the Playbook

- **Fairness and Impartiality**: 2.3, 4.1.6, 4.2.1, 4.5.2
- **Transparency and Explainability**: 2.3, 4.2.1, 4.2.3, 4.3.1, 4.5.2
- **Accountability and Compliance:** 2.3, 4, 4.1.2, 4.4.1, 4.5.2
- **Safety and Security:** 2.3, 4.3.1, 4.5.2, 4.5.3
- **Privacy**: 2.3, 4.3.1, 4.5.2
- **Reliability and Robustness:** 4.3.1, 4.4.1, 4.4.2, 4.5.2

3.3. Future Directions and Trends in AI

The next wave of AI technology is set to revolutionize the way CMS operates, from improving the beneficiary experience to streamlining operations. This section will explore how emerging technologies can make healthcare more efficient and accessible. It will start with a high-level introduction of several AI-driven emerging technologies that may be pertinent to CMS, followed by specific use-cases of these technologies in Section 3.3.2. The technologies covered in this section may not be employed in CMS or expected in the near future but are introduced to spur the reader's imagination.

3.3.1. Emerging Technologies

Emerging technologies offer exciting potential for CMS by enhancing healthcare delivery, operational efficiency, and the overall experience for employees and customers. AI can help people navigate through the medical landscape regardless of their language skills or educational levels. For medical providers, hospitals can be redesigned in ways not apparent to humans by using digital twins (which are explained later in this section) to make them more welcoming and efficient. For patients, the future could hold innovations like smart prosthetics that enhance mobility and independence. These examples seem attainable with current and near-future technology, and the possibilities extend far beyond our current imagination.

Hardware Improvements: No discussion about emerging technologies can be complete without mentioning the computation power needed for AI. A trend since the early days of computing is Moore's law (Intel 2023), which states that computational power doubles every two years, which still holds true to this day. Other advancements, including the creation of inference chips and ternary quantization, make even faster compute possible. Inference chips are built specifically for running large language models (LLMs) and take advantage of their architecture to do it more efficiently. While quantization is not a hardware improvement, it accomplishes the same improvements. Ternary quantization increases speed by 10 times and reduces memory requirements by a factor of 8, which allows LLMs to become larger and faster (Ma, et al. 2024).

Wearables: The future of technology involves integrating it more closely to our biology and having it do tasks that only humans can currently do. The implementation of wearable technology will allow people to monitor their blood oxygen, heart rate, respiratory rate, and more, which can help them assess when they need preventative care and empower them to attain a healthier lifestyle (Apple n.d.). They could also be used in a medical setting and replace the variety of tools to measure vital signs, such as thermometers and sphygmomanometers, which are cuffs to measure blood pressure. These devices are likely to become more advanced and affordable, while potentially gaining the ability to diagnose less apparent markers, such as anxiety and depression levels (Ahmed, et al. 2023). The future of wearables is not limited to just diagnosis, but they could also provide health management plans tailored to the users' unique data over time. Personalized and data-driven nutritional recommendations, exercise routines, and medication adjustments are all attainable.

Multi-Modal AI: Combining LLMs with vision models and web search APIs allows AI to access and interpret text, image, and video and search on the internet, unlocking many possibilities. These AIs not only output text but can also create images, videos, and audio (Google 2023). Multi-modal AI will allow humans to unleash their creativity. These tools will automate and streamline the creation of standard documents, presentations, reports, and stakeholder communications. By eliminating repetitive tasks through AI, productivity will skyrocket by allowing people to focus on higher-level tasks and creative thinking, while enabling quick iteration to test ideas that would take hours to accomplish unassisted. While these applications will require large and comprehensive datasets, along with expensive hardware and power consumption, these modalities are all technologically within reach.

Spatial Computing: Future systems will integrate the digital and physical worlds, enabling devices to recognize and use 3D spaces. They will employ gesture, voice, and movement to allow users to interact with digital content integrates and overlays the physical environment using advanced sensors for spatial awareness and AI for data interpretation. Challenges with spatial computing include the limitations of the hardware on the headset, which can restrict processing power and user experience, and its reliance on cloud computing, which poses concerns about data privacy and security (PwC 2024).

Digital Twins: Virtual replicas of physical objects, systems, or processes that enable real-time monitoring and simulation, digital twins use data from sensors in the physical world to update and change the replica according to the real-world counterpart (McKinsey & Company 2023). Digital twins are used to predict performance issues, optimize operations, and help inform decision making, although data privacy, integration complexity, and data analysis and interpretation issues still limit applications of this technology. Using AI can help these simulations more accurately reflect real-world data and allow greater customizability. AI can change behaviors to respond to different starting conditions and personalize the experience based on user input and reaction.

Blockchain for Data Provenance: Blockchain provides an immutable and transparent set of records, and this decentralized ledger system ensures that once information is entered, it cannot be altered. By storing access information, it offers a clear and traceable history of data transactions, this enhances regulatory compliance and patient trust for healthcare providers and enables secure data sharing among authorized entities. Consequently, improved collaboration and potentially expedited diagnoses and treatments occur, leading to better patient outcomes and more efficient healthcare delivery. The integration of AI can enhance blockchain security by identifying and neutralizing threats in real-time. Additionally, AI can optimize mining operations by predicting lower energy cost periods, in order to reduce expenses and environmental impact. However, the downsides include high energy consumption for blockchain operations and potential scalability issues as the ledger, A digital transaction record, grows.

3.3.2. Potential CMS Applications

The integration of AI can significantly enhance the experiences of both CMS employees and beneficiaries. For staff, AI-assisted tools such as multi-modal AI, automation, and digital twins can streamline tasks, reduce manual workload, and facilitate more accurate and efficient decision making.

These technologies are at the point where their deployment is both feasible and useful, thanks to advancements in both computational power and data availability.

For beneficiaries, AI can lead to improved healthcare outcomes through personalized medicine and better access to information. By automating routine tasks, AI allows human workers to focus on more nuanced and innovative activities, which can improve job satisfaction and operational efficiency. Overall, the application of AI technologies within CMS operations promises to revolutionize the healthcare landscape, making services more responsive, efficient, and tailored to individual needs. However, as outlined in section 3.2.4 there are privacy concerns associated with these technologies. Therefore, these predictions are meant to inspire future AI initiatives at CMS, with acknowledgement that significant issues must be resolved before applying AI to patient data.

- **Multi-Modal AI for Customer Support:** By integrating multi-modal AI into customer service, CMS can provide instant 24/7 responses to common inquiries, reducing wait times and increasing satisfaction. These AI-driven systems can handle a vast array of questions related to subjects like eligibility criteria and coverage details, which enables human agents to focus on more complex cases. They can also be trained on CMS policies, guidelines, and frequently asked questions to ensure accurate and up-to-date responses. This approach also empowers beneficiaries with immediate access to vital healthcare information, aligning with CMS's commitment to improving healthcare access and information transparency. This is one of the more realistic potential use cases as LLM’s can already perform a lot of these tasks but require more work before being deployed autonomously.
- **Enhanced Patient Information Processing:** Multi-modal AI could significantly improve CMS’s ability to understand complex patient data from multiple sources like text, images, and medical records that could enable more accurate and personalized healthcare services. This also facilitates early intervention strategies by identifying subtle health trends. Improved patient outcomes and more efficient use of resources are all possible with increase AI integration.
- **Wearables for Preventative Care**: Wearables offer a complementary approach to the capabilities of multi-modal AI and by monitoring vital signs like blood oxygen, heart rate, and respiratory rate, CMS could consider supporting the development of machine learning tools that help manage the health of beneficiaries who choose to opt-in. This approach, which provides real-time alerts and detailed health information to medical professionals, could reduce the risk of severe illnesses and promote a healthier lifestyle, aligning with CMS objectives while potentially lowering patient care costs.
- **Documentation/Presentation Development:** Multi-modal AI can transform how documentation and presentations are developed within CMS. Having access to the internet and the capability to understand text, images, and videos allows AI systems to generate content that is informative, visually appealing, and tailored to the specific needs of different users, enhancing its overall effectiveness.

- **Report Generation:** For report generation, multi-modal AI can automate the labor-intensive process of collecting and analyzing information, writing reports, and updating healthcare records. This capability would accelerate the production of these reports and enhance their accuracy. AI can identify trends and anomalies in healthcare data that might not be apparent to human analysts, and these insights can then be added into the reports to support decision making and policy development.
- **Policy Impact Analysis**: By creating digital twins of healthcare ecosystems, CMS can analyze the potential impacts of policy changes and system modifications before they are implemented. Due to the nature of CMS's work, by allowing the impact of decisions to be seen before they are taken, catastrophic risks and consequences can be avoided. Digital twins are also a proven technology employed by healthcare providers and companies, like Phillips and Siemens, for a wide variety of tasks (Katsoulakis, et al. 2024). Hospital management, device design, personalized medicine, etc. are all processes in use, and so employment for CMS use cases may be feasible.

3.4. Recommended AI Design Framework for Working in CMS

Since AI is growing rapidly, and emerging technologies that exist now were science-fiction ten years ago, CMS's approach must consider how to *keep up* with AI. How close do agencies need to stay to the cutting edge to not be left behind in the next few years? How much risk is necessary to spur innovation? There are no correct answers, and making any decision requires thinking about the tradeoffs.

Section 3.2 outlined the principles important to CMS for implementing AI; this section focuses more on the design and technical aspects of AI integration. This section also codifies many of the themes throughout the rest of the Playbook into succinct steps to follow in parallel with the guiding principles of AI.

NASA uses Technology Readiness Levels (TRLs) (Manning 2023) to assess when the risk is low enough to launch a rocket. While the stakes of deploying AI prematurely are not as high, using some of these concepts for AI development and deployment in CMS will help gauge the pace and risks associated with using AI. (Lavin 2022)

Other than risk-avoidance, using TRLs can help an organization avoid technical debt, scope creep, and model misuse/failures, which all have expensive consequences. Machine Learning Technology Readiness Levels (MLTRLs) offer a modified framework that defines a principled process that ensures "robust, reliable, and responsible ML and data systems" (Lavin 2022) and a common framework for people across many teams and organizations to work collaboratively on ML/AI technologies. By having an agreed-upon grading schema for assessing the maturity of an AI technology, and for how or when that technology fits within a product or system, everyone can communicate more effectively and transparently about it.

In essence, by using MLTRLs, a team can minimize the technical debt and risk associated with the delivery of an AI/ML project by helping the development team ask the necessary questions and tackle issues before they become readily apparent. For CMS uses, the MLTRLs are adjusted to emphasize RAI and consist of ten steps, summarized below in Table 8.

Table 8. Adjusted MLTRLs for RAI

Level	Level Overview	Description
Level 0	**Ideation and Research**	This is the brainstorming phase, where the concept is outlined, and initial research and discussions take place.
Level 1	**Begin Coding**	Coding starts here, where the focus is on gathering sample data and creating research-level code to explore the initial idea. Due to the importance of the principles of Responsible AI, regulatory and ethical standards should also be kept in mind before going further.
Level 2	**Proof of Principle**	This phase evaluates the feasibility for deployment, determining what requires further research. This is also when expanded data collection occurs.
Level 3	**Prototype Development**	Most AI/ML projects live within Levels 3-9. The transition to developing a prototype involves creating well-designed, architecturally sound code, which resides at this stage.
Level 4	**Proof of Concept**	This critical phase involves demonstrating the technology's capabilities, addressing ethical considerations, identifying potential issues, and assessing the project's feasibility beyond the project team with stakeholders and product managers. This stage also presents a decision point to continue, pause, or cease the project.
Level 5	**Integration into Larger Systems**	At this stage, the model or algorithm becomes a component of a broader product or application, extending ownership beyond the initial AI/ML team. The data at this point will be close to complete, and pipelines should also be in place. Steps past this point require more resources to be pushed to production.
Level 6	**Application Development**	Focusing on preparing the application for users, this level involves developing production-caliber code, ensuring comprehensive test coverage, and refining APIs.
Level 7	**Integration Testing**	At this point there is a need for both infrastructure and ML engineering expertise so that deployment to production can handle the number of users and quantity of information that will be transferred. This level also emphasizes extensive testing to confirm the alignment of inputs and expected outputs.
Level 8	**Mission Ready**	The project undergoes final evaluations to ensure readiness for deployment, including performance logging, A/B testing, and logging of data distributions and model performance. After review of project requirements, this is the decision point to deploy or wait.
Level 9	**Deployment and Monitoring**	Following deployment, the focus shifts to monitoring the model for performance, maintenance, and adjusting for any data or model drift. These are not strict steps to follow but are a general guideline for a successful AI/ML project lifespan. Following them can help avoid unforeseen issues while getting all stakeholders involved and lending their expertise to see the project to fruition.

These are not strict steps to follow but are a general guideline for a successful AI/ML project lifespan. Following them can help avoid unforeseen issues while getting all stakeholders involved and lending their expertise to see the project to fruition.

4. Implementation and Operation of AI at CMS

Chapter 4 of the playbook navigates the implementation and operation of AI using an approach for AI projects that consists of five major steps (Figure 4): Gathering Requirements & Conducting User Research, Understanding AI Technology & Tools, Engineering AI Models, Evaluating Performance & Determining Metrics, and Governing AI. After an introduction to the roles within an AI project team (Table 9), the sections within Chapter 4 will each focus on one step in the project approach. These sections will provide topics most pertinent to consider for that step, generally including how different roles are involved, what resources can and should be leveraged to learn more, and key action items. The final section in Chapter 4 summarizes the challenges and best practices across all steps.

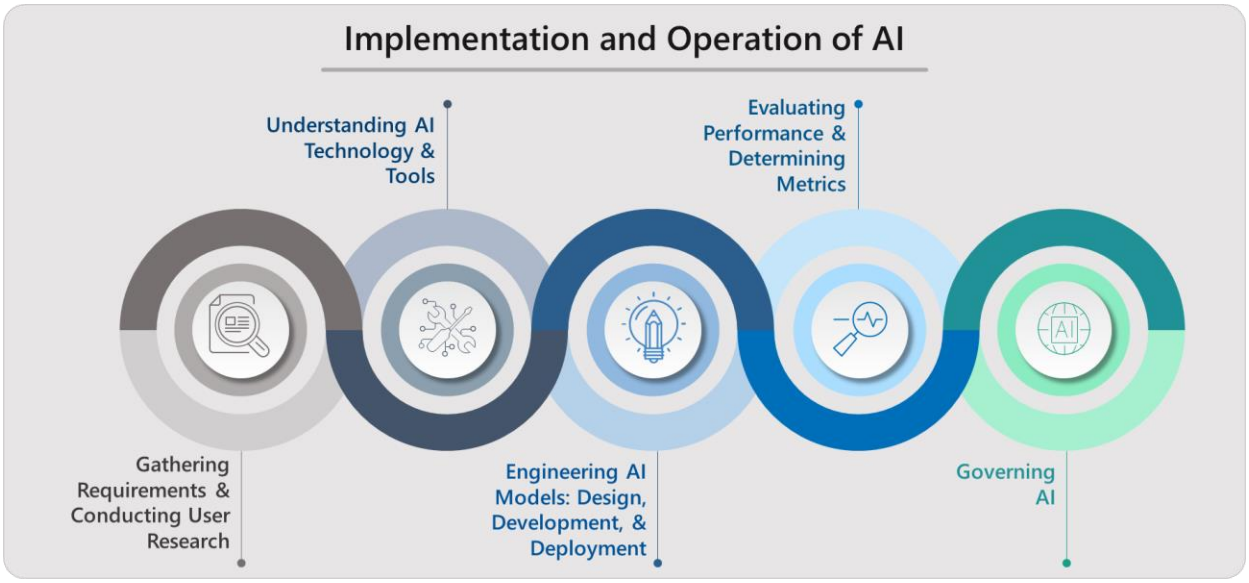


Figure 4. Approach for AI Projects

To pursue the implementation and operation of AI at CMS, AI project team members must understand the roles and responsibilities required to execute an AI project. On a smaller team with a limited budget, it is likely that one person will be responsible for more than one role. For example, a project manager might also act as the Human-Centered Design (HCD) researcher. The descriptions of team members' roles and responsibilities, as well as information for each step of the AI project approach in this section, provide foundational information to jump-start each team member's process regardless of experience level in that role.

Table 9. Roles and Responsibilities of an AI Project Team

Role	Responsibilities	Relevant Steps in an AI Project
Product Manager	Act as the CMS federal employee liaison between stakeholders and the AI project team. They are responsible for defining the vision of the project or product and translating this vision into a prioritized list of features and requirements.	The Product Manager should be involved in every step of the AI Project, but their contributions are especially important during the Research phase when requirements and features are determined (see Section 4.5).

Role	Responsibilities	Relevant Steps in an AI Project
Project Managers	Oversee the operations of the entire AI lifecycle by coordinating multidisciplinary team members and managing team capacity and resources.	Project managers should be involved in every step of the AI Lifecycle, by supporting operations and team members.
Internal Agency Stakeholders	Represent various entities with vested interests in AI projects. They include executives, clients, regulatory bodies, technical system owners, and end users. Other stakeholders could include teams that represent groups, divisions, or components that the solution would impact. Their input and feedback shape project requirements and determine success criteria.	Stakeholders' contributions are especially important during the Research phase when requirements and features are determined (see Section 4.2). As the project progresses, the team may consult each stakeholder as decisions are made that are relevant to the stakeholder's entities and interests.
External Agency Stakeholders	Depending on the scope and impact of the AI project, these stakeholders can include community members, beneficiaries, and health-care providers. Their input and feedback can shape project requirements and determine success criteria.	Stakeholders' contributions are especially important during the Research phase when requirements and features are determined (see Section 4.2). As the project progresses, the team may consult each stakeholder as decisions are made that are relevant to the stakeholder's entities and interests.
Domain Experts	Provide industry-specific knowledge that guides AI projects. They help define project goals, validate model outputs, and ensure alignment with real-world scenarios.	Domain experts' contributions are important throughout all phases. As the project progresses, the team may consult each domain expert as decisions are made that the expert can provide input on.
HCD Researcher	Specialize in understanding the real-world needs, behaviors, and expectations of end users regarding the AI project or AI-based tool. They conduct research and translate user needs into features through human-centered design methods.	The HCD researcher's contributions are especially important during the Research phase when requirements and features are determined (see Section 4.2). As the project progresses, the HCD researcher will continue to conduct user testing to support the team in iterating on the AI project or AI-based tool.
Data Scientists	Are at the forefront of AI projects, responsible for developing and implementing machine learning algorithms to extract insights from data, as well as ensuring that the AI-based tool meets the criteria for RAI.	Data scientists should be involved in every step of the AI Lifecycle, but their contributions are especially important in the phases of Understanding AI Technology and Tools (see Section 4.1), Engineering AI Models (see Section 4.3), Evaluating Performance and Determining Metrics (see Section 4.4), and Governing AI (see Section 4.5).

Role	Responsibilities	Relevant Steps in an AI Project
Development Security Operations (DevSecOps)	Embed security from the earliest stages of the AI Lifecycle, ensuring that security considerations are integral to the AI project's design, development, and operational processes and enhance the system's defense against vulnerabilities and attacks. This role may not be played by a member of the project team, rather it is more likely played by an external team, with whom the project team will have to collaborate.	These team members make critical contributions during the research phase, where security requirements and protocols are established to safeguard the AI systems (see Section 4.1 and Section 4.3.1). When collaborating with an external DevSecOps team, the AI project team should continue to engage the team throughout the AI project lifecycle (see Section 4.3 and Section 4.4).
Developers & System Architects	Translate AI models into functional software solutions while ensuring system architecture. They collaborate to ensure seamless integration into existing systems and develop user interfaces for interaction.	Developers should be involved in every step of the AI Lifecycle, but their contributions are especially important during the Research phase when requirements and features are determined (see Section 4.2). System architects design the overall structure and framework of the system, ensuring scalability, efficiency, and reliability. They should be engaged in the Research phase (see Section 4.2), the Engineering phase (see Section 4.3) and the Evaluation phase (see Section 4.4).
End Users	Are the recipients of the AI-based tool and are responsible for providing feedback to the team on the tool's performance, usability, and features.	End users participate during the Research phase when requirements and features are determined (see Section 4.2). As the project progresses, the team should continue to gather feedback from end users when conducting usability testing, and to inform future iterations of the tool.

4.1. Gathering Requirements & Conducting User Research

Once the AI project team has defined the roles that team members will play, the next step is to deepen the team's understanding of the problem they are setting out to solve by gathering requirements and conducting user research. These insights will help to inform the team on whether they should move forward with an AI project by evaluating the business and users' needs, exploring opportunities for AI to meet that need, and assessing the feasibility of the proposed solution.

This section introduces human-centered design research for AI projects as a critical method for understanding the problem, the people, and the technical requirements to design and develop a human-centered AI project or AI-based tool.

Human-centered design is a problem-solving framework that helps make systems and products more responsive to the people or customers who use them. This framework and the methods for conducting research are described in the "Human-Centered Design: Discovery Stage Field Guide" published by the U.S. General Services Administration (GSA) (GSA n.d.). Implementing human-centered design research

ensures that the AI-based solution meets the needs of the organization and the users, is human-centered, and is responsible (See Section 3.2).

4.1.1. What is Discovery and Evaluative Research?

Conducting research in the context of an AI project can be broken up into two phases: discovery research and evaluative research. During the discovery research phase, the team asks intentional questions to define the problem that it hopes to address, uncovers constraints, and names the requirements needed for an effective solution (User Interviews n.d.).

Discovery research for AI-based tools or software should also seek to identify any stakeholder or user needs related to RAI and consider requirements or concerns around fairness, explainability, security, and privacy early in method selection and design plans. Section 4.1.3 discusses gathering requirements in more detail.

After conducting discovery research and building an initial prototype of the AI-based tool, the team conducts evaluative research to understand if the proposed solution meets stakeholders' and users' needs using methods such as usability testing (User Interviews n.d.). Conducting usability testing early in the software development lifecycle allows the team to gather feedback and implement incremental changes in an iterative fashion, enabling the team to think big and start small. Without the steps of conducting discovery research and then evaluating the tool with stakeholders and users along the way, the team risks incurring technical debt and developing a solution that users do not want or need or that may not meet the technical, operational, and RAI requirements.

4.1.2. How to Conduct Research

Define Team Roles and Responsibilities

The first step in conducting research is defining which members on the AI project team will take on which roles and responsibilities when uncovering requirements. Typically, the research phase is led by a HCD researcher in collaboration with the product manager, lead data scientist and lead developer. Each team member is responsible for uncovering requirements related to their domain.

Conducting research within a complex subject area requires members of the team to develop domain expertise if they are not familiar with the subject. In preparation for interviews with end users and stakeholders, team members will need to be open to developing an understanding of AI models and infrastructure, human-centered AI interactions, and the AI project lifecycle to formulate meaningful questions for their interviewees and make sense of detailed responses that interviewees will provide. Team members can develop their domain expertise by reading about relevant terminology and news from industry leading companies and organizations, watching videos, and drawing on what they learn from their interviews with subject matter experts to lead them to new and relevant concepts.

Create a Research Plan

When preparing to conduct research, a best practice is to create a research plan (User Interviews n.d.). A research plan outlines the goals, objectives, and logistical considerations of the research. It is an evolving document and should be used to support the communication and tracking the team's research tasks. **If the team is unclear on whether an AI-based tool is the appropriate solution for the problem, determining this should be one of the team's research goals.** The following steps are part of a robust research plan:

1. **Identify research goals:** Determine the learning objective from the research and align these goals with broader business aims. During the discovery research phase, the team's research goals will concentrate on expanding the team's understanding of the problem, constraints, and requirements by narrowing its focus. During the evaluative research phase, the research goals will evolve to concentrate on refining the proposed AI-based tool to meet users' needs.
2. **Develop research questions for stakeholders, end users, and technical system owners**: Formulate specific, practical, and actionable questions that your research will address. A fundamental question for both stakeholders and end users could be, "If this AI project or tool was perfect in every way, what would it allow you to accomplish?" It's crucial to include technical system owners and DevSecOps team members in the initial interviews to uncover insights into the practical requirements and constraints that might affect the implementation.
3. **Choose the right research methods:** Select methodologies that best fit the research questions and goals, considering the scope and resources available. The most common methodologies during discovery research are conducting desk research, observations, one-on-one interviews, and focus groups. The most common methodology to use during evaluative research is usability testing. (See Section 4.1.4.).
4. **Plan a study design:** Detail the execution of the chosen methods, including research strategies, timelines, and scheduling. Be mindful that identifying and scheduling interviews with the appropriate end users and stakeholders can take weeks. Plan to ensure the team stays on schedule.
5. **Prepare to share and reshare findings:** Outline how the research team will report and share the insights gained with the rest of the team and stakeholders, ensuring the results are actionable and relevant. The most common way to do this is by creating a slide deck containing the findings and next steps and then presenting the results. It is important that findings are revisited or reshared as requirements evolve or as new information is uncovered.

Conduct the Research

After defining team roles and responsibilities and creating the research plan, it is time for the AI project team to execute the research. Ensure that the team members involved in the research schedule time internally to recap meetings with stakeholders and end users to review insights from the research. Once interviews or other chosen methodologies are complete, set aside time with the team to analyze and synthesize the research findings.

Determine if AI is the Right Solution

Once the team has conducted research, it should revisit determining if an AI-based tool is the correct solution for the problem. By interviewing stakeholders, technical system owners, and end users, the team should have uncovered requirements (business, functional, technical, data, operational, and ethical) that might impact the feasibility of an AI-based solution. AI-based solutions are complex, and sometimes simpler alternatives may be more cost effective. (See Section 4.1.3.)

Analyze & Synthesize Research Findings, Create Deliverables

A popular process for analyzing research findings is Affinity Mapping. This activity can be conducted using an online tool such as Mural. Team members work to group units of feedback transcribed onto digital sticky notes according to common themes so that they can quickly and easily grasp user sentiment. After identifying high-level themes and sentiments, this information can be transitioned into

deliverables such as user/stakeholder personas, journey maps, or service blueprints to make the information easier to digest (See Section 4.1.4). These deliverables can provide context for determining a set of requirements and features for the AI-based tool. While not all project research will lend itself to the deliverables mentioned above, at the very minimum, research findings should be compiled into a research deck and findings should be presented to the team and stakeholders (Krause 2022).

4.1.3. Gathering Requirements During Discovery Research

Conducting initial discovery research should begin to reveal the following requirements (Rosala 2020). Conducting evaluative research later in the process will allow the team to refine the AI project or AI-based tool so that it more closely meets user needs. The following Table 10 organizes the types of requirements that the team should hope to uncover, and which team member is responsible for uncovering it.

Table 10. AI Project Requirements Gathering

Requirement Type	Information to Understand	Who to Ask	AI Project Team Member Responsible for Gathering Information
Business Requirements	What are the business goals and objectives that the product/service/solution must achieve to be considered successful?	AI project stakeholders (See Section 4.1.4)	AI Product Manager (See Section 4)
Functional Requirements	What features and functions are necessary for the end user to interact with the product/service/solution effectively?"	End users of the AI project, tool, or software that your team is proposing to build (See Section 4.1.4)	HCD Researcher (See Section 4)
Technical Requirements	What technical considerations must be addressed to ensure functionality, reliability, performance, and security?	Technical stakeholders, system architects, DevSecOps team members (See Section 4.2)	Product Manager in collaboration with developers, and system architects on the AI project team (See Section 4)
Data Requirements	What data is necessary for training the AI model(s)? What is the quality and quantity of the data? How it will be collected, processed, and maintained?	AI project stakeholders or end users who are familiar with the data that is available (See Section 4.3)	Data scientist (See Section 4)
Operational Requirements	What background operations are essential to maintain the continuous functioning of the product or process over time, ensuring its efficiency, reliability, availability, and security?	AI project stakeholders, technical system owners, and DevSecOps team members responsible for systems adjacent to, or impacted by the proposed AI project, tool, or software	A combination of the AI project team members who are best positioned to understand both technical requirements and affected system owner needs

Requirement Type	Information to Understand	Who to Ask	AI Project Team Member Responsible for Gathering Information
Ethical Requirements	What ethical principles and guidelines should the AI-based system adhere to? How will the system promote fairness, accountability, and transparency, and avoid causing harm or bias? (See Section 3.2.4)	AI project stakeholders, technical system owners, DevSecOps team members, and end users to determine how the system may or may not adhere to ethical principles and guidelines	Product Manager in collaboration with developers, and system architects, and data scientists on the AI project team (See Section 4.5)

The next section will focus on how to identify and engage relevant stakeholders and users and which research methods to use to get effective results. For more information on gathering technical and data requirements, see Section 4.2 and Section 4.3.

4.1.4. Engaging Stakeholders & End Users

Who are Stakeholders & End Users?

For an AI project to succeed, the team must identify and engage stakeholders and end users throughout the project lifecycle (Mortensen 2021). Stakeholders represent various entities with vested interests in the AI project. They include executives, clients, regulatory bodies, and technical system owners, who may also be end users of the resulting AI-based tool. Their input and feedback shape project requirements and determine success criteria. End users are the recipients of the AI-based tool and are responsible for providing feedback to the team on the tool's performance, usability, and features. As they conduct research, the team may find that stakeholders and end users have conflicting wants and needs; the team must be able to uncover these effectively and implement a solution that meets both business (stakeholder) and functional (user) requirements. The following are methods that will support the team in effectively uncovering the information it needs.

Identifying and Prioritizing Stakeholders

To be classified as a stakeholder, a person or group must have some interest or level of influence that can impact the project. Use the following considerations in Table 11 to identify a list of relevant stakeholders.

Table 11. Considerations for Identifying Stakeholders

Stakeholder Type	Considerations for Identifying Stakeholders
Authority over Resources	Identify key decision makers. Look for individuals or groups with power to allocate budget and labor.
Domain Knowledge	Identify SMEs. Look for individuals or groups who have knowledge of technical limitations, similar projects being conducted elsewhere in the organization, or connections to important points of contact.
Leadership or Influence	Identify formal and informal agency leaders who oversee AI programs, vision, or have a wide social reach at CMS. Identify individuals or groups whose programs may be impacted by the AI project.

Stakeholder Type	Considerations for Identifying Stakeholders
Ownership of Technical Systems	Identify owners of tools or systems that may be affected by the AI project's implementation.
Technical Teams	Identify data scientists, engineers, or developers who must implement any changes resulting from the AI project's implementation.

After identifying a list of stakeholders, the AI project team can begin to prioritize these stakeholders (Smith 2000). This can be done on a spreadsheet that lists each stakeholder's name and considers the questions from Table 12, below.

Table 12. Considerations for Prioritizing Stakeholders

Question to Consider	Considerations for Prioritizing Stakeholders
To what extent does this stakeholder have influence over the project's success?	Consider budget, project milestones, and who can advocate for the project at an executive level.
To what extent are this stakeholder's expectations critical to meet?	Consider regulatory bodies or teams that are dependent on the proposed AI-based solution.
Who stands to gain or lose the most from the project?	Stakeholders with a lot to gain from the project's success will be more motivated to support it, while stakeholders who perceive risks will require management to address concerns.
Who can provide essential knowledge, expertise, or resources?	Prioritize stakeholders such as product managers, data scientists, and AI experts who can provide context for data.
What is this stakeholder's priority in relation to other stakeholders?	Teams have limited time, attention, and communication efforts. It is important to consider which stakeholders are most important to engage.

When refining the stakeholder list, consider the stakeholder's **importance**, or whether the project can be successful if this stakeholder's interests and needs are not addressed. Additionally, consider the stakeholder's **influence** or their relative power over and within a project. These factors should be considered side-by-side, with each having equal priority in the decision whether to continue engaging with stakeholders (Smith 2000).

Over the course of the AI project lifecycle, those stakeholders with less impact on, or interest in, the project can be removed from consideration.

Engaging Stakeholders Throughout the Project Lifecycle

Once the team has identified and prioritized its stakeholders, they should continue to engage them throughout the lifecycle of the project. The engagement methods described in Table 13, below, should be repeated more than once, ensuring complete and updated information, priorities, and risk management. The identification and prioritization of stakeholders should also be repeated, as more may be identified as the direction of the project takes shape or through continued user and operations research.

Table 13. Types of Stakeholder Engagement

Engagement Type	Description
Informative Engagement	The intent of this engagement is keeping stakeholders aware of status, concerns, and upcoming events. There are many forms that this engagement can take, including newsletters, Jira updates, and scrum meetings.
Information Gathering	This type of engagement entails approaching stakeholders to gain insight from them on the topics of risk mitigation, points of contact, subject matter expertise, and other key areas. This can take place through focus groups, surveys, and approval votes at board meetings.

Identifying End Users

Just like stakeholder research, end-user research is crucial for AI project success. It helps the team to uncover knowledge gaps and ensures the AI-based tool aligns with user needs (The User Experience Research Field Guide n.d.).

Below, you will see three primary ways to identify user groups in Table 14: by role, by demographic, and by user needs.

Table 14. Identifying User Groups

User Group Type	Description
User Role	Consider the various functions or roles users might have in relation to the AI-based tool. For example, the users of a business intelligence tool might include executives, administrators, and business analysts. Each role may want to engage with the tool differently. Executives might expect a quick, high-level summary on the interface, while business analysts may use filters to gather detailed information.
User Demographics	Consider factors such as age, gender, income, education level, and geographic location. Demographics may influence preferences, behavior, and technology use. For example, younger users might prefer a mobile over desktop interface.
User Needs & Pain Points	User needs and pain points cut across role and demographic and focus on the "why" of the user's interaction with the AI-based tool. These groups are derived based on the intended use of the product and should be defined before design and development begins.

Types of Stakeholder and User Research

Once the team has identified the appropriate stakeholders and end users, the next step is to determine which methods work best to elicit the information needed. There are many ways to conduct stakeholder and user research. Keep in mind that talking to one or two users is always better than talking to zero users. If the team needs to prioritize one kind of research method, user interviews are the most valuable for gathering information from both stakeholders and users. See examples of research methods in the table below.

Table 15. Examples of Stakeholder and User Research Methods

What to Use	What It Is	Why to Use It	When to Use It
Desk Research	Reviewing existing information and data from various sources, including internal agency documentation, artifacts from past or related projects, academic papers, market research, and competitor analysis	Gathers existing insights to inform project direction and design decisions without starting from scratch	Primarily in the discovery phase of the discovery phase to build a foundation of understanding
User Interviews	One-on-one sessions, in person or virtual, with a variety of the service or product's potential end users	Helps the team to understand the user's motivations, feelings, and how they interact with a product or service	In all project phases, both discovery and evaluative, but especially: • Before there is a clear concept for the product or service • Once a model has been developed After launch, to evaluate the final product
Observations	Watching how users interact with the service or product in their natural environment without interference	Provides real-world insights into user behavior and the actual usage context of the service or product	Throughout the project as needed, particularly useful in the discovery phase to gather unbiased data
Usability Testing	Asking users to interact with a model in a specific way and observing the result	Identifies unknown flaws or inefficiencies in product design; counters cognitive bias by proving how users really interact with the product	During the evaluative phase, after initial designs or prototypes are created, to identify design flaws and refine the user experience.
Five-Seconds Testing	Displaying a visual element of the model for a few seconds and asking users about their first impressions	Creates opportunities to improve the visual elements of the product or service, to ensure that they give the correct impression of the product's intended use	At any time that a new graphical or textual layout is added to the model, as evaluative research
Card Sorting	Writing words or phrases on cards and asking users to label and categorize them	Ensures that information architecture is done in a logical, methodical way, or reveals inconsistencies in that architecture	In the early discovery stages of a project, before the information architecture is in place

What to Use	What It Is	Why to Use It	When to Use It
Focus Groups	Leading small groups of users through questions or exercises and gathering feedback on their thoughts, feelings, and reactions	Provides insight into what users want from a system; can be used to discuss interactions that take days or weeks and cannot be directly observe	Throughout the lifecycle of a project, in both discovery and evaluative phases

Operations Research & Engaging Technical System Owners

In addition to identifying and engaging stakeholders and users, teams must also conduct operations research to identify existing dependencies between applications, tools, and business processes that will be impacted by the proposed AI-based tool or software. Technical system owners are stakeholders or end users who are the points of contact for these dependent applications or tools.

This kind of research helps to reduce business uncertainties and improve the coordination between different departments and teams. When conducting operations research, teams should create a diagram mapping the current state of operations as they uncover information from owners of the affected systems. Identify where the process begins, where it ends, where the gaps are, and how affected systems could be impacted by the implementation of the proposed AI-based tool.

Application of Results

After the first iteration of this process, the project team should have a viable list of potential stakeholders and have met with the most highly prioritized stakeholders to begin the process of information gathering. They should understand the role and influence of each stakeholder or stakeholder group for future information gathering and risk mitigation. The team will have a documented understanding of stakeholder needs and interests, translated into a set of requirements, targets, and goals to address those needs and interests baked into the project plan. This understanding, just like the list of stakeholders, should be reviewed and updated frequently.

Additionally, the project team should have key deliverables based on their user research, as described in Section 4.1.5. These deliverables will assist in identifying gaps in the team's knowledge and create a better customer experience by helping the team to understand the motivations behind user behaviors, uncover problems that need solving, and develop relevant solutions. It can also give the team insight into the user experience and identify opportunities to improve it.

Involve Stakeholders and End Users Early

Stakeholder input informs the design and development of the project; the earlier the input can be gathered, the less likely technical debt will be incurred later in the project lifecycle. Technical debt refers to the accumulation of inefficient design or infrastructure within a project. This metaphorical debt incurs interest in the form of decreased productivity and duplication of effort due to having to reimagine the design later in the project.

Therefore, project teams should involve stakeholders and end users in a project from its earliest stages to ensure alignment of objectives, priorities, and expectations. Stakeholder and end user involvement also facilitates a comprehensive understanding of requirements, constraints, and user needs, thus enabling the development of a robust architecture and design. Additionally, early involvement fosters a

sense of ownership and accountability among stakeholders, promoting a collaborative approach to problem solving and decision making throughout the project lifecycle. Ultimately, by integrating stakeholder and end user input early on, teams can proactively address potential issues, minimize technical debt, and deliver products that meet user expectations and business objectives.

4.1.5. Documenting Research Findings

Project teams will apply the results of user and operations research throughout the project lifecycle, from the initial design to testing the final product and making any adjustments. Results can take many forms, as shown in this list of possible deliverables (Interaction Design Foundation n.d.).

Personas are fictional representations and generalizations of a cluster of target users who exhibit similar attitudes, goals, and behaviors in relation to the product. These personas should be based on real data obtained from the user groups they are meant to represent and should be developed in the early stage of the projects and revised throughout design and development.

Information mapping can take multiple forms, but all are visualizations used to build a common understanding of user experiences when interacting with a product or service. Information mapping can be used to show the current state of a product, or to describe the ideal future state as a goal-setting and requirements-gathering mechanism.

A **service blueprint** is a detailed diagram that visually represents the process of delivering a service, highlighting the interactions between those receiving the service (end users), those supporting the implementation of the service, and digital systems. It illustrates the activities visible to the end user and activities that occur behind the scenes. It is primarily used to plan service processes and improve service quality.

Prioritized features, which are derived from all other research results, represent the most critical functions or improvements that directly address user concerns and align with their goals. They are ranked based on their impact on user experience, feasibility of implementation, and business objectives.

4.1.6. Designing Human-Centered AI Interactions

By conducting research to understand the needs of end users, the AI project team is on the path to designing and developing a human-centered AI-based tool. (See Section 3.2.1) Once the team has identified the requirements and prioritized features of the solution, they will have to further consider how the end users will interact with the tool. The following resources are industry standards for human-centered user experience design. If the team determines from user research that its AI-based tool will require a user interface, consider the following resources.

Microsoft's Guidelines for Human-AI Interaction

The Microsoft Guidelines for Human-AI Interaction are widely regarded as the industry standard for teams that design and develop interfaces for AI-based tools (Microsoft n.d.). The resource is available for free online and comes with a workbook to support team collaboration. It also provides design patterns and examples on how to apply the guidelines. These guidelines, when applied to designing and developing an AI-based tool, will help ensure that the team is considering how users might interact with the tool in a variety of contexts.

Web Content Accessibility Guidelines 2.1

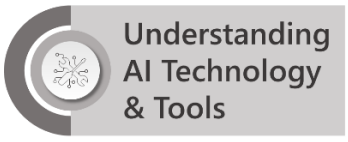
Accessibility is foundational to human-centered AI**.** The design and development of the AI interfaces should follow the Web Content Accessibility Guidelines (WCAG) 2.1, which define how to make web and software content more accessible to people with disabilities (WCAG 2.1 2023). Ensuring software is accessible occurs in the both the design and development phases of the software development lifecycle.

Section 508 of the Rehabilitation Act of 1973 is a U.S. federal law that requires federal agencies to ensure that their electronic and information technology is accessible to people with disabilities. WCAG and Section 508 are often used interchangeably when referring to the need to ensure web and software content is accessible. However, it is important to understand the difference. WCAG 2.1 is a specific set of guidelines that support software and web designers and developers to comply with the Section 508 law.

4.1.7. Key Action Items for Gathering Requirements & Conducting User Research

- Create a research plan: Clearly define the purpose of the research, including what the team plans to learn and how the insights will inform the design process.
- Identify stakeholders and users: Identify and recruit diverse participants who represent both the end users and key stakeholder groups.
- Select appropriate research methods: Choose research methods that best suit the context and objectives of the project, incorporating both discovery and evaluative methods.
- Conduct research sessions: Engage with users and stakeholders using the research methods selected and capture relevant data.
- Analyze findings and synthesize results: Analyze the collected data to identify themes, patterns, and opportunities. Prioritize findings based on their impact on user experience and project goals.
- Determine if your AI-based tool requires a user interface: If it does, refer to resources such as Microsoft's Guidelines for Human-Centered AI Interaction and WCAG 2.1.

4.2. Understanding AI Technology & Tools



AI initiatives rely on a suite of robust tools that are indispensable in crafting algorithms that can learn, predict, and assist in making data-driven decisions. This section will first dive into the key categories of AI tools while explaining their role, significance, and how they streamline the development for AI project teams. It will then provide information on available infrastructure options at CMS for AI.

AI tools enable users the ability to create actionable insights from their raw data. They range from machine learning libraries to AI collaboration spaces that facilitate open source sharing and innovation. Additionally, code repositories work in conjunction with Integrated Development Environments (IDEs) to support development and testing of AI models while also aiding collaboration through version control. Whether for data preparation, model development, or deployment, the AI tools a project team selects should align with the unique demands of the project.

Selecting the right tools for AI projects involves ensuring that these tools not only meet case-specific technical needs, but also align with organization-wide goals and ethical values. As CMS advances its AI capabilities, it must make sure AI is used responsibly. This means choosing tools and toolkits that not only have high performance, but also meet the CMS's six RAI domains (see Section 3.2.4).

4.2.1. Criteria and Guidance for Selecting CMS AI Tools and Resources

This section details criteria and guidance for selecting AI tools and resources with the goal of alignment with CMS's vision and ethical standards. It provides a shortlist of key considerations for infrastructure, explores the methodology for tool selection, and breaks down the essential categories of AI tools, providing insights into their functions and how they integrate into the AI workflow.

Preparing for Infrastructure

Table 16 below outlines several high-level questions for practitioners as you plan your approach to infrastructure.

Table 16. Infrastructure Planning Considerations

Category	Checklist Items
Aligning with Project Objectives	• Does the tool and infrastructure meet the project needs? • Do they support CMS' strategic objectives? • Are the expected outcomes achievable with the selected tools?
Considering Future Needs	• Is your infrastructure scalable to accommodate potential growth? • Can the chosen tools adapt to evolving technological standards? • How will updates and maintenance be managed long-term?
RAI	• Does the tool clearly define its approach to ethical AI? • Are there mechanisms to ensure fairness and transparency? • Does the tool handle bias detection and correction?
Technical Suitability	• Does the infrastructure support the necessary computational and data requirements? • Are the tools compatible with existing CMS systems? • Does the technical setup ensure high availability and reliability?
Security And Compliance	• Does the infrastructure meet all relevant data privacy and security standards? • Is there a clear protocol for compliance with regulatory requirements? • Are there controls in place to handle data breaches or security incidents?
Cost Efficiency	• Is the cost structure of the tools and infrastructure sustainable for the project's budget? • Are there financial efficiencies or benefits to adopting certain models (e.g., cloud-based)?
Ease of Integration	• How easily can the tools and platforms integrate with current workflows and systems? • Are there existing solutions within CMS that can be leveraged? • Does the technology support interoperability?
Environmental Considerations	• How can the technical and computational requirements for the project be aligned with energy-efficient and sustainable infrastructure solutions? • What measures can be taken to minimize the agency's environmental footprint during the deployment and ongoing operation of the project?

Methodology for Tool Selection

Choosing the appropriate AI tools starts with analyzing the project needs, including the array of expected outcomes, the state and size of the data, the computational requirements, and the project timeline. Further, it's suggested that CMS project teams refine their selection process using strategic and ethical considerations. For example, when deciding between machine learning libraries, teams should consider not only their performance but also their support for transparent and explainable AI. To ensure tools align with these values, teams should look for features that support:

- Fairness: Tools that provide mechanisms to detect systemic errors in the decision-making process (Ferrara 2023).
- Transparency and Explainability: Resources that ensure the decisions made by an AI system can be described and reproduced (Baker and Xiang 2023).

Table 17 provides criteria and their details when considering these tools:

Table 17. Key Criteria for Selecting AI Tools

Criteria	Description
Expected Outcomes	Detailing the goals of the AI project, such as improving service, increasing operational efficiency, or enabling informed decision making.
Data Characteristics	Evaluating the state and size of the data, which includes considering data volume, variety, velocity, and veracity to determine computational load and type of tools required.
Computational Requirements	Assessing the processing power needed based on the complexity of tasks, the algorithms to be used, and the overall scope of the project. This involves evaluating whether local processing is sufficient or if cloud-based resources are necessary. Key decisions include choosing between Central Processing Units (CPUs), Graphical Processing Units (GPUs), and Tensor Processing Units (TPUs), which vary in their processing capabilities.
Project Timeline	Setting realistic expectations for the duration of development phases, including model training, testing, and deployment, to ensure timely project completion without compromising quality.

AI Tool Categories and Frameworks

From programming languages tailored for AI projects, to machine learning frameworks that make model building more efficient, and data preprocessing tools that clean and organize data for use, the AI field can be a lot to navigate. This section details the essential categories of AI tools, provides insights into their functions, explains how they fit into the broader AI workflow, and offers examples. Below is an overview of these categories:

- **Programming Languages:** Foundational to AI development, programming languages enable coders to construct algorithms and develop models. They offer the structure needed for developing sophisticated AI applications (Coursera Staff 2024).
- **Frameworks:** Frameworks act as powerful tools in AI that simplify the development and deployment of models, offering pre-built components, algorithms, or functionalities. They range from machine learning, which streamline tasks like regression and classification, to deep learning, which enable the creation of deep neural networks that handle complex data and computation.

Many frameworks inherently support scalability and interoperability, ensuring AI projects can adapt and integrate with various systems and expand as needed (Mungoli 2023).

- **Data Management Tools:** Data management tools encompass a wide range of functionalities, each crucial to the AI workflow, including preprocessing, big data technology, data cataloging, data provenance, and data storage. They prepare data for modeling, process large datasets, and offer secure storage solutions. Additionally, integrating data cataloging and provenance tools ensures well-organized, traceable, and accountable data management (Ghai 2022). By appropriately using these data management tools, project teams can ensure that AI systems are built on a foundation of well-grounded, data-driven AI.
- **Collaboration Tools:** Throughout the AI workflow, collaboration tools facilitate teamwork, knowledge sharing, and project management. These tools span from managing codebases to sharing models and insights within the community. They include version control for tracking changes and enabling simultaneous contributions, as well as documentation tools for maintaining detailed project information.
- **Development Platforms:** These platforms offer integrated environments that streamline the entire AI development process, from creation to deployment. The tools and services provided on these platforms enhance project team productivity, collaboration, and the deployment of AI models. Services like AWS SageMaker (Amazon SageMaker n.d.) or Google Cloud AI (AI and machine learning products n.d.) provide comprehensive tools for building, training, and deploying machine learning models. They feature built-in algorithms, one-click model training, and deployment capabilities, along with seamless access to scalable computing resources.
- **Visualization and UI:** Important for communicating insights and making AI applications user friendly, these tools provide intuitive interfaces for users to interact with, explore, and make sense of complex data and AI models. By providing intuitive interfaces for interaction and exploration, AI project teams can enhance the user experience and promote human-centric AI (Unwin 2020).

As the various tool categories essential for AI development are detailed, it is also helpful to explore some of the specific technologies that can be employed. The following Table 18 lists example tools and platforms that correspond to the categories outlines above, providing a quick reference to aid in the selection process for project teams.

Table 18. Tools and Platforms Supporting AI Development

Category	Example Tools
Programming Languages	Python, R, Scala, etc.
Frameworks	scikit-learn, TensorFlow, PyTorch, Keras, etc.
Data Management Tools	Preprocessing – Pandas, NumPy, etc. Big Data Technology – Apache Hadoop, Spark, etc. Data Storage – Amazon S3, Google Cloud Storage, etc.
Collaboration Tools	Version Control – Git, SVN, etc. Documentation – Markdown, Confluence, etc. Model Sharing and Collaboration Platforms – Hugging Face, GitHub, etc.
Development Platforms	AWS SageMaker, Google Cloud AI, Databricks, etc.
Visualization and UI	Plotly, Dash, Gradio, Streamlit, etc.

To learn more about the AI tools CMS is using, refer to the CMS System Census (CMS 2024) and HHS AI Inventory.

4.2.2. Infrastructure Availability within CMS

Hardware and software provide the foundation for artificial intelligence algorithms learn and operate. The performance and reliability of AI models from development to scaling relies heavily on the underlying infrastructure. Because of this reliance, project teams must tactically approach infrastructure selection for their unique needs. In CMS, AI projects leverage infrastructures that typically include flexibility and computational power for project teams. Such infrastructure is a blend of elements that allow AI optimization, such as:

- **Computing Hardware:** This refers to the physical components essential for AI projects to execute software instructions and process data. Some of the hardware components include Central Processing Units (CPUs), Graphical Processing Units (GPUs), and Tensor Processing Units (TPUs). Each of these offer specific advantages for project teams when powering their AI-based solution. However, there is also a significant cost associated with acquiring and maintaining this computing hardware, which can impact the overall budget of AI projects.
- **Software Infrastructure:** AI functionality requires software infrastructure like machine learning frameworks for developing AI models. Containerization technologies can standardize the operational environment for consistency across model development, testing, and production. Additionally, code repositories can support collaborative development and maintain version control.
- **Data Storage:** When developing AI models, it's important to have access to large amounts of data. There are different types of storage options, like data lakes, data warehouses, and data lakehouses, that each offer unique organization, security controls, and scalability of stored data. CMS utilizes cloud-based data storage solutions like the Integrated Data Repository (IDR), a data warehouse that runs on Snowflake and allows for storage scalability and compute capacity.
- **Network Infrastructure:** AI projects have many disparate elements that need to be connected. Network infrastructures, like Virtual Private Clouds (VPCs) and Content Delivery Networks (CDNs), can provide secure connection across resources.

Deployment of these components is far from straightforward- requiring both in-depth planning and a cross-functional understanding of the project's specific needs.

Monitoring and Managing Resource Consumption

Project teams engaged with AI infrastructure must prioritize the effective management and monitoring of resource consumption. It is essential to deploy tools that facilitate real-time observation and management of resources, spanning hardware, software, data storage, and network infrastructures. Major platforms like Amazon Web Services (AWS) and Google Cloud offer robust monitoring solutions (e.g. AWS CloudWatch and Google Stackdriver, respectively). These tools provide critical insights into resource utilization and system performance, aiding in cost efficiency, capacity planning, performance optimization, as well as benchmarking and benefits analysis. Teams are encouraged to integrate these tools early in the project lifecycle to leverage their full potential in maintaining an efficient operational environment.

Framework for Choosing Infrastructure:

During the early stages of developing an AI project, teams should assess, evaluate, and implement a technical infrastructure that best meets the needs and goals of the project. The landscape of the technical infrastructure for AI projects consists of three primary models: cloud-based, on-premises, and hybrid solutions. Each model offers distinct advantages and limitations, influenced by factors like cost, scalability, performance, and data management. Table 19 presents these three primary technical infrastructure models.

Table 19. Defining Infrastructure Models

Model	Description
Cloud-Based	Model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (Mell and Grance 2011).
On-Premises	Housing hardware and software resources locally on the physical premises.
Hybrid Solutions	Combines cloud-based and on-premises solutions, allowing organizations to keep critical operations local while also taking advantage of the flexibility and scalability of the cloud.

To navigate the complexity of selecting the appropriate infrastructure for an AI project, it's recommended teams evaluate their projects' unique requirements against the capabilities of each infrastructure model. Consider the following decision-making framework:

- Simple Questions to Guide Your Infrastructure Decision
  - What are the data privacy and security requirements?
  - How critical are scalability and flexibility to the project?
  - Does the infrastructure my project needs already exist at CMS?
  - What are the budgetary constraints?
- Use Cases and Examples
  - **Cloud-based:** Ideal for AI projects requiring rapid scaling, such as deploying a chatbot designed to handle varying volumes of user queries.
  - **On-premises:** Best suited for projects with strict data-handling requirements, like processing sensitive health records, where the data must remain within a controlled environment.
  - **Hybrid:** Better for scenarios where a project must balance sensitive data handling with the need for rapid scalability, such as an AI-driven platform analyzing datasets with both public and sensitive information.
- Alignment with CMS's Existing Infrastructure
  - CMS currently has infrastructure capabilities that AI project teams can evaluate and determine if they meet their needs. Leveraging these existing platforms can speed both the development process and obtaining an Authorization to Operate (ATO).

Selecting the right infrastructure is a foundational step in developing AI projects and one that teams should carefully consider. Before making a decision, it is beneficial for project teams to consider factors such as a projects unique needs, alignment with organizational goals, and a responsible approach to technology deployment. By following a structured decision-making framework, teams can navigate the complexities of infrastructure selection, ensuring their projects are built on a solid, scalable, and secure foundation.

Factors Influencing Infrastructure Choice

When deciding on a project's infrastructure, it is important to weigh the multifaceted considerations that will steer the course of an AI initiative. By considering the factors in Table 20, project teams can chart a path that aligns with their operational and computational needs, ensuring that their AI infrastructure is designed for a successful project.

Table 20. Comparing Infrastructure Models Across Various Factors

Factors	Cloud-Based	On-Premises	Hybrid
Cost	Pay-as-you-go model that reduces upfront costs.	Involves initial investment in hardware, facilities, and staffing.	Balanced cost profile with management needed to optimize the costs effectively
Scalability	Provides flexible scalability for varying computational demands.	Scaling is limited by physical capacity, requiring planned expansion.	Offers a mix of on-premises stability and cloud flexibility.
Security and Compliance	Features advanced security measures, though responsibilities are shared with the provider.	Tighter control and security for handling sensitive data.	Tailored security, where sensitive operations can be kept in-house.
Performance	Improving performance, but with potential latency issues or bandwidth constraints.	Typically offers robust performance with configurable latency settings.	Balances on-premises control with the agility of cloud resources.
Operational Expertise	Simplifies technical tasks, reducing the need for detailed IT handling.	Requires a skilled IT workforce for infrastructure management.	Demands workforce that is knowledgeable in managing both cloud and on-premise solutions.
Environmental Impact	Typically, lower direct energy usage due to efficient large-scale data centers that often utilize renewable energy sources.	Higher energy and cooling demands due to less efficient infrastructure; potential for greater direct environmental impact.	Combines the environmental efficiencies of cloud services with the customization and control offered by on-premises setups.

4.2.3. Open Source in AI Development

Embracing open-source libraries and platforms can help teams maintain a transparent and collaborative environment, while also cutting down on costs. Open-source tools allow for the sharing of knowledge and advancements, which ultimately promotes a collective approach to solving AI challenges. This approach not only broadens access to innovation by making cutting-edge technology available to everyone, but it also fosters a culture of continuous improvement and peer review. Additionally, it encourages the development of more secure and robust solutions, as vulnerabilities can be quickly identified and addressed by the community. Through open-source usage, teams can contribute to a greater network of technology built on the principles of transparency and innovation.

In 2023 CMS established the Open Source Program Office (OSPO), the first of its kind in the federal government. OSPO was created to provide enhanced access to the code and software that CMS already has and will continue to develop while decreasing the time to delivery and barriers to capable personnel. This move highlights CMS's dedication to advancing its technology and open-source software (OSS). For more information on OSPO's open-source initiatives and how to engage with these resources, please visit their repositories within the Digital Service at CMS (DSACMS) GitHub.

4.2.4. Key Action Items for Understanding AI Technology & Tools

- Review and align with guiding principles: Ensure all AI Project tools and toolkits align with the guiding principles in section 3.2 to ensure compliance with CMS's mission and ethical standards.
- Determine project requirements: Clearly define and document the specific needs of your project, including data privacy, scalability, performance, and any other special considerations specific to the project.
- Utilize the decision-making framework: Follow the structured decision-making framework provided in this section to make informed choices about AI infrastructure, weighing factors like cost, scalability, security, etc.
- Assess CMS infrastructure and platform offerings: Evaluate CMS's existing infrastructure capabilities to determine which platforms best fit the project's needs and facilitate efficient development.
- Plan for future scalability and interoperability: Ensure that the selected AI tools and infrastructure are not only suitable for the project requirements but are also scalable and interoperable to meet future needs and expansions.

4.3. Engineering AI Models: Design, Development & Deployment

4.3.1. Understanding AI Project Lifecycle

Once the AI project team has started to gather requirements, understand user needs, and determine the technical tools needed for their project as detailed in Sections 4.14.1 and 4.2, it is time for the team to start addressing the AI models they will use. This section will utilize examples from both traditional AI approaches and generative AI tools (GATs) based on LLMs to illustrate various aspects of the AI project lifecycle.

This phase demands a structured and iterative approach to adeptly navigate the inherent complexities of AI projects. Teams must understand the broader operational context, particularly as projects evolve from small-scale pilots to full-scale deployment, shaping the implementation of agency-wide AI programs. Each stage of the AI project lifecycle introduces specific structural frameworks and technological support, promoting both consistency and scalability. Integral to this approach is the consideration of appropriately scaled and interoperable AI as described in Section 3.2.3. This principle underscores the importance of a gradual, adaptive approach aligned with the agency's maturity level and immediate requirements.

As shown in Figure 5, below, the step of Engineering AI Models includes the phases of (1) Model Research & Design, (2) Model Development, and (3) Model Deployment. These phases are integral to the process of building an AI system, focusing specifically on the exploration and refinement of models. The phases ensure that each model is optimally designed, developed, and prepared for integration. This is crucial for realizing desired outcomes and harnessing the full potential of AI technologies.

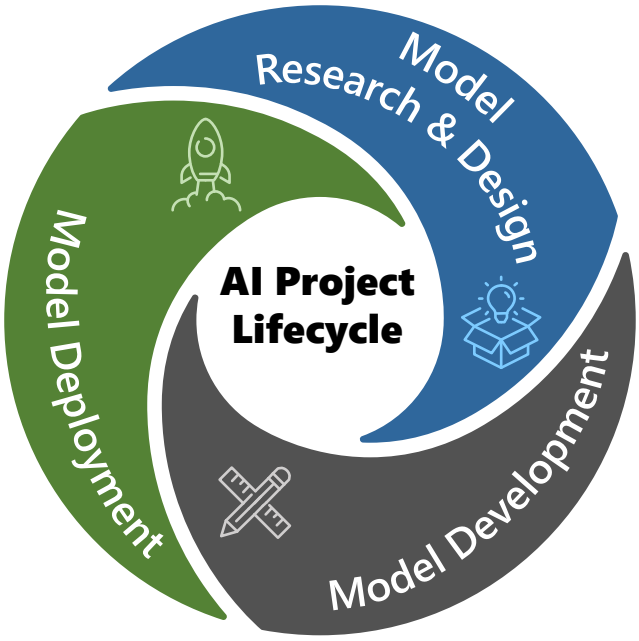


Figure 5. AI Project Lifecycle

4.3.2. Navigating the AI Project Lifecycle

Model Research & Design

The Model Research & Design phase is where the team begins to explore and design models based on the business requirements, user needs, and technical constraints defined during discovery research described in Section 4.1. It is crucial for the team to be aligned on the insights from the research and have a collective understanding of the problem the team is trying to solve to move forward with selecting and designing effective models. To select and design effective models practitioners begin with gathering relevant data and preparing the data for analysis and model development.

Data Gathering and Exploration: Data serves as the foundation of any AI-based solution. Without a clear understanding of the required data and its makeup, AI models cannot effectively utilize it. Engaging in activities such as gathering, cleaning, and analyzing data, along with ensuring data provenance, contributes significantly to a more well-grounded and data-driven model design phase.

Thorough data cleaning minimizes the impact of outliers and irrelevant data points, while good exploratory analysis ensures the selection of features that most accurately represent the underlying patterns in the data. By prioritizing these activities, teams can ensure that their AI models are built on reliable and representative data, aligning with the principle of Well-Grounded and Data-Driven AI.

Particularly when dealing with generative AI/LLMs, additional considerations come into play, such as the diversity of data sources and the risk of bias in training data, especially in language-centric tasks where

nuances and cultural context are paramount (Ramesh 2023), (Arsanjani 2023). Throughout this process, adherence to data ethics and privacy considerations is paramount, ensuring appropriate data access, collection, and usage (Lee, Zankl and Chang 2016).

Data Wrangling and Preparation: Data preparation is often the most challenging and time-consuming phase of the project lifecycle, yet it is crucial for model effectiveness and performance. This step involves transforming raw datasets, including unstructured data, into a structured format suitable for model development (General Services Administration (GSA) n.d.). Key tasks include data cleaning, transformation, feature selection, and data splitting, all aimed at ensuring data quality, consistency, and compatibility with AI algorithms. In Table 21, below, you'll see typical data processing techniques for traditional AI models.

Table 21. Data Preprocessing Techniques for Traditional AI Models

Data Cleaning	Data Transformation	Feature Selection	Data Splitting
Addresses missing values, outliers, and inconsistencies to enhance data quality and readability.	Converts data into a format suitable for ML models like normalizing numerical features, encoding categorical variables.	Identifies and selects relevant features based on their importance and contribution to the model.	Divides dataset into training, validation, and test sets to facilitate model training and evaluation.

Reference: Bite-sized AI

For generative AI/LLMs, preprocessing techniques may include tokenization, handling long-range dependencies, and fine-tuning model architectures to suit specific language generation tasks (Tam 2023), (Large Language Models Explained 2024). Additionally, teams should pay careful attention to ethical considerations and bias mitigation strategies to ensure fair and unbiased model outputs (Arsanjani 2023). In Table 22, below, you'll see typical data processing techniques for Generative AI or LLM models.

Table 22. Data Preprocessing Techniques for Gen AI/LLMs

Tokenization	Long-Range Dependencies	Fine-Tuning
The process of breaking down text into smaller units called tokens that can be fed into LLMs. These tokens could be words, sub-words, or characters depending on the tokenization strategy used.	Enabling the model to capture and understand the context and dependencies between distant tokens effectively. This is crucial for generating coherent and contextually relevant text.	Adjusting the parameters and structures of pre-trained language models to adapt them to specific tasks or domains.

To illustrate the importance of the research & design phase, consider the Office of Human Capital (OHC) Division of Workforce Analytics and Accountability's initiative to enhance work effectiveness by making various raw datasets accessible to employees. This initiative led to the development of the OHC AI Pilot (OHC Time to Hire Calculator n.d.). The objective of the pilot is to develop a user-centered prototype – the Time-To-Hire-Calculator – that will allow a hiring manager to calculate how long it would take to hire a candidate for a specified position and provide alternative paths to reduce the number of days until hire. Following a structured and iterative approach, the solution employed research and design principles to gain insights into existing data, design process flows, and ensure user comprehension and responsible data utilization.

Model Development

The model development phase is a pivotal stage in the AI life cycle, where the conceptualization of AI-based solutions transitions into tangible implementations. This phase focuses on experimenting with data to determine the most suitable model for translating these concepts into actionable implementations. Model development uses an Agile approach to continually train, test, evaluate, and retrain different models, refining them through iterative fine-tuning and incorporation of user feedback (General Services Administration (GSA) n.d.).

Model Selection: Before model development, the project team makes purposeful considerations to determine the most suitable AI model for the business need. They can employ various techniques, such as random train/test split, cross-validation, and bootstrap, to select the best model aligned with the project's purpose and criteria (such as such as performance, robustness, or complexity) (Model Selection n.d.). For GAT/LLMs, teams should consider models capable of handling complex language structures and generating human-like responses, potentially revolutionizing tasks like content generation and natural language understanding (Saltz, The GenAI Life Cycle 2024). Table 23 discussed examples of model selection techniques:

Table 23. Model Selection Techniques

Random Train/Test Split	Cross-Validation	Bootstrap
Splitting the data into train and test sets to select the model exhibiting the best performance on the test set.	Training and evaluating models on multiple resampled train and test sets, ensuring robust performance.	Sampling data points with replacement to select the model with the best performance.

Model Training: This step involves feeding the model sufficient training data to facilitate machine learning. Consistent training significantly improves the model's prediction rate, with the accuracy of the training dataset being critical for model precision (What Is Model Training? n.d.). With GAT/LLMs, training may involve specialized techniques such as leveraging large-scale pre-trained models and fine-tuning them for specific tasks, ensuring adaptability and effectiveness in generating high-quality outputs (Arsanjani 2023), (Saltz, The GenAI Life Cycle 2024).

Testing & Evaluation: Rigorous testing of the developed models on new datasets allows teams to evaluate performance and interpretability. In Table 24, there are several model performance metric examples. While these evaluations commonly apply metrics like precision, accuracy, F1 score, recall, specificity, and Area Under Curve (commonly known as AUC) score, these are not the only possible metrics. Teams should consider other types of evaluation, such as groundedness, coherence (measures related to readability), and bias testing, for assessing model reliability and robustness (Health and Human Services (HHS) n.d.), (What is AI Project Cycle? n.d.), (Saltz, What is the AI Life Cycle? 2023).

Table 24. Model Performance Metrics

Metric	Description
Recall (Sensitivity)	Measures the model's ability to correctly identify positive instances out of all actual positives.
Specificity	Measures the model's ability to correctly identify negative instances out of all actual negatives.
Precision	Measures the accuracy of positive predictions made by the model.
Accuracy	Measures the overall correctness of the model's predictions.
F1 Score	A balanced measure of a model's precision and recall.
Area Under Curve (AUC) Score	Provides an aggregate assessment of a model's performance across various classification thresholds.

In line with the AI Explorers Use Case presented in Section 2.1, the model development phase exemplified by the Multi-Model Interface for Enhanced User Accessibility showcases the innovative application of generative AI-based solutions. Focusing on leveraging LLMs for CMS, the project demonstrated the potential of these models through a proof-of-concept demo. The team aimed to illustrate how a single website could host multiple models tailored to specific datasets, ensuring accessibility and user interaction within CMS' infrastructure. Throughout this phase, the team optimized model performance within hardware constraints, employing previous-generation GPUs and AWS "inferentia" inference chips. Their iterative approach involved integrating model monitoring tools, empowering users to validate responses and build trust in AI-generated content. While formal testing procedures were limited, extensive experimentation and validation ensured the functionality and effectiveness of the models.

Model Deployment

As the final phase of the AI project lifecycle, the model deployment phase represents the culmination of efforts, transitioning from pilot implementation to full-scale integration into operational environments. This phase encompasses various activities, including the initiation of continuous maintenance and monitoring, ensuring the ongoing performance and reliability of the deployed AI model. Note that not all pilots will progress to this stage, and those that do may not undergo every aspect of the deployment process in its entirety. This approach aligns with the Appropriately Scaled and Interoperable AI principle, emphasizing a tailored deployment plan based on each project's maturity level and specific requirements. Furthermore, this phase must incorporate transparent communication of model limitations, potential biases, and the steps taken to address them at every step, fostering trust and accountability among end users and stakeholders (Arsanjani 2023).

Implementation and Scaling: This step requires meticulous planning to ensure the seamless incorporation of the developed AI models into existing infrastructure, aligning with the guidance provided in Section 3.1 of the Playbook (Violino 2021). These efforts must consider compatibility with

existing workflows, scalability, data governance, security, and user training for successful implementation (AI Services Integration with Existing Business Systems 2024).

For AI models, deployment entails configuring the infrastructure to support hosting, scaling, and management of the model, facilitating its smooth operation and integration into existing systems. Deployment may also employ containerization to package the model for production environments, with specialized endpoints enabling scalable hosting and real-time request processing (The Lifecycle of Generative AI – In Simple Steps n.d.).

Following successful deployment, the focus shifts to content generation and delivery, where the model generates and delivers new content to end users through APIs or applications. Input prompts and conditional parameters enable the customization of outputs, while post-generation processing may involve editing or formatting results as necessary. Effective delivery mechanisms ensure the dissemination of AI-generated materials to target audiences via appropriate communication channels (Ramesh 2023).

Monitoring and Maintenance: Starting with the critical aspect of maintaining model security and robustness, continuous improvement is fundamental to AI deployment. Models must evolve and remain effective in dynamic, real-world environments. Continuous monitoring tools play a crucial role in post-deployment operations by tracking latency, error rates, and resource consumption, ensuring the reliable and stable operation of deployed AI models (The Lifecycle of Generative AI – In Simple Steps n.d.). Following this, regular updates and improvements based on performance data, user feedback, and evolving business needs will help maintain model relevance and effectiveness over time (ML Model Maintenance: Best Practices for Ensuring Accurate and Reliable Models 2023), (Appen 2021). In the realm of generative AI, this also includes ongoing monitoring for concept and distribution shifts, feedback-driven optimizations, and updates to safeguard performance in dynamic, real-world environments (The Lifecycle of Generative AI – In Simple Steps n.d.).

Moreover, evaluation of model performance should not be limited to the training and testing phase alone. Before deployment, rigorous evaluation ensures that the model meets predefined criteria for accuracy, reliability, and interoperability. However, evaluation should not stop there. Continuous monitoring and re-evaluation post-deployment are essential to detect any drift in performance, adapt to evolving data patterns and maintain the model's effectiveness over time.

Transitioning to the deployment phase in the Multi-Modal Interface for Enhanced User Accessibility use case, the focus shifted to implementing the developed models within CMS's infrastructure. Despite the absence of formal measures for continuous maintenance, the team remained vigilant in addressing challenges, particularly regarding hardware compatibility and software updates within the AWS environment. The deployment aimed to showcase the capabilities of the models to business leaders, enabling them to envision real-world applications and solutions.

4.3.3. Key Action Items for Engineering AI Models: Design, Development & Deployment

- Define the objective and scope of the AI project (see Section 3.2).
- Determine the infrastructure requirements for developing and deploying AI models (see Section 3.1).
- Gather and preprocess data necessary for model training and evaluation.
- Select and experiment with ML algorithms or techniques based on project requirements.

- Train the chosen model using training data and evaluate its performance using validation data.
- Deploy and integrate models into production environments and test the deployed model thoroughly to ensure reliability, accuracy, and robustness.
- Implement mechanisms for continuous improvement, including retraining models with new data and updating algorithms to incorporate advancements.

4.4. Evaluating Performance & Determining Metrics

"You can only manage what you measure" is a management saying and a good tenet for AI. However, there are many things in AI that need to be managed despite not being straightforwardly measurable.

This section explores key performance indicators (KPIs) to measure and manage AI projects, and how to account for things not currently quantifiable. It also introduces specific considerations for generative AI and explains why it is valuable to remember learnings in impact cases studies.

4.4.1. Key Performance Indicators

What are Key Performance Indicators?

KPIs are quantifiable measures that help track projects to specific goals. Ideally, they should be selected and calibrated to determine which levels are indicative of success or failure after goals have been chosen but before the project has been implemented.

How do KPIs apply to AI generally? While AI algorithms are technical matters, AI-based solutions are multi-faceted and contain technical, business, ethical, and other perspectives. For an AI-based solution to succeed, it needs to do well in each facet.

Who is involved?

Successful creation and implementation of KPIs require perspectives from a wide range of stakeholders. Because KPIs are generally the domain of product managers, it makes sense to have them lead and manage these metrics, but not in a vacuum. While challenges associated with bringing stakeholders together exist, a constant and varied set of perspectives can ensure the right measures drive an AI project. Table 25 shows the most essential stakeholder groups and the kinds of perspectives they may bring to KPIs. Stakeholders for a given project may vary, but this core group is likely to be a constant to minimize the chance that something important gets neglected.

Table 25. List of Stakeholders and the Perspectives They Can Bring to KPIs.

Stakeholder	Perspective
Product	Product teams focus on overall product success, which they generally view through several dimensions. For instance, they may track the overall market through number of competitors and specific product performance through the number of customers who become repeat clients. Product teams are likely the most experienced and adept at using KPIs. Overall, they can help ensure that product performance is tracked properly. They can also share best practices, such as tools to track metrics.

Stakeholder	Perspective
Legal/Compliance	Legal and compliance teams interpret, translate, and provide context to internal and external rules and regulations – a landscape that non-experts may find hard to navigate. Their perspective can help ensure that organizations are following the rules and adhering to regulatory provisions. Legal and compliance teams can be strong contributors to ethical considerations as that is often an important element of their specialization.
Human Resources	Human resources can also help ensure that projects follow organizational rules. Moreover, given their focus on the people within an organization, they can help avoid unnecessary tensions with internal employees. For example, they may ensure that perspectives on a given project encompass all levels at an organization. This may be tracked by the proportion of employees by level who have had a chance to make their voices heard.
Management	Management focuses on organizational priorities. They can help ensure AI/ML align with and contribute to institutional goals. This may be tracked through the number institutional priorities aligned with an effort.
Data Science	Data scientists' perspective is in thinking how data can be valuable. They consider how valid and clean data is, and how that information can be used to drive decisions or predictions. They also consider the performance of models and algorithms, and how they can be measured and improved. Their KPIs could focus on false positive rates over time, for example.
SMEs	SMEs off offer in-depth knowledge in specific areas. They help ensure that all technical work done by the data science teams makes intuitive sense.
Customers	If a project has a customer, their perspective will be the biggest driver of success, and project teams should include this perspective early. This can be done through elements of human-centered design, including surveys. Metrics could seek to capture how many customers have been polled and what their opinions are.
External	External stakeholders include the public and the environment. These are easy stakeholders to miss, and their perspective must be inferred by other stakeholders. This can be accomplished through questions such as "What impact could training this model have on the environment?" Answers may often not directly measurable and will have to be proxied. For example, metrics may focus on energy consumption and amount of pollutants generated.

Keys to KPI Success

Overall, there are general keys to unlock successful KPI development. First, ensure a mix of general and specific KPIs. General KPIs can include number of minutes saved overall, and specific ones can account for how much time was saved on weeknights. Both perspectives can be informative. The broader perspective can reveal if the overall project is succeeding, while the specific one can highlight ways to improve the project going forward.

Relatedly, KPIs should be use-case specific. While there are generalities that apply to almost all projects (e.g., resources saved), each project will need measures that capture its intricacies. Examples could include the number of stage 2 tumors found in lung scans or the proportion of questions answered satisfactorily by a chatbot.

Another key is using KPIs to ensure adherence to organizational data governance policies. Elements that should be tracked include data quality, usage, and protection.

Monitoring and accountability are as important as choosing metrics. What is the point of setting a KPI if it has no impact? Put someone in charge of monitoring. Ideally this person (or group) should have authority to coordinate corrective actions for KPIs going astray. Empower the monitor with tools, such as dashboards – even better when they have automated alerts.

A final key is balancing the costs and benefits of capturing metrics. More is not always better, and too many metrics may be distracting. Some metrics may capture exactly what is needed, but collecting the data may be unachievable. Metrics that require additional user effort may have the unintended consequence of making people less likely to use the tool or service. Consider how often people stay on the phone to answer "a brief survey." It is best to brainstorm KPI ideas, refine them, then select those that are likely to be attainable and impactful.

Model KPIs

Model KPIs are metrics related to model performance. Most frequently they are metrics related to accuracy (e.g., Receiver Operating Curve (ROC Curve)) or to runtime performance (e.g., throughput and latency). This section focuses on several key metrics that need specialized attention.

The metric used to validate how well the model gets things right – what technicians call validation – needs very careful attention. Choosing the wrong metric may mean that the work being done on a project is solving the wrong problem. Take, for instance, a model designed to detect fraud where only 1% of transactions are fraudulent. If the model simply classifies all transactions as non-fraudulent, it would technically achieve 99% accuracy, yet it would fail to identify any fraud. In such cases, focusing on minimizing the false negative rate is crucial, as the cost of missing a fraud instance is typically much higher than that of a false alert. The way to prevent or rectify this scenario is to review published literature and to consult with experts to ensure that the selected metrics make sense for a given problem.

Prevalence is the name of a metric for how frequently what the model aims to find exists in the data. For some tasks, such as finding rare events, the prevalence in the real world is, by definition, very low. Because it is hard to train models on datasets with very low prevalence, training datasets for these use cases often have much higher prevalence than what exists in the real world. The remedy to this depends on the specifics of the problem. Ensure this issue is discussed between data scientists and their teams.

Model drift is unique to AI and needs to be accounted for through KPIs. Whereas traditional software algorithms are static, AI models are dynamic. AI models are trained on a dataset at a specific point in time. However, the data on which the model is performing inference may start to resemble the training dataset less, and AI model performance tends to get worse over time—a concept known as drift (Białek 2023). To account for this, a general rule of thumb is to retrain AI models every six to 12 months. While this point may seem trivial, it is often lost in practice: one cannot account for drift unless it is measured. Ensure KPIs are used to track model performance over time. If third parties are involved, ensure contracts at least give the government access to the data, but preferably ownership of it.

Business KPIs

Business KPIs capture metrics related to efficient use of resources and profitability. The primary objective is to determine if the overall effort is worthwhile. From a practical perspective, AI projects should be

seen through a business lens. KPIs should account for costs, benefits, and risks. Costs can range from cloud compute billing to hours spent labeling data. Benefits can include resources saved, or number of machines that did not need to be replaced due to preventative maintenance. Risks can capture number of unforeseen issues that have come up, or number of times models made mistakes.

Ethical KPIs

Ethical KPIs are metrics that track ethics compliance. These are probably the hardest to manage as there often is no clear-cut way to measure them, but they are no less important than other considerations. One approach is to try to measure by proxy things that are related to ethics. An example would be to measure the emotional intelligence (EQ) for chatbots (Paech 2024) on the assumption that chatbots with high EQs are less likely to display bias or be unethical. Section 3.2.4 includes several examples and useful resources for responsible AI.

Custom KPIs

There are many other project factors that could be measured, depending on the project and specific use case. For example, a custom KPI could be the number of Spanish speaking senior citizens in Alabama that indicated the Medicare website chatbot successfully answered their blood thinner prescription drug coverage question. Projects that try to fix one issue, such as how to better help Spanish speaking citizens in the South through web-based solutions, may need this level of granularity.

4.4.2. Generative AI Considerations

What are Generative AI Considerations?

Amongst AI capabilities, a distinction for generative AI can be made from "traditional" models. Traditional AI has many uses including classifying data, making recommendations, and making predictions. Generative AI is used to create novel and seemingly creative responses based on use input. Generative AI is effective in a number of use cases, such as powering chatbots or creating works of art. However, this technology comes with risks that require special considerations to empower safe and effective usage. This section will briefly discuss these risks.

Who is involved?

All stakeholders who interact with generative AI, and are impacted by it, have a stake in how it is used. Because there are risks without existing safeguards (e.g., hallucination – discussed below), data science and ethics teams need to devote extra attention to ensure users interacting with these tools have safe, productive, and honest interactions.

What is the process?

The process for understanding and addressing considerations specific to Generative AI begins with reviewing the role of foundational models, their characteristics, and how to assess them; then covers issues specific to these models; and finally ends with best practices for managing these models.

Role of Foundation Models

Modern GATs are different from traditional AI/ML chiefly due to their reliance on foundation models – that is, very large generalist models trained on massive datasets that are used for general purposes. Traditional AI/ML models would seek to train a model from scratch or would recalibrate an existing model to a specific use case (a process called transfer learning). To retrain a foundation model to a

specific use case is more complex and expensive. Instead, other techniques are often used to adapt them to a specific problem (e.g., prompting), or the model is used as is (e.g., completions).

Understanding how transparent and reliable foundation models are requires reviewing characteristics of the industry and of the technology behind them. When it comes to accessing LLMs, there are two main ways to get to foundation models: open source and proprietary. The general theme is that proprietary models are more opaque, but open-source models are no cure-all (McKinzie 2024).

While companies releasing foundation models have an incentive to create high-quality, unbiased models, their results may not always meet expectations – after all, it is not uncommon to see headlines about models misbehaving (Vigliarolo 2024). Those who opt for proprietary models cannot afford to be any less vigilant than those who choose open source.

Some foundational models are considered better than others based on either performance or innovation. There are several benchmarks on a variety of topics that have become standards. For example, some test how well models perform on grade-school-level math problems, others test how they perform questions that require graduate level domain understanding. Table 26 summarizes several prominent LLM benchmarks to showcase the kinds of assessments the AIML community relies on to differentiate LLM models. However, performing well on these benchmarks is not conclusive. Rarely does a model lead in every category, and some argue that the tests can be gamed. Moreover, performing well on these tests does not guarantee success. For instance, the Falcon family of models performed well on benchmarks, but that performance failed to translate to humans' evaluation of the models (Sanseviero 2024).

Table 26. Notable LLM Benchmarks

Practice	Description
Winogrande	Tests for commonsense reasoning with a collection of over 44,000 fill-in-the-blank questions with binary answers.
GSM8K	Assesses multi-step mathematical reasoning using over 8000 grade school math word problems requiring between two and eight steps to solve using a series of basic arithmetic operators.
HumanEval	Uses 164 programming problems to assess how well models write code.
EQ Bench	Assesses emotional intelligence using 60 dialogue examples for which models have to ascertain the intensity of emotional states.
GPQA	Assesses how well models can answer difficult questions requiring domain expertise using 198 multiple-choice questions crafted by graduate-level domain experts.

Generative AI Model Issues

Generative AI models have three issues that need special attention: hallucinations, groundedness, and emergent capabilities.

Hallucination is the term coined for models confidently providing incorrect answers – confabulation would be a more accurate term. This is probably the first-time humanity has had to deal with a tool that makes things up. There are technical approaches to combating this, such as reading comprehension and generation (RAG) where the model retrieves and incorporates external data, given to it, to verify its answers. While promising, this approach cannot incorporate all the information it will ever need.

Groundedness is much harder to track. Did the LLM respond positively because it has a bias for "yes" in general? Would it have answered differently if the question were asked a different way? How would the response have varied given a larger back and forth with the model? Currently, there are no agreed upon ways to assess these things, let alone conclude on their answers.

Emergent capabilities refers to the concept that as generative AI foundation models get bigger, they are able to do things for which they were not explicitly trained (Jason 2022). This is not universally accepted. Some argue that these emergent capabilities were in the training data all along but were unnoticed because the training datasets are so large (Lu 2023). Regardless, the potential for emergent capabilities only elevates the level of risk. In the past, something designed for one purpose may have been used for another, but never before have new capabilities sprouted out of nowhere from existing tools.

Managing Generative AI Best Practices

The Department of Commerce, implementing the Executive Order on the Safe, Secure, and Trustworthy Development of AI, is publishing guidance intended to improve the safety, security, and trustworthiness of AI systems – with a notable focus on generative AI. These include, as of April 2024, four draft publications and a challenge series from NIST, and a request for public comment from the U.S. Patent and Trademark Office (Office of Public Affairs 2024). Readers should refer to the most recent federal guidance on AI as they continue to be released, especially for evolving generative AI developments.

From CMS, the document Generative AI Tools (GAT) CMS Uses and Risks (Requires CMS Cloud access), offers general best practices for managing generative AI. Table 27 highlights and summarizes several of these. The CMS Confluence page Evaluating Trustworthy LLM and GAT-Driven Projects at CMS also lists GAT-specific considerations that will be helpful in guiding responsible practices and evaluation of generative AI. These are summarized in Table 28 (in section 4.5.2). Please refer to these linked sources for a more comprehensive discussion and to find additional supporting resources.

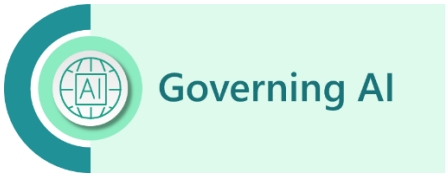
Table 27. Highlights and Summaries of Major Points from the Document GAT CMS Uses and Risks

Practice	Description
Understand fundamentals	Stakeholders and users should have at least a baseline understanding of how models operate, including their limits of understanding. If possible, this understanding should also promulgate to those impacted by the tool without directly interacting with it.
Assess suitability and impact	AI models are trained statically but operate in dynamic environments, so the need to ensure suitability is constant.
Assess data and inputs	PII and other sensitive data must be protected and secured to ensure compliance and protection of information.
Implement human guardrails	Generative AI models have distinct issues, such as hallucinations, that cannot yet be eliminated algorithmically. Implementing human oversight is necessary to prevent issues, such as models spreading misinformation.
Communicate with stakeholders	Communication is necessary to ensure alignment with organizational goals and maintain project momentum.

4.4.3. Key Action Items for Evaluating Performance & Determining Metrics

- Select KPIs at project inception and monitor them throughout. Empower those in charge of monitoring to enact changes to mitigate issues.
- Monitor generative AI tools for hallucinations and assess for groundedness.

4.5. Governing AI



While Chapter 4 of the Playbook thus far has focused on the research, design, and implementation of AI, this section describes how AI governance can support alignment of AI efforts with the guiding AI principles.

AI governance is a cross-cutting function that directs the development and use of AI across the agency. To govern AI is to inform and enforce the agency's values through clear and established policies and procedures; however, to try to list all of these within this playbook would be unproductive. Over time, the collection of governing controls will continue to be defined and mature for an organization. What this section introduces instead is an AI governance framework (introduced in Figure 6) that advocates for collaboration from different roles, principles to guide methodologies, and standard review processes. These elements operate in concert to align AI efforts with the Playbook's guiding principles for AI.

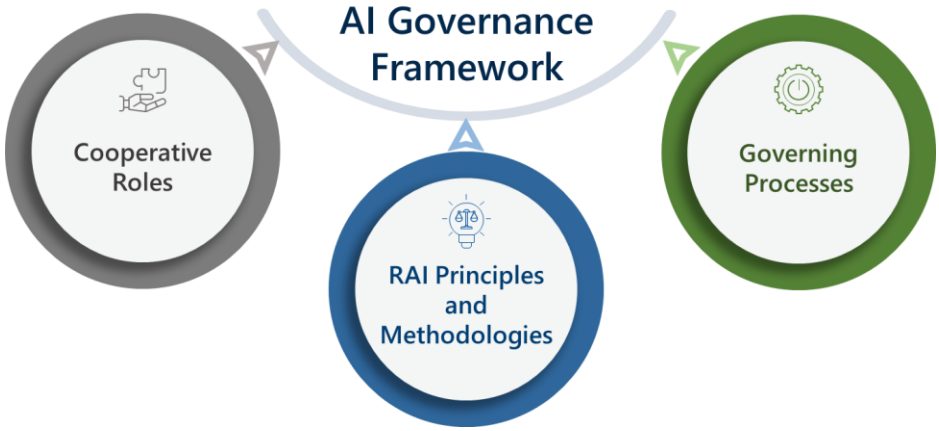


Figure 6. AI Governance Framework

Similar to varying levels of maturity for appropriately scaled and interoperable AI (see Section 3.2.3), AI governance can address different levels of implementation. Organizational AI governance can be perceived through environmental, organizational, and AI system layers which coexist to cover different requirements and practices (Mäntymäki, et al. 2023). The examples within this playbook's AI governance framework align most with the organizational and AI system layers.

4.5.1. Interplay of Different Roles for AI Governance

AI governance cannot be accomplished through any individual effort. An effective governance approach requires oversight and cooperative involvement from many different roles to be valuable and successful. The values of stakeholders, end users, and domain experts are the drivers behind the guiding RAI principles.

These roles are indirectly involved in the establishment of these principles and may participate in evaluation and monitoring through feedback.

DevSecOps and other regulatory or executive stakeholders, such as the HHS CAIO (Young 2024), AI Governance Boards, and various CMS components, are involved in establishing the policies, procedures, and controls for different domains across the governance principles (e.g., policies for data management, cybersecurity, open source). They often maintain oversight of or hold decisive responsibilities within major review processes.

Product managers and project managers aid in the communication, coordination, and enforcement of governance methodologies and processes for their teams. These teams – typically consisting of HCD researchers, data scientists, and developers – follow and contribute to governance practices throughout their individual workstreams.

Ideally, all roles through their involvement and feedback will help the agency organically improve their AI governance framework into one that is comprehensive and aligned with their existing responsibilities.

4.5.2. Responsible AI Principles and Methodology

Responsible AI Principles

An AI governance framework requires principles that can serve as an anchor guiding the human-centric approach to an organization's overall AI adoption and operation. These principles should guide the identification of risks, measurement of risk severity and tradeoffs, and development of mitigation strategies (e.g., policies, tools, techniques). The governance framework in this section derives these principles from the Playbook's guiding RAI principle (see Section 3.2.4). While adopting AI, teams should commit to advancing the principles described across the six RAI domains.

> To serve **fairness and impartiality** as an RAI governance principle, CMS can aim to advance and integrate of health equity in our AI efforts by minimizing the impacts of biases and promoting accessibility.
>
> To serve **transparency and explainability** as an RAI governance principle, CMS can ensure awareness and buy-in from stakeholders to build trust and ease integration to truly maximize AI capabilities.
>
> To serve **accountability and compliance** as an RAI governance principle, CMS can accept the responsibilities that accompany AI implementation as a federal agent and shall adhere to these responsibilities.
>
> To serve **safety and security** as an RAI governance principle, CMS can always protect against, rather than foster, harm that may target or arise from AI vulnerabilities.
>
> To serve **privacy** as an RAI governance principle, CMS can respect both legal and moral privacy rights and aims to preserve this in AI applications.
>
> To serve **reliability and robustness** as an RAI governance principle, CMS can employ proper means to ensure teams effectively progressing AI capabilities with accurate and reliable results.

Applying the RAI Principles

Once the guiding RAI principles have been established, governance tasks can be embedded into procedures and assessments throughout an AI project. Operationalizing AI governance does not simply occur in a single step or all at one time. Incorporating phases from the AI project steps serves as a starting point for the identification of risks and requirements that are much more practical for teams and review boards to track.

Different use cases and AI capabilities will have a unique set of needs and considerations. The HHS Trustworthy AI Playbook (HHS Trustworthy AI Playbook 2021), which these RAI principles nearly replicate, is a great resource for general considerations and guidance on applying the principles to an AI project. Table 28 provides a use case example where RAI principles have solicited additional areas of RAI consideration for LLM and GAT risks. These considerations are discussed in more depth in the AI Explorer's CMS-enterprise Confluence, where they may be dynamically updated as guidance evolves over time.

Table 28. Responsible AI Principles Use Case: LLM/GAT-specific Considerations

RAI Principle	LLM/GAT-specific Considerations
Fairness and Impartiality	• Managing **bias and representation** issues from training data and user biases to generative content with inclusivity and well-represented perspectives. • Promoting **equitable use** by accommodating varied language capabilities and accessibility features like voice-to-text and text-to-voice applications.
Transparency and Explainability	• Guiding **appropriate use** by sharing user instructions for effective system use; intended purposes for the AI; strengths, weaknesses, and limitations of the models; terms and conditions associated with product use; and explanations of overreliance risks. • Sharing **model information** about sources, evaluation, modifications, data, and open-source availability. • Steering the **perception of AI-generated content** by marking provenance (e.g., watermarks or labels) while carefully considering the impact and impression that might be derived by users and stakeholders.
Accountability and Compliance	• Managing **supply chain vulnerabilities** that may pertain to provenance, packages, models and model data, and protected data. • Understanding the early **legal landscape** around AI and generative AI, including review and enforcement of terms and conditions for use of GATs.
Safety and Security	• Securing against risks from **user prompts**, including prompt injection and implications of in-context learning. • Recognizing and moderating **harmful output**, which can include representation and toxicity harms; misinformation and integrity harms; information and safety harms; and malicious use. • Managing **access and authorization** pertaining to model access implications and excessive agency (heightened risk from system autonomy without human confirmation).
Privacy	• Recognizing risks of **data disclosure** from sensitive training data, inversion attacks, and prompt inputs. • Communicating **individual privacy rights** to inform users about use of their data and level of control, while also considering implications of public data.
Reliability and Robustness	• Validating **groundedness** of models by identifying its commitment to its own ground-truth "reality" and effectively citing sources. • Enhancing the **coherence** of generated outputs by improving context, context window lengths, and role-based evaluation. • Proactively monitoring unique model **vulnerabilities** such as model collapse and model staleness.

Identification and measurement of risks and considerations will guide subsequent governance tasks throughout a project. To show this, Table 29 maps example governance tasks (e.g., T40) to the primary RAI principle they contribute to and the AI model engineering lifecycle phase (see Section 4.3) during which they are most likely to be addressed. Notice the progression from assessment to design and then to monitoring. Monitoring encompasses evaluation strategies across all RAI principles. It is vital for AI governance to enforce monitoring and evaluation both for current and emerging risks. This methodology promotes adaptability of mitigation strategies. Appendix E provides a fuller version of Table 29 with the names of all tasks listed in "Additional", and all can be cross-referenced from the original source, Artificial Intelligence Governance and Auditing (AIGA) Governance Framework (Mäntymäki, et al. 2023).

Table 29. AI Governance Tasks Mapped to RAI Principles and AI Model Engineering Lifecycle

	Research and Design	Model Development	Model Deployment
Fairness and Impartiality	Example: T40. AI system harms and impacts pre-assessment Additional: T32, T41, T43, T44	Example: T45. AI system impact metrics design Additional: T46	Example: T47. AI system impact monitoring Additional: T48
Transparency and Explainability	Example: T49. Transparency, explainability and contestability (TEC) expectation canvassing Additional: T3, T19, T44, T50	Example: T51. TEC monitoring design	Example: T53. TEC health checks Additional: T52
Accountability and Compliance	Example: T31. Data sourcing Additional: T17, T23, T54, T55, T56, T60, T61, T62	Example: T63. Compliance monitoring design Additional: T64, T65	Example: T28. Algorithm version control Additional: T66, T67
Safety and Security	Example: T20. Algorithm technical environment design Additional: T5, T21, T40, T42, T44	Example: T26. Algorithm verification and validation Additional: T27, T59	Example: T29. Algorithm performance monitoring Additional: T30
Privacy	Example: T44. AI system impact minimization Additional: T4, T40	Example: T37. Data health check design Additional: T57	Example: T39. Data health checks
Reliability and Robustness	Example: T22. Algorithm operational metrics design Additional: T18, T21, T23, T33, T34, T40, T44	Example: T24. Algorithm performance monitoring design Additional: T25, T26, T27, T35, T36, T37, T58	Example: T38. Data quality monitoring Additional: T28, T39

4.5.3. Example AI Review Processes

When implementing AI governance across an AI project, it is important to have umbrella processes that can provide avenues for training, review, feedback, and approvals from delegated parties. Figure 7 illustrates an example process for RAI review.

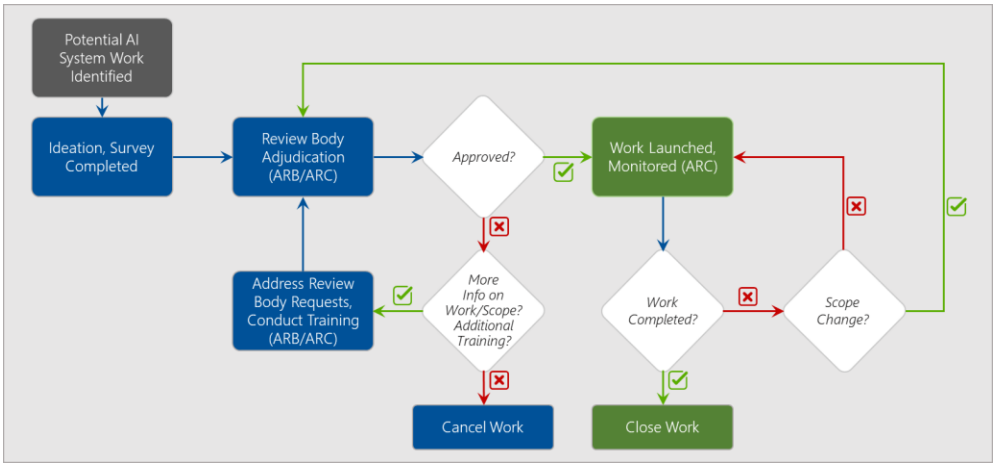


Figure 7. Responsible AI Review Process

In this sample process, review bodies such as AI review boards (ARBs) or AI review committees (ARCs) are situated to review, inform/guide, adjudicate, and help monitor potential AI system work. These reviews often have steps to elicit and answer critical questions about results, checks, and balances of AI-based systems that are pertinent to governance responsibilities or specific governance tasks like those in Figure 7 (above) (Noblis 2023). The simple interactive structure of this sample allows teams to easily modify the steps and entities with their own subject matter and key players. Day-to-day nuances and ongoing refinement of responsibilities in the realm of AI reflect the need for flexibility. The Generative AI Tools (GAT) CMS Uses and Risks research spotlight (link only accessible on CMS network/VPN) provides examples where different use cases and risks can be associated with relevant best practices and mitigation strategies (AI Explorers 2023). These GAT examples underpin areas where RAI review processes or individual steps may be administered to impose them.

CMS AI Review Resource Contacts

For audiences searching for support on maintaining compliance with evolving governance processes, the following parties can help guide product managers and project teams to the right information or contacts. Startup AI programs may connect with the OIT at CMS to help identify and roadmap review process iterations that may be most appropriate for their use case.

- Technical Reference Architecture: TRA team (tra-admin@cms.hhs.gov)
- IT Strategy and System Architecture: Technical Review Board (TRB) (CMS-TRB@cms.hhs.gov)
- Data Privacy and Protection Policies: ISPG (Privacy@cms.hhs.gov or Security@cms.hhs.gov)
- Open Source: Digital Service at CMS (OpenSource@cms.hhs.gov)
- Target Life Cycle: CMS IT Governance (IT_Governance@cms.hhs.gov)
- IT Governance CMS Slack Channel: #cms-it-governance
- AI Community CMS Slack Channel: #ai_community
- OIT's AI Explorer team (ai@cms.hhs.gov) can help find the right people for any of your AI needs.

4.5.4. Procurement Processes

Standing up new projects to build out the necessary infrastructure, systems, products, tools, or even algorithms that collectively drive AI adoption is not a realistic course of action for the agency. Most individuals do not directly engage in procurement activities, however, may still provide support to the contracting officers (COs), contracting officer's representatives (CORs), and Office of Acquisition and Grants Management (OAGM) in understanding AI-related needs. This subsection describes a high-level technical evaluation process which encourages tenets from the guiding AI principles as it demonstrates key considerations to be made for acquisition of AI-related procurements. Those searching for information or assistance with the agency's formal acquisition process should refer to external resources such as the IT Procurement and Acquisition Toolkit (CMS Connect).

The conceptual procurement evaluation process begins with assessing organizational readiness, translating findings to program requirements, assessing alternatives based on several generic and AI-specific conditions, then reinforming the procurement cycle with post-implementation retrospection. Figure 8 depicts this process.

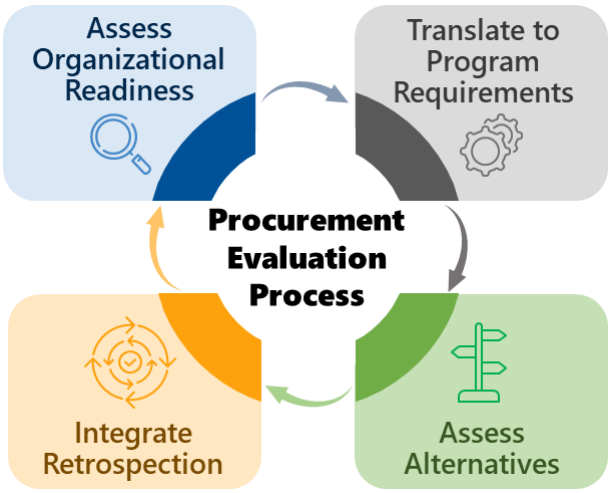


Figure 8. High-level Procurement Technical Evaluation Process

Assess Organizational Readiness

The first phase of the procurement process is an assessment of organizational readiness. Organizational readiness is a measure of the maturity of the agency's own AI goals, and related processes and technologies, such as having:

- An AI steering committee, teams, or individuals assigned to lead AI initiatives
- Technology platforms and digital foundations that lend to easy data access and AI model building
- The necessary technical, data science, and management skills for AI
- RAI principles that are enforced by governance processes
- Operational, reputational, and technological risk management (i.e., standards and processes for identifying, assessing, and mitigating risks associated with the deployment of AI)
- AI use-case pilots

- Sustainability and environmental impact targets (i.e., accounting for energy consuming and carbon footprint of AI systems)
- Budget allocated for AI and IT efforts

Procurement officials will want support from a wide range of stakeholders, SMEs, and enterprise support groups (e.g., IT, legal, and business operations) to understand the gap between the agency's current and desired capabilities. This information will offer insight into which investments will have the most impact.

Translate to Program Requirements

After identifying the strengths and gaps in the agency's current AI readiness, the gaps should be translated to program requirements. Different skills will be needed for novel product development, organizational transformation, and operating and developing models at scale. Each requires a different approach to defining expectations that can be put into procurement contracts and different skill sets to effectively execute.

At this phase, procurement officials may work with individual CMS components, product managers, and even AI project teams to identify the specific outcomes to target and which types of procurement (e.g., skills, approaches, tools) will most closely align with the current and target maturity levels of the organization (consider the Appropriately Scaled and Interoperable AI guiding principle). The human-centered design and discovery research practices from Section 4.1 and research into CMS tools and infrastructure for AI from Section 4.2 will offer insight into procurement types, such as technology platforms, analytical models, intellectual property, or R&D services.

AI-related procurement opportunities should be prioritized for the needs and expectations of diverse ranges of stakeholders, with special attention for those supported by RAI principles, such as tools for evaluating ethical use, fairness, or AI explainability.

Assess Alternatives

Acquisition teams will conduct market research for the items/services identified and justified as clear target capabilities warranting investment. Market research support often comes from program office staff and can also be requested from the GSA (CMS 2024). When assessing AI tool or capability alternatives, teams should consider the toolkit criteria and guidance from Section 4.2, in addition to the following (GSA Artificial Intelligence Center of Excellence (AI CoE). n.d.):

- Compatibility, interoperability, and integration: IT and AI-based tools/services should be compatible with existing and planned future systems, and capable of integration without significant overhauls. The agency should consider open standards and APIs to facilitate interoperability for AI-based tools.
- Scalability and performance: Procurements should at least be capable of handling the expected workload, and ideally could be scalable for future growth.
- Cost-effectiveness: The agency should seek the most cost-effective solution that considers both upfront costs and long-term operational expenses, while not compromising on quality and compliance.
- Market conditions: The rapidly evolving AI market continues to see new startups and approaches, encouraging CMS to adopt an agenda that favors defined outcomes over short-term

AI projects or use cases to allow for shifts as market conditions change and new capabilities become available.
- Compliance and risks: The agency must ensure that AI providers comply with federal data protection standards, such as the Federal Information Security Management Act (FISMA) and privacy laws like the Privacy Act of 1974, among other security, privacy, legal, and regulatory considerations.
- Vendor reliability and support: The agency must also assess vendor reliability based on metric such as their financial stability, track record, level of support, and training they offer.

Integrate Retrospection

Following the implementation of the acquired tools and capabilities, the agency should evaluate the effectiveness and return on investment of the procurement, as well as retrospect the process that led to its acquisition. Retrospectives should include the SMEs and business contributors from the procurement process, as well as any direct users who were part of the procurement process. The team's retrospectives should aim to identify and evaluate the successes, challenges, and potential improvements from the procurement in question (GSA Artificial Intelligence Center of Excellence (AI CoE). n.d.). Teams should use this process to consider the following questions:

- Did we inaccurately measure the organizations readiness? Did we not have the right prerequisite capabilities, or perhaps did we duplicate efforts?
- Was this outcome truly useful and contributive to the agency's AI maturity, knowledge, or capabilities? In what ways did this innovation ensure longevity and relevance?
- Were there risks that we overlooked during the assessment of alternatives?
- Did we provide adequate acquisition support for development teams?
- What key clauses and language worked and what language caused problems, either in communication with AI project teams and provider-contract terms, or the translation between them?

CMS Procurement Resources

Procurement can be a tricky field to navigate, especially for those less experienced in surveying the market or in designing for agency-wide growth. The Federal Acquisition Regulation (FAR) and GSA's Acquisition Gateway are great external resources to learn more about federal procurement and acquisition. Within CMS, the IT Procurement and Acquisition Toolkit on ServiceNow (CMS Connect) can provide helpful resources to employees and contractors alike. Additionally, CMS federal employees may connect with the IT Acquisition Community at CMS through the community's private Slack Channel for tips, templates, and best practices, such as from CMS's IT Contracting Cookbook (permissions-restricted CMS Confluence site).

4.5.5. Key Action Items for Governing AI

- Leverage the RAI principles to identify risks, measure risk severity and tradeoffs, and develop strategies for mitigation, evaluation, monitoring, and adapting to changing circumstances or criteria.
- Engage with compliance support teams across CMS to inform governance practices and request formal review of use cases.
- Engage with CMS IT and procurement communities to discuss or share AI experiences, challenges, and suggestions so that CMS can continue to collaboratively build its AI maturity.

4.6. Implementing and Documenting Best Practices

Section 4 of the playbook outlines a conceptual approach for AI projects, detailing the steps for implementing and operating AI systems. Implementing AI poses challenges, particularly for teams with limited experience. However, pursuing AI projects within the agency enables collective learning from best practices, enhances the agency's AI capabilities, and will ultimately improve the services the agency can deliver to the public.

This section provides an overview of common challenges teams may face as they implement the overarching AI Project steps and best practices recommended in this document to address those issues. Additionally, it includes a template to guide teams in documenting their AI projects as an AI case study. Documenting these projects enables teams to reflect on their experiences and share their insights with the CMS AI community, fostering collective learning throughout the agency.

4.6.1. Common Challenges and Best Practices

The following Table 30 is a summary of challenges and pitfalls AI project teams may encounter within each step of this playbook's approach for AI projects and their corresponding best practices.

Table 30. AI Project Challenges and Best Practices

AI Project Step	Description	Common Challenges	Best Practices
Gathering Requirements & Conducting User Research	A successful AI project requires a deep comprehension of the underlying business problem. During this stage, the team must gather information on business requirements, technical constraints, and user needs. The team should aim to effectively understand the business problem and identify the people who need to be involved over the course of the project.	The project lacks broad and continuous stakeholder involvement resulting in the misalignment of business requirements and user needs. This could potentially lead to low adoption of the proposed AI-based tool and project failure.	Identify business stakeholders, technical stakeholders, system architects, DevSecOps team members, technical systems owners, and end users early in the project. Create a plan for continuously engaging stakeholders and users for feedback and inputs as the project progresses and requirements inevitably evolve.
Understanding AI Technology & Tools	The team must carefully evaluate and select AI tools and technologies that align with the specific requirements and goals of their project.	The team may select AI tools that are incompatible with existing systems or do not meet the specific needs of their projects. This can lead to inefficiencies or failures in project outcomes. Teams might struggle with managing the complexity of various AI technologies, which could hinder the effective integration and utilization of these tools.	The team should conduct a comprehensive needs analysis that includes understanding requirements such as data characteristics, computational needs, and expected outcomes (i.e., improving service, increasing operational efficiency, or enabling informed decision making). Teams should also establish criteria for tool selection that prioritize transparency, fairness, and compliance with CMS's ethical standards, ensuring the tools align with both technical and organizational goals.

AI Project Step	Description	Common Challenges	Best Practices
Engineering AI Models: Design, Development, & Deployment	The team must design, develop, and deploy AI models by using a structured, iterative approach. This incorporates model research, data preparation, model testing and continuous improvement.	The team may face challenges in balancing the technical requirements with the scalability and interoperability of the models across different deployment stages. Additionally, ensuring that the AI models are ethical and unbiased presents a significant challenge, especially when dealing with diverse data sets and complex model behaviors.	The team needs to develop the models in stages, continuously improving and iteratively training and testing the models. This ensures they meet operational needs and follow ethical standards. They must also use thorough methods to reduce bias and strong testing steps to make sure the models are fair and accurate in diverse scenarios, keeping them trustworthy and reliable.
Evaluating Performance & Determining Metrics	The team must establish KPIs to measure and manage the performance of their AI projects. The team should also develop a method to address and evaluate aspects of the project that are not directly measurable, integrating both quantitative and qualitative data.	The team might find difficulty in selecting appropriate KPIs that accurately reflect the complexity of AI projects, which encompass technical, business, and ethical aspects. Another challenge is the dynamic nature of AI systems, where performance can change over time due to factors like model drift, making continuous monitoring and adjustment of KPIs necessary.	The team should involve a diverse group of stakeholders in the selection and ongoing review of KPIs to ensure all perspectives are considered and the KPIs remain relevant across all project facets. Establishing a monitoring system equipped with dashboards and automated alerts can facilitate the tracking and management of KPIs, allowing for quick adjustments as the project evolves.
Governing AI	The team must enforce RAI principles throughout both practice and product, leveraging the policies and processes within the organization's governance framework for guidance.	The RAI principles span many considerations that are not mutually exclusive and will vary for different projects. The team may also struggle to identify or understand the policies, processes, or other governance controls that may be applicable or required of their specific use case. Between level of effort, evolving policies, and negligence, AI governance can be difficult to measure.	The team should carefully assess all RAI principles at every step of the lifecycle and collaborate with governing entities and review committees in the agency for support.

4.6.2. Writing an AI Case Study

By writing an AI case study, an AI team can reflect on its process and document the AI project's development and outcomes. An AI case study outlines the project's goals, actions taken, participants involved, and lessons learned. This documentation can be completed during or after an AI project is completed and should involve contributions from all team members. It provides insights that enables other teams to reproduce the work as well as spread knowledge within and beyond CMS. AI case studies highlight a project's successes and challenges, offering valuable learning opportunities that can guide future projects and drive innovation. See Appendix A for previously documented AI case studies.

AI Case Study Template

The following Table 31 is an AI case study Template that the AI team can follow to ensure that they are documenting all aspects of their project.

Table 31. AI Case Study Template

Category	Description	Questions to Consider
Project Overview	When introducing your AI case study, provide an overview of the business use case, project logistics (including funding), and methods used to measure the project's success.	• What was the project goal and how was it established? • How was the project funded? • Was the project goal achieved? Why or why not? What can be learned? • Did the project goal evolve over time? • How was the success of the project measured? • Was a cost/benefit analysis conducted? How did the outcome compare to initial projections?
Stakeholder and User Engagement	Stakeholder and user engagement is integral to the success of a project. Detail who your key stakeholders and users were and to what extent the project met their requirements and needs.	• Who were the key stakeholders? (Stakeholders can range from individuals to teams and components.) • How were stakeholder needs and expectations aligned to project objectives? • Who were the end users of the proposed solution and what were their needs? • How were their needs met by the project?
Ethics	Provide information on how ethical considerations and compliance requirements were managed throughout the project. Consider RAI principles and the involvement of DevSecOps.	• How were ethical concerns identified and managed throughout the project? • How were compliance requirements identified and managed throughout the project?

Category	Description	Questions to Consider
Data Management	Address the acquisition, usage, and governance of data used in the project. Describe how data was sourced, managed, secured, and maintained throughout the project lifecycle.	• How was the data collected or generated, and was synthetic data used? • What data was left out? • What permissions were required for the datasets, and where was the data stored? • What efforts were involved in data cleaning and security measures implementation? • Who annotated the data and how were they compensated? • Who had access to the data, and what are the plans for data disposal or ongoing access? • Where is the data stored?
AI Model Design, Development, and Deployment	Describe the process of creating, refining, and deploying machine learning models, from initial feature extraction and metric selection through to final implementation and ongoing evaluation.	• What metrics were chosen to evaluate model performance, and how were they validated? • How was feature extraction conducted, and what tools and principles guided this process? • Describe the types of models developed (proprietary or open source), where development occurred (on-premises or cloud), and the key resources used (hardware and software). • How were models transitioned from development to deployment, including changes in hardware or software and hosting details? • What mechanisms were put in place to monitor ongoing model performance? • What guardrails were established to prevent unwanted behaviors such as bias or hallucinations in the model? • How often are models retrained, and how does this align with initial expectations and resource allocation?
Impact and Lessons Learned	Evaluate the broader effects of the AI project on stakeholders, users, and CMS. Document outcomes and insights and focus on the practical and strategic lessons that can inform future AI initiatives.	• What were the primary impacts of the AI project on the organization and its stakeholders? • What were secondary downstream impacts? • What key lessons were learned from the project's successes and challenges? How were these lessons elicited? • How can the insights gained from this project be applied to improve future AI initiatives within the organization?
Resource Sharing and Collaboration	If possible, provide information on repositories and resources to facilitate the reuse and further development of the project outputs by other teams. This promotes transparency and fosters a collaborative environment for AI innovation across CMS.	• Was the data or code open-sourced? • Where can other teams access the GitHub repository or other data resources associated with this project? • What documentation is available to help others understand and utilize the shared resources effectively? • Are there any restrictions or guidelines that others should be aware of when accessing and using the project materials?

4.6.3. Pursuing AI Organizational Maturity

Adhering to and sharing best practices throughout the AI project lifecycle is critical for promoting AI innovation, advancing AI organizational maturity, and fostering collaboration across the agency. By documenting each step of the AI project, teams not only refine their approach through reflection but also contribute valuable insights to the CMS AI community. This collective learning environment, supported by systematic documentation enables AI teams to leverage past experiences and improve the effectiveness of future AI projects. These practices are vital for developing AI-based solutions that not only align with agency goals but also prepare CMS to meet the evolving expectations of the public.

5. Appendices

Appendix A. Case Studies

The following case studies were completed within CMS to better understand AI technologies, strengthen infrastructure, grow employee technical skills, and pursue AI organizational maturity.

The first case study, “Using Generative AI Solutions to Create a Multi-Model Interface for Enhanced User Accessibility” was conducted in 2024 and follows the proposed AI Case Study format detailed in Section 4.6.2.

The following case studies, “Ontology Development” and “OHC AI Pilot for Time to Hire Prediction” were completed in 2022. They were originally published in the CMS AI Playbook V2 and are represented in the original format.

A.1. Using Generative AI Solutions to Create a Multi-Model Interface for Enhanced User Accessibility

A.1.1. Project Overview

The AI Explorers team sought to understand how CMS's technology and infrastructure could support AI-based solutions to make documents more accessible to diverse readers. To illustrate this, they used the "Medicare and You 2024" Handbook as an example of a CMS-specific document to showcase their solution. This AI-based solution would help users more easily access and understand information from the handbook and other similarly comprehensive documents.

Their hypothesis was that LLMs could enhance document accessibility by streamlining the process of retrieving information and delivering Handbook content relevant to the user through a chatbot interface, with minimal instances of model hallucinations. While the team trained models using the Handbook data, the purpose of this project was to showcase how training LLMs on CMS-specific documents could not only enhance the accessibility of other documents within CMS but also facilitate the deployment of models within the CMS infrastructure.

The solution developed by the team consists of four AI-based capabilities. These four capabilities are documented in the AI Explorer’s CMS-enterprise Confluence.

Table 32. Solution Capabilities to Increase Medicare Handbook Accessibility

Solution Capability	Potential Impact	Model and Tools Used
Medicare Handbook Chatbot	The team developed a chatbot user interface (UI) to accompany the “Medicare and You 2024” Handbook text that could answer questions users have about the information contained in the Handbook. This would allow users to easily search for specific answers rather than reading and scrolling through the document.	The chatbot answers questions by grounding an open-source LLM, Mistral 7B, with RAG (Retrieval-Augmented Generation). The model's response and speed were evaluated using Trulens, a tool assisted by LLMs. The UI was developed with Gradio and Plotly Dash.

Solution Capability	Potential Impact	Model and Tools Used
Document Upload and Query Tool	This tool allows the user to extend the LLMs capabilities beyond what it was trained on. Leveraging a UI users can add their own documents to the application, enabling the LLM to learn from a wider range of data. The tool provides users with an easy way to find information about their personal documents rather than manually searching the document themselves.	This capability uses Mistral 7B and allows users to upload their own documents for RAG querying.
Chatbot Variation: Text-to-Speech	This chatbot variation allows users to communicate with the chatbot verbally, increasing the accessibility of the Handbook information for users who may have low-vision or blindness.	This chatbot uses OpenAI's Whisper model for speech to text translation and then uses Mistral 7B to provide a response to that translation. The model's response is then processed using Microsoft's SpeechT5 model for speech synthesis (text-to-speech).
Foundational LLM and QA	This tool empowers users to test the capabilities of an LLM model using only the data it was initially trained on. Additionally, the tool gives users the ability to alter the LLMs parameters allowing for fine-tuning and customization of the model's responses.	Mistral 7B operating without RAG, providing a baseline for comparison.

A.1.2. Stakeholder and User Engagement

Key stakeholders in this project included the AI Explorers product manager, CMS Office of Information Technology (OIT) leadership, OIT data scientists, and the Open Source Program Office. Demonstrating progress to various stakeholders within OIT allowed the team to initiate discussions and lay the groundwork for potential future projects.

Additionally, conducting this exploratory research allowed the AI Explorers team to engage with the Knowledge Management Platform (KMP) team and develop the solution capabilities within the KMP AI Workspace. This collaboration allowed both teams to understand what technologies are possible to develop within the KMP AI Workspace infrastructure.

Since the primary goal of this research was to explore the team's understanding of how technology and infrastructure could support AI projects within CMS, potential users of the Medicare Handbook solution capabilities were not engaged at this time. However, should development of this solution continue, future user research is required to ensure that it meets user needs.

A.1.3. Ethics

Regarding ethics and responsible AI, our team referenced the Health and Human Services Trustworthy AI Playbook as well as the latest research that covers the evaluation of LLMs (Health and Human Services (HHS) n.d.) (Guo, et al. 2023). The team also used Facebook AI Research (FAIR) and Trulens to understand the trustworthiness of our models.

The Medicare Handbook Chatbot required an additional level of responsible AI implementation that would apply to the user interface. To ensure Human-Centered AI interactions, our team applied the Microsoft Human-AI Experience (HAX) Toolkit guidelines to determine how the chatbot's features should work (Microsoft n.d.). The following guidelines can be seen in the chatbot's interface below.

Table 33. Examples of Microsoft Human-AI Experience Guidelines

Guideline	Description
G1. Make clear what the system can do	Set expectations for the user as to what the system was designed for.
G2. Make clear how well the system can do what it can do	Let users know that the system can make mistakes.
G3. Show contextually relevant information	Let users see examples of what they can ask the chatbot.

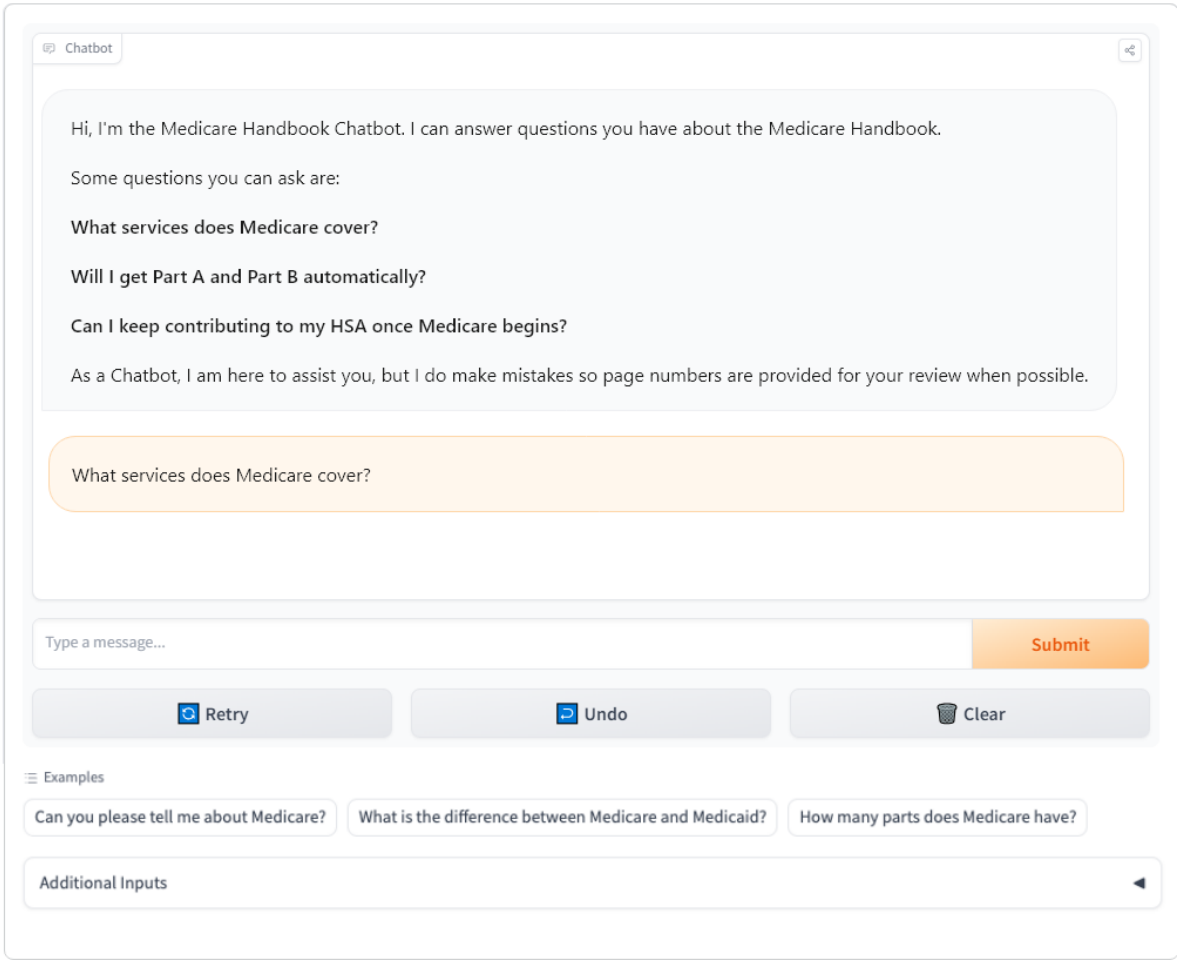


Figure 9. Medicare Handbook Chatbot Welcome Screen with Guidelines 1-3 Applied

A.1.4. Data Management

For the Medicare Handbook use case, the data management strategy for the RAG models involved collecting and generating data points to evaluate model quality and performance. Key metrics, including queries (prompts), responses, latency, and timestamps, were gathered, occasionally supplemented by additional metadata, without collecting any user-specific data. The Medicare Handbook Chatbot application leverages the "Medicare and You 2024" handbook to provide contextual information for user queries via retrievers.

The data collection process included all available information without any specific exclusions. Managing datasets required access to the AWS KMP AI Workspace, ensuring all data remained secure. The handbook file, although publicly accessible and free from sensitive details, was securely stored in AWS S3 under strict safety measures. This confirmed the environments viability for future projects with increased sensitivity needs.

The AI Explorers team carried out data annotation, offering human evaluations of model performance. Access to the data was limited to the internal team, and being the data wasn't considered sensitive, there was no need to discard or remove any information.

A.1.5. AI Model Design, Development and Deployment

For this project, the team explored open-source LLMs, such as Mistral 7B, that come pre-trained and ready to use. The team implemented a RAG approach, enabling the LLM to rely on documents provided to it, like the "Medicare and You 2024" Handbook. This method helped ground the AI's responses and reduced hallucinations or incorrect answers that can occur when a model generates incorrect answers based on its training data.

To evaluate the quality and relevance of the RAG model responses, the team developed a list of relevant and non-relevant questions about the Medicare Handbook, ranging from direct recall about the document to multi-step reasoning. They then graded the responses using human evaluation and TruLens. They used an LLM, GPT 4 Turbo to assess the following three key metrics:

Table 34. Metrics Gauging Relevance of Model Responses

Metric	Description
Context Relevance	This metric tests the effectiveness of the RAG setup's retriever by checking if the context it retrieves from the handbook is relevant.
Groundedness	This measures the factual accuracy of the chatbot's responses, according to the information in the source documents. .
Quality Assurance (QA) Relevance	This assesses whether the responses address the posed questions.

Guardrails were put into place with prompting and hallucinations were reduced by using a RAG setup, so that the LLM is very much encouraged to use data from the documents supplied to it. The Foundational LLM and QA application provides the user parameter adjustment functionality. These parameters include Max New Tokens to control the LLM response length, Temperature to manage the randomness of outputs, Top-p (Nucleus Sampling) to adjust the model's likeliness of word and phrase choice, and Repetition Penalty to alter model redundancy and text variety. As this was a playground, the same hardware and hosting was used for both development and deployment since large-scale production deployment was not in scope.

The images below show latency, cost, and performance evaluation scores on two open-source LLM's hosted on AWS, Llama-2-13B and Mistral 7B, and gpt-3.5-turbo through the OpenAI API. There are some caveats to the results. Context relevance, groundedness, and QA relevance are lower than the other metrics because some of the sample evaluation questions were not in the Medicare handbook so the returned responses could not provide an answer, which is desired behavior.

The speed was also negatively affected by running SOTA models on older hardware on AWS cloud. Since then, newer models have been released that offer improved accuracy and faster run-times. Overall, the chatbot effectively handled queries related to Medicare content across various levels of difficulty.

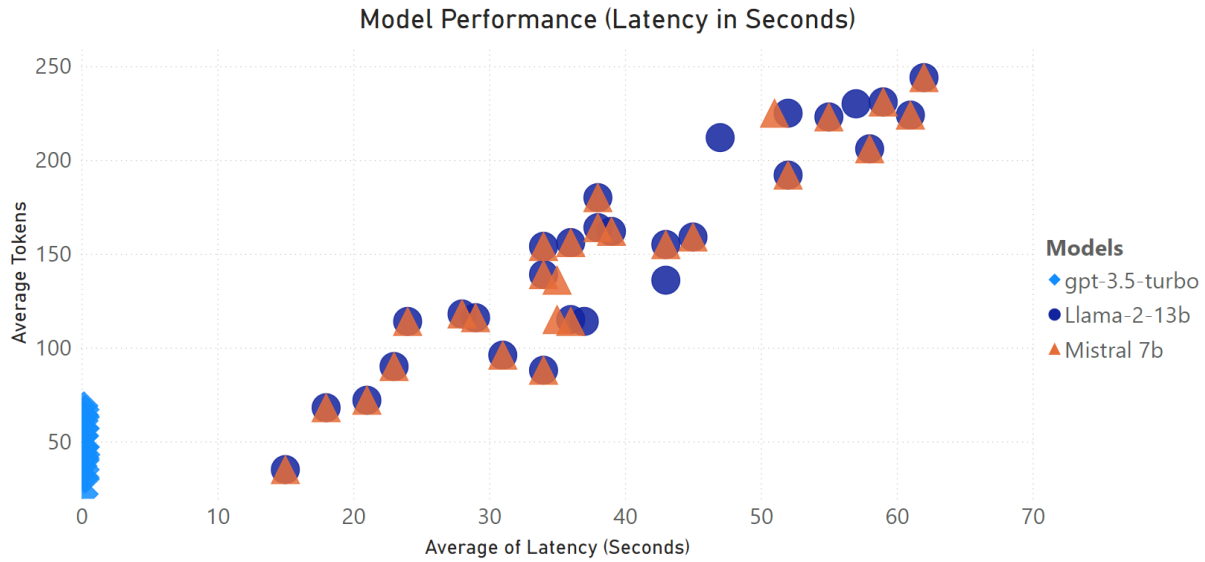


Figure 10. Comparison of Three Models' Abilities to Generate Tokens Measured by Latency.

GPT 3.5 (the light blue dots on the bottom left) has the best performance because it is accessed by an API and run on SOTA hardware by OpenAI. Llama and Mistral have comparable performance but inferior speed because they were run on local not SOTA hardware.

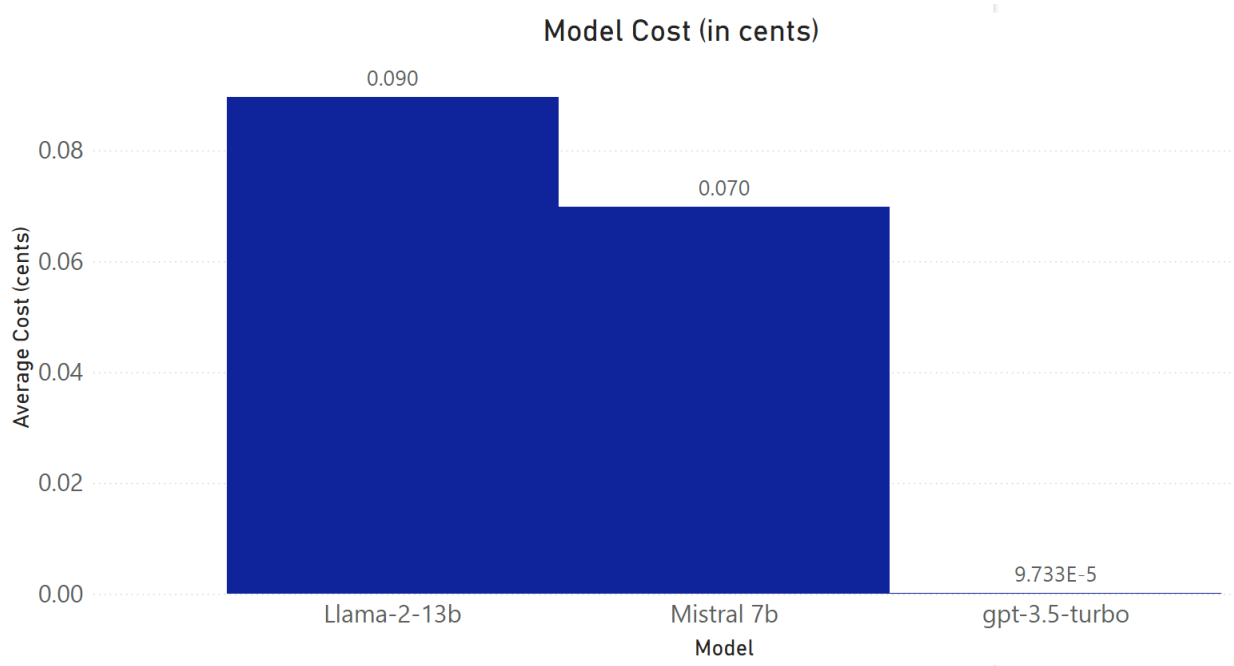


Figure 11. Estimated Cost (in Cents) Based on the Cost of Hosting the Tool/Model Divided by Average Time Taken for a Model Request vs. gpt-3.5-turbo.

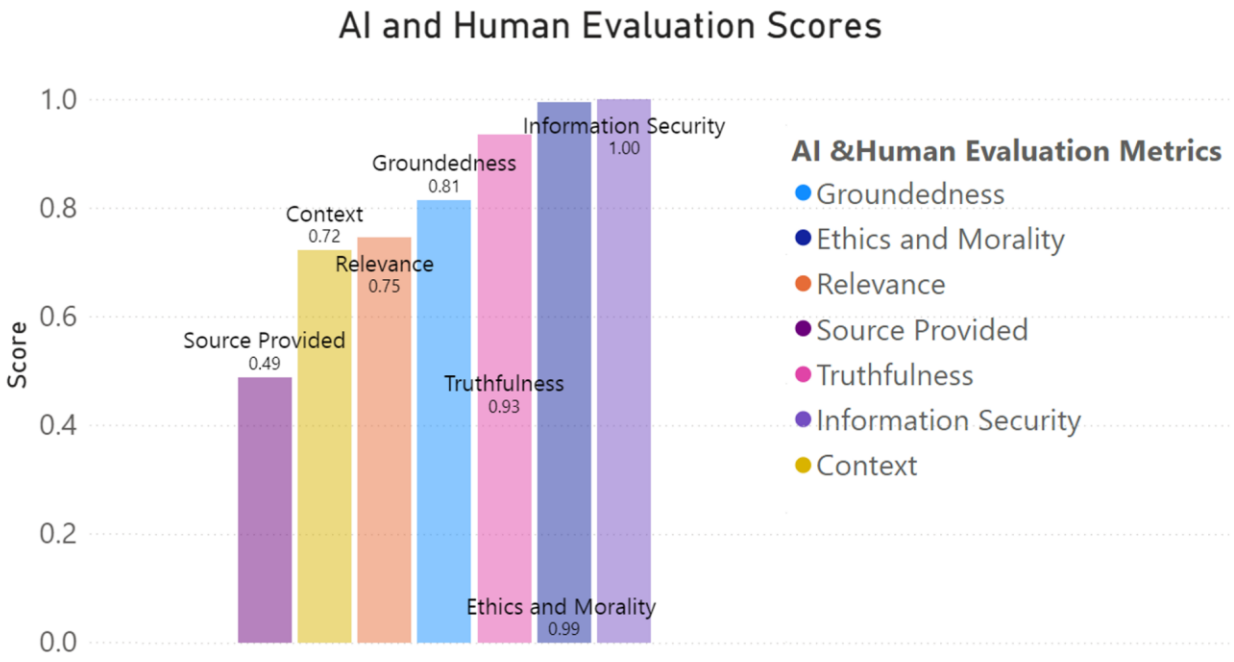


Figure 12. While There Are Some Discrepancies in Performance Between the Three Models, Based on Individual Measures, Overall They Had Similar Performance to This Figure.

A.1.6. Impact and Lessons Learned

The impact and lessons learned from this AI project were significant in understanding what was possible to deploy within the CMS infrastructure and the level of effort required. Through the team demoing their progress to a variety of stakeholders, the project fostered discussions within the Office of Information Technology (OIT) about technical possibilities and CMS's existing infrastructure.

Through this AI project, the team discovered that limited computational availability required the team to use older GPUs, resulting in slower application performance. When evaluating LLM performance, the team used TruLens alongside human evaluations. Though TruLens proved effective, further assessment is needed to fully understand its quality as an evaluation tool. Furthermore, by collaborating with the KMP team, this project became the first to deploy within the KMP AI Workspace. This partnership enabled the team to make recommendations that have been incorporated into the platform.

A.1.7. Resource Sharing and Collaboration

The use of open-source tools was a top priority for the project team. To promote transparency and reuse of code, the project team intends to open-source the code on GitHub and share all other details in the repository. Additionally, the project team modeled the code and repository after the CMS Open Source Policy and the template used by U.S. Federal open source projects.

To operate TruLens users must posses a valid OpenAI API key. The provided scripts on GitHub include designated sections for users to enter their individual API key. This is a mandatory requirement for accessing and using these services.

A.2. Ontology Development

A point project with Manifold and CPMS kicked off in June 2021 to explore two example data sets: one from the ServiceNow ticketing system and one from the knowledge base articles used for self-service support. The goal was to understand if an ontology could be built from concepts contained in these data sets.

A small team of four Data Scientists and Machine Learning Engineers (MLEs) explored the data sets to see what information could be extracted from those data sets. We will focus here on the overall discovery process.

A.2.1. Initial Data Exploration

A.2.1.1. Goals

As is typical for most exploratory analysis, our goals in this phase were as follows:

- Develop an initial understanding of the data (scale/size, relevant columns, degree of missing data).
- Identify trends and patterns and potential future approaches.
- Evaluate feasibility of technical approaches to solve initial business problems.
- Create a set of findings and open questions to begin a dialogue with organizational stakeholders (e.g., business owners, etc.).

While the exact goals and mode of implementation may vary from use case to use case, these goals can serve as a general template for the initial exploration.

A.2.1.2. What We Did

- Examine quantitative variables (ticket duration / resolution time, number of tickets, etc.).

- Examine various data groupings (category, subcategory, assignee, contact type, customer, CMS subdivision, etc.).
- Examine temporal data.
- e.g., category prevalence over time, unusual events like sudden spikes in term frequency, seasonal trends, and obsolete vs. new terms.
- Natural language processing (NLP) analysis of unstructured text (short description, description, close notes, etc.) using term frequency/inverse document frequency (TF-IDF).

A.2.1.3. Outcomes and What We Learned

- Identified key trends and patterns, such as:
  - Password resets and account locks were the most common ticket type.
  - Resolution time varied by category and sub-category.
  - Many subcategories were rare.
  - Among cases, financial/enrollment tickets took the longest total time.
  - Among incidents, security tickets took the longest.
  - Distribution of category/time varied greatly from account to account.
  - De-duplicating tokens would be necessary to see more patterns.
  - Several categories showed seasonal trends (e.g., training in the fall, remote work, and access at the start of the pandemic, etc.).
- Developed a set of action items for future phases.
  - Selected categorical variables useful for future analyses and comparisons (e.g., identifying category as a key signal).
  - Need for robust, de-duplicated concepts or entities instead of just raw n-grams.

A.2.2. Refine the Findings: Concepts, Entities, and Vocabularies

A.2.2.1. Goals

Based on the results of the exploratory phase, we established the following goals:

- Develop a set of concepts and entities to:
  - facilitate future work on ontology learning and taxonomy representation.
  - de-duplicate and refine results from exploratory phase.

A.2.2.2. What We Did

- Resolve repeated phrases (n-grams) for the same concept (e.g., reset, password, reset password, etc.): the remaining high-frequency phrases capture “concepts” or “entities” in the data.
- Visualized concepts as a graph to show links between concepts based on co-occurrence in tickets.
- Used graph clustering to extract sets of related concepts and their associated categories.

A.2.2.3. Outcomes and What We Learned

- Developed vocabularies of key concepts and entities (one each for cases and incidents), which were useful for all future NLP analysis.
- Unsupervised graph clustering algorithms found concept groups relating to training/education, policy, access, technical assistance, and more.

After this phase, the team met with several subject matter experts (SME) and business owners to discuss findings to date, and to learn what specific use cases might be investigated in the time remaining on the project.

A.2.3. Explore Additional Use Cases

A.2.3.1. Goals

- Based on extended consultation and discussion with OIT project leads and business owners, the team established the following three areas as goals.
- Explore ability to predict ticket resolution time, reassignment, failure, etc.
- Continue ontology learning by attempting to learn entity relationships.
- Identify links between knowledge articles and cases that would value from the knowledge for accelerated resolution.

A.2.3.2. What We Did

Near the start of this phase of the project, the team acquired an additional dataset: approximately two hundred "tier zero" knowledgebase articles.

- Attempt relationship extraction by searching for sentences with multiple concepts.
- Identify connections between articles and tickets (e.g., suggest articles based on ticket descriptions).
- Developed a new vocabulary based on articles, using TF-IDF (the article dataset was too small to reliably use the entity and concept process described above).
- Build interpretable decision tree prediction models to predict resolution time (binarized) and reassignment.

A.2.3.3. Outcomes and What We Learned

- In entity relationship learning for ontology learning, structured data is ideal. ServiceNow ticket data was found to be too noisy, unstructured, and informal to extract relationships between different entities.
- Prediction performance was generally good, and the team's use of interpretable decision trees highlighted key variables that were associated with fast ticket resolution (common descriptions for known categories, etc.).
- The knowledgebase articles did not cover the most common use cases for incidents and cases. Therefore, most tickets did not have a reliable "best article" suggestion (some of these examples were due to the need for help desk staff intervention, e.g., password reset).

A.2.4. Wrapping Up

As is common in point projects, the results were both a body of knowledge developed from the exploration and a set of potential areas for future focus.
The primary project learnings include:

1. Methodologies to extract taxonomy/ontology from unstructured data. This methodology could be generalized in the future to other unstructured datasets.
2. Validated the methodology with a second unstructured dataset to determine our ability to detect relationships between multiple datasets for the creation of future ontologies.
3. Generated several preliminary predictive models with actionable results.

A.2.4.1. Future Focus Areas Can Include:

- Expanding the number of datasets to allow for development of a true ontology.
- Developing automatic outlier/anomaly detection to understand users' needs.
- Pre-incident alerting and user notification (e.g., "We've seen an increase in VPN issues today; if that is your issue, can we point you to a given resource?").

A.3. OHC AI Pilot for Time to Hire Prediction

A.3.1.1. Introduction

The Office of Human Capital (OHC) AI Pilot team built a proof-of-concept "Time-To-Hire Calculator" that provides clarity into the hiring process for Hiring Managers and supports shorter hiring timelines. Using open-source tools and data from USA Staffing reports we processed the data, trainedML models, designed an explainable and user-friendly interface, then deployed these interwoven components securely to the AWS cloud – all while drawing on human-centered design processes to verify that the product fit the needs of users.

A.3.1.2. Challenge

The OHC Division of Workforce Analytics and Accountability has access to a variety of raw datasets and wanted to explore ways to make those datasets accessible to CMS employees and support more effective work. With the AI Explorers Program, the Division wanted to address the following challenges.

1) Develop a user-centered prototype from a raw dataset by processing the data through ML models and deploying the prototype in the cloud.

2) Ensure the process taken to develop the solution is repeatable and applicable to future datasets.

A.3.1.3. Solution

Our solution to these challenges consisted of the following steps:

1) **Exploratory Analysis**: Understanding the datasets and initial Proof of Concept prototype, and building a new extracting, transforming, and loading (ETL) process.
2) **HCD, UX, and Interface Development**: Identifying business questions, conducting user interviews and testing, iterating on wireframes, and developing the interface prototype.
3) **(ML Model Development**: Building, training, testing, and evaluating various Python ML models, while incorporating RAI and Explainable AI (XAI).
4) **Deployment to the AWS Cloud**: Deploying the tool to the AWS Commercial Cloud in a CMS Cloud environment to provide a path to making the application accessible to users. The team utilized AWS managed services and documented CMS Cloud configuration and CMS security requirements to provide a secure deployment model and a path towards achieving ATO.

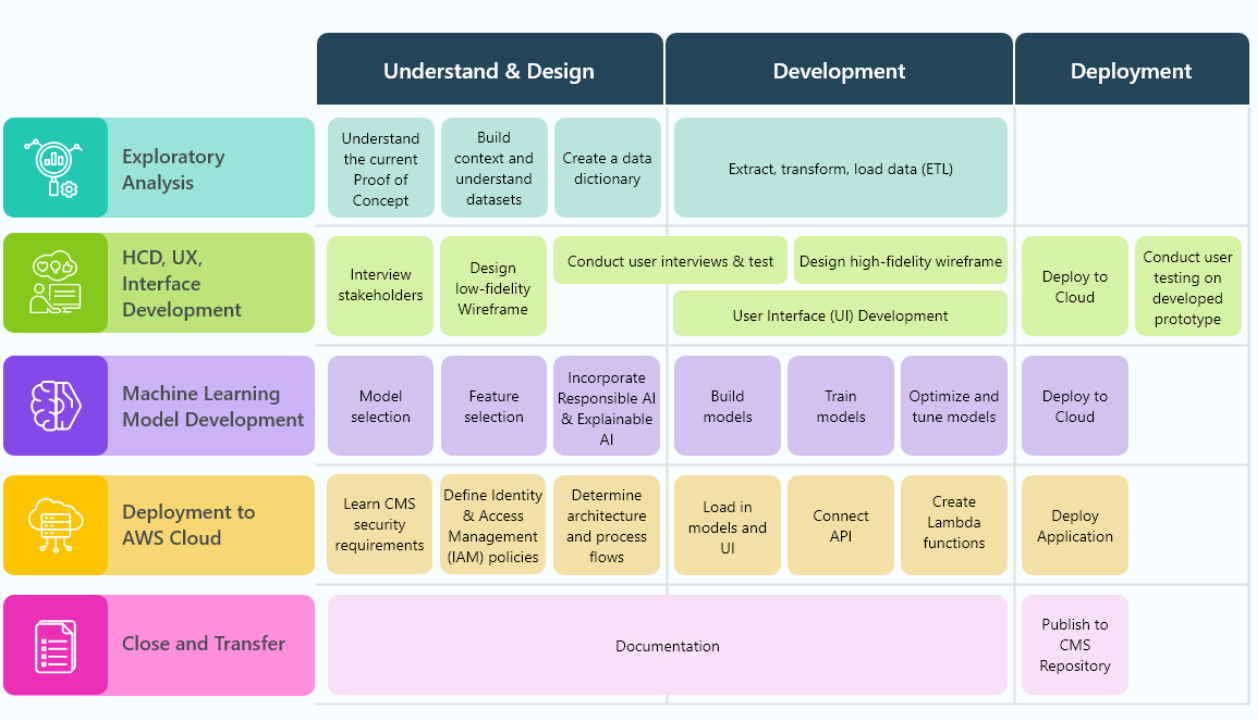


Figure 13. Process to Create a Product Solution

A.3.1.4. Impact

As a result of this project:

1) OHC has a technically functional and user-centered "Time-to-Hire Calculator" prototype that allows Hiring Managers to assess how long it would take to hire a candidate for a specific job description. Along with the product, we are providing documentation, research, and recommendations for future versions of the product.
2) OHC and the AI Explorers program were able to participate in and learn from the solutioning process. The process is well-documented and can be repeated for future projects involving large datasets, AI and ML modeling, and human-centered design.

A.3.2. Exploratory Analysis

A.3.2.1. Goals

- Understand OHC's project goals, current processes, data collection, and data insights.
- Assess the initial proof of concept, datasets, and user interface and determine what improvements could be made.
- Streamline and automate the ETL process.

A.3.2.2. What We Did

- Reviewed the current proof of concept (PoC) and related materials (e.g., datasets, ETL flows, ML and UI scripts) to establish a full understanding of the baseline functionality and resources.
- Produced a new data dictionary for the project, and a new ETL process comprised of three Python-based scripts: 'generate_certificate_files.py', 'generate_time_to_hire_regression_data_file.py', and 'TTH_Summary_Stats.ipynb'.

A.3.2.3. Outcomes and What We Learned

- **Understanding the PoC**: Familiarization of the initial PoC's design, structure, and process flows supported the re-engineering process necessary for enhanced pilot functionality.
- **Standardization**: Creation of a data catalog, to include both available and utilized elements, helped track all applicable features and data provenance. For example, standardization of terms utilized under 'Job Title' supported the transition to statistical-based input features.
- **Designing a New ETL**: Consolidation of data processing into a single Python-based platform to perform all ETL actions and generate required output files allowed for easier transition to AWS-based platform for pilot.

A.3.3. HCD, UX, Interface Development

A.3.3.1. Goals

- Understand how potential users of the product (hiring specialists and hiring managers) go through the hiring process, what tools they use, tasks they need to complete and what their pain points are.
- Understand to what extent the data collected and product interface meets users' needs.
- Develop a user-centered working prototype that integrates data, models, and user interface.

A.3.3.2. What We Did

- **Context Building**: Conducted initial review of material and interviewed a PoC stakeholder to define design and business questions and corresponding interface features.
- **Wireframe Iterations and Testing**: Conducted user interviews/testing and XAI/RAI research, then used the feedback and insights to iterate on interface wireframes.
- **Product Interface Development**: Followed a structured control system for code tracking and collaboration between design experts and interface developers to combine user feedback, interface design, backend integration.
- **Internal Prototype Testing**: Conducted user testing on the developed prototype using members of our internal team to identify any remaining general, functional, or stylistic pieces of feedback, then prioritized or backlogged items for the remainder of the pilot.

A.3.3.3. Outcomes and What We Learned

- **Collaboration is Key**: Scheduled time for collaboration between developers and designers is key in ensuring that all team members understand any constraints and challenges that other team members may face in their work stream and how that may impact the product.
- **Robust Datasets to Enhance the User Experience**: Asking users to input their own data into the application to make up for a lack of data and resultant low-accuracy predictions can decrease user confidence in the tool's usefulness, as well as the data's accuracy.
- **Provide Logical Context when User Testing**: When conducting user testing, the logistics of the wireframe and testing should be clearly scripted out in the research plan and shared with the user, including any limits within the prototype or wireframe.

A.3.4. Model Development

A.3.4.1. Goals

- Analyze the original PoC and the initial model development.
- Provide baseline analysis of current model evaluations for future comparison as the models become further developed and altered.
- Utilize ML best practices and tools to develop thorough models and choose the one with the highest yielding results, that also provide value to users.

A.3.4.2. What We Did

1) **Analyze and Evaluate Initial ML Models**: Created reusable processes and templates in our approach to analyze and evaluate the initial ML models. These include a model selection process, Notebook template and Model cards.
2) **Model Iterations**: Used initial analysis and evaluations to iterate and improve on the models, including efforts toward feature selection and hyperparameter tuning.
3) **Model Selection and Deployment**: Made a final selection of models and processes to use based on evaluation results, client recommendations, and pilot time constraints.

A.3.4.3. Outcomes and What We Learned

- **Model Selection**: A Linear Regression model was deployed for predicting time to hire. This model performed well against the other models tested (on metrics R^2, adj. R^2, MSE, RMSE) and accommodated the client's preference for a simpler and more interpretable model. XGBoost, our chosen classification model, was backlogged and excluded from our prototype.
- **Feature Selection**: The time to hire model is trained using the five hiring phases. The application uses statistical averages filtered by position title and grade, rather than ML, for model inputs due to time constraints of the pilot. While a sufficient workaround for the pilot, this highlights the importance of feature selection and its impact on model performance.
- **Creating ML Resources for Development and Comprehension**: The model team created reusable templates and structured workflows for selecting, training, and iterative logging and evaluation of model performance to streamline future model development processes.

A.3.5. Cloud Development

A.3.5.1. Goals

1) Provide a secure deployment model for the OHC AI Pilot Hiring Assessment Tool.
2) Deploy Version 1 of the OHC AI Pilot Hiring Assessment Tool on AWS.

A.3.5.2. What We Did

- **Accessing the CMS AWS Environment**: Collaborated with CMS Support to obtain access to the AWS sandbox environment, then set up Identity and Access Management (IAM) roles (demo provided) within the environment..
- **Designing the Model Architecture**: Surveyed AWS services approved for use in the CMS cloud system and modeled the architecture of our pilot application.
- **Deployment to AWS**: Transferred our application to the cloud by deploying the data ETL process, web app server hosting, and data transmission process. The deployed application successfully

populates with model results from user input and access requires the user to be signed into the CMS Cloud VPN.

- **ATO & Security**: Researched the process for obtaining ATO in the CMS Cloud, with emphasis on security considerations laid out within the Security Impact Analysis (SIA) and the intent to leverage an existing ATO boundary. See: provided SIA and ATO Process Overview

A.3.5.3. Outcomes and What We Learned

- **Cloud and Security Capabilities**: The team was not as experienced in CMS Cloud expertise resulting in a slow start to the cloud deployment phase. As a result, several tasks had to be backlogged for future recommendation. Nonetheless, we gained significant capabilities and outlined a security plan that strengthened our competence in the cloud field.
- **Putting the Pieces Together**: Deployment to the cloud required full collaboration between CloudOps and the rest of the team to connect the previously generated and researched content together as one working system. We were successful by engaging in high responsiveness amongst the team and conducting focused working sessions as needed.
- **Technical Grievances**: There were several challenges and technical grievances that we faced while working in the CMS AWS environment which required additional efforts to work around. We endorse prompt communication with CMS Cloud Support regarding any roadblocks, especially where there are significant time restraints such as during a short pilot.

A.3.6. Future Considerations

This pilot produced a working prototype that is both technically functional and meets users' needs. In order to continuously improve the product and prepare for it to be used at scale by CMS employees, we provide an extensive list of future enhancements for consideration.

As with any production, all steps making up the solution to this pilot have room for improvement and expansion for the next iteration of the deployed application. Key suggestions that will support the serviceability and scaling of the application include improving the models and datasets, building out our proposed interface and cloud architecture, securing the application in line with an ATO, and reviewing our HCD findings regarding the user need for a real-time tracking tool.

A.3.7. Resources

Unabridged documentation of the OHC AI Pilot and Time-To-Hire Calculator, including implementation details, artifacts, and future considerations, can be found on the CMS AI Explorer Program Confluence site under Awarded Projects and the CMS Github repository.

Appendix B. Reference Materials

2024. February 25. https://redresscompliance.com/ai-services-integration-with-existing-business-systems/.

Ahmed, Arfan, Sarah Aziz, Mahmood Alzubaidi, and Jens Schneider. 2023. "Wearable devices for anxiety & depression: A scoping review." *National Library of Medicine.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9884643/.

n.d. *AI and machine learning products.* https://cloud.google.com/products/ai.

AI Explorers. 2023. "Generative AI Tools: CMS Uses and Risks." *Technical Reference Architecture.* December 12. https://www.cms.gov/tra/Research_Spotlights/RS_231212_Generative_AI_Uses_and_Risks.htm.

n.d. *Amazon SageMaker.* https://aws.amazon.com/sagemaker/.

Appen. 2021. February 26. https://www.appen.com/blog/ai-model-maintenance-guide-to-managing-model.

Apple. n.d. *Healthcare.* https://www.apple.com/healthcare/apple-watch/.

Arsanjani, Ali. 2023. March 21. https://dr-arsanjani.medium.com/the-generative-ai-life-cycle-fb2271a70349.

Baker, Stephanie, and Wei Xiang. 2023. "Explainable AI is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence." *arXiv.org.* December 4. https://arxiv.org/pdf/2312.01555.pdf.

Białek, Jakub. 2023. "Understanding Data Shift: Impact on Machine Learning Model Performance." *nannyML.* 03 14. Accessed 2024. https://www.nannyml.com/blog/types-of-data-shift.

Centers for Medicare and Medicaid Services. n.d. "CMS Cross Cutting Initiative: Data to Drive Decision Making." *CMS.* https://www.cms.gov/files/document/data-drive-decision-making.pdf.

n.d. *Cloudely.* https://cloudely.com/the-lifecycle-of-generative-ai-in-simple-steps/.

CMS. 2024. *2024 System Census Survey Information.* Accessed 03 2024. https://share.cms.gov/Office/OIT/EADG/DEA/SitePages/FY24%20System%20Census%20Survey%20Information.aspx.

—. n.d. *About CMS.* Accessed April 8, 2024. https://www.cms.gov/about-cms.

CMS AI Explorers. 2024. *CMS AI Explorers Program.* April 3. Accessed April 8, 2024. https://confluenceent.cms.gov/display/APP/CMS+AI+Explorers+Program.

—. 2024. *CMS Medicare Handbook Chatbot.* 03 05. Accessed 04 16, 2024. https://confluenceent.cms.gov/display/APP/CMS+Medicare+Handbook+Chatbot.

CMS. n.d. *AI Health Outcomes Challenge.* Accessed April 4, 2024. https://www.cms.gov/priorities/innovation/innovation-models/artificial-intelligence-health-outcomes-challenge.

—. n.d. *Artificial Intelligence at CMS.* Accessed April 8, 2024. https://ai.cms.gov/.

—. 2024. "IT Procurements and Acquisitions: Market Research." *CMS Connect.* January 17. https://cmsitsm.servicenowservices.com/connect?page=kb_article&sys_kb_id=4a7c35a58773b1504297ba28cebb3513.

—. n.d. *Person-Centered Care.* Accessed March 2024. https://www.cms.gov/priorities/innovation/key-concepts/person-centered-care#:~:text=Person-centered%20care%20allows%20patients%20to%20make%20informed%20decisions,to%20them%2C%20and%20are%20accountable%20for%20their%20care.

2024. *Couchbase.* January 10. https://www.couchbase.com/blog/large-language-models-explained/.

Coursera Staff. 2024. "AI Programming Languages: What to Know in 2024." *Coursera.* March 21. https://www.coursera.org/articles/ai-programming-languages.

2023. *deepchecks.* July 24. https://deepchecks.com/ml-model-maintenance/.

Department of Homeland Security. 2024. "Artificial Intelligence Roadmap 2024." March 14.

DoD Responsible AI Working Council. 2022. "U.S. DoD Responsible AI Strategy and Implementation Pathway." June. https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf.

n.d. *Domino.* https://domino.ai/data-science-dictionary/model-selection.

Dunnmon, Jared, Bryce Goodman, Peter Kirechu, Carol Smith, and Alexandrea Van Deusen. 2021. "Responsible AI Guidelines." *Defense Innovation Unit.* November 10. https://www.diu.mil/responsible-ai-guidelines.

Endra. 2023. *Challenges in AI Adoption: Technical, societal, and organizational barriers.* 08 15. Accessed 03 2024. https://medium.com/aimonks/challenges-in-ai-adoption-technical-societal-and-organizational-barriers-abdd4614073e.

2020. "Federal Data Strategy Data Governance Playbook." *Federal Enterprise Data Resources.* July. https://resources.data.gov/assets/documents/fds-data-governance-playbook.pdf.

Ferrara, Emilio. 2023. "Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies." *arXiv.org.* December 7. https://arxiv.org/ftp/arxiv/papers/2304/2304.07683.pdf.

General Services Administration (GSA). n.d. "AI Guide for Government: Understanding and managing the AI lifecycle." https://coe.gsa.gov/coe/ai-guide-for-government/understanding-managing-ai-lifecycle/.

Ghai, Jitesh. 2022. "UNLOCKING THE POWER OF AI WITH DATA MANAGEMENT." *Capgemini.* March 2. https://www.capgemini.com/insights/expert-perspectives/unlocking-the-power-of-ai-with-data-management/.

Gonzales, Aldren, Guruprabha Guruswamy, and Scott R. Smith. 2023. "PLOS Digital Health." Accessed April 2023. https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082.

Google. 2023. *What is Multimodal Search: "LLMs with vision" change businesses.* August 21. https://cloud.google.com/blog/products/ai-machine-learning/multimodal-generative-ai-search.

GSA Artificial Intelligence Center of Excellence (AI CoE). n.d. "AI Guide for Government." *GSA - IT Modernization Centers of Excellence.* Accessed February 2024. https://coe.gsa.gov/coe/ai-guide-for-government/introduction/index.html.

GSA. n.d. *HCD Discovery Guide .* Accessed April 9, 2024. https://www.gsa.gov/system/files/HCD-Discovery-Guide-Interagency-v12-1.pdf.

Guo, Zishan, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi Linhao Yu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. *Evaluating Large Language Models: A Comprehensive Survey.* Accessed May 6, 2024. https://ar5iv.labs.arxiv.org/html/2310.19736.

Health and Human Services (HHS). n.d. *hhs.gov.* https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf#page=96&zoom=100,0,0.

2021. "HHS Trustworthy AI Playbook." https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf.

IEEE Global Initative. 2017. "Ethically Aligned Design, Version 2." *Institute of Electrical and Electronics Engineers Global Initiative on Ethics of Autonomous and Intelligent Systems.* https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.

Intel. 2023. *Moore's Law.* September 18. https://www.intel.com/content/www/us/en/newsroom/resources/moores-law.html#gs.848w35.

n.d. *Intellipat.* https://intellipaat.com/blog/what-is-ai-project-cycle/.

Interaction Design Foundation. n.d. *A List of the Most Common UX Deliverables.* Accessed April 9, 2024. https://www.interaction-design.org/literature/topics/ux-deliverables#a_list_of_the_most_common_ux_deliverables-9.

Jason, Wei et al. 2022. "Emergent Abilities of Large Language Models." *Cornell University, arxiv.* 10 26. Accessed 2024. https://arxiv.org/abs/2206.07682.

Katsoulakis, Evangelia, Qi Wang, Huanmei Wu, Leili Shahriyari, and Richard Fletcher. 2024. *Digital twins for health: a scoping review.* March 22. https://www.nature.com/articles/s41746-024-01073-0.

Krause, Rachel. 2022. *Creating Engaging Reports & Asynchronous Presentations.* April 3. Accessed April 9, 2024. https://www.nngroup.com/articles/engaging-reports-presentations/.

Lavin, Alexander et al. 2022. "Technology readiness levels for machine learning systems." *Nature Communications.* 10 20. https://www.nature.com/articles/s41467-022-33128-9.

Lee, Wanbil W, Wolfgang Zankl, and Henry Chang. 2016. "An Ethical Approach to Data Privacy Protection." *ISACA.* https://www.isaca.org/resources/isaca-journal/issues/2016/volume-6/an-ethical-approach-to-data-privacy-protection.

Lu, Sheng, Bigoulaeva, Irina, Sachdeva, Rachneet, Tayyar Madabushi, Harish, Gurevych, Iryna. 2023. "Are Emergent Abilities in Large Language Models just In-Context Learning?" *Cornell University, arxiv.* 09 04. Accessed 2024. https://arxiv.org/abs/2309.01809.

Ma, Shuming, Hongyu Wang, Lingxiao Ma, and Lei Wang. 2024. *The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits.* February 27. https://huggingface.co/papers/2402.17764.

Manning, Catherine G. 2023. *Technology Readiness Levels.* September 27. https://www.nasa.gov/directorates/somd/space-communications-navigation-program/technology-readiness-levels/.

Mäntymäki, Matti, Matti Minkkinen, Teemu Birkstedt, and Mika Viljanen. 2023. "Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance." *arxiv.* January 31. https://doi.org/10.48550/arXiv.2206.00335.

McKinsey & Company. 2023. *What is digital-twin technology?* July 12. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-digital-twin-technology.

McKinzie, Brandon et al. 2024. "MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training." *Cornell University, arxiv.* 03 22. Accessed 2024. https://arxiv.org/abs/2403.09611.

Mell, Peter, and Timothy Grance. 2011. "The NIST Definition of Cloud Computing." September. https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf.

Microsoft. n.d. *Microsoft HAX Toolkit.* Accessed April 9, 2024. https://www.microsoft.com/en-us/haxtoolkit/.

Mortensen, Ditte Hvas. 2021. *How to Involve Stakeholders in Your User Research.* Accessed April 9, 2024. https://www.interaction-design.org/literature/article/how-to-involve-stakeholders-in-your-user-research.

Mungoli, Neelesh. 2023. "Scalable, Distributed AI Frameworks: Leveraging Cloud Computing for Enhanced Deep Learning Performance and Efficiency." April 26. https://arxiv.org/pdf/2304.13738.pdf.

Noblis. 2023. "Artificial Intelligence (AI) Field Guide for Public Sector Enterprises." *Noblis.* Accessed 03 2024. https://noblis.org/aiguide/.

n.d. *Oden Technologies.* https://oden.io/glossary/model-training/.

Office of Public Affairs. 2024. "Department of Commerce Announces New Actions to Implement President Biden's Executive Order on AI." April 29. https://www.commerce.gov/news/press-releases/2024/04/department-commerce-announces-new-actions-implement-president-bidens.

n.d. "OHC Time to Hire Calculator." https://github.com/CMS-Enterprise/ai_explorers/tree/main/OHC%20Pilot.

Paech, Samuel J. 2024. "EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models." *Cornell University - arxiv.* 01 03. Accessed 2024. https://arxiv.org/abs/2312.06281.

PwC. 2024. *Tech Translated: Spatial computing.* February 7. https://www.pwc.com/gx/en/issues/technology/spatial-computing.html.

Ramesh, Rohit. 2023. January 31. https://blog.segmind.com/the-lifecycle-of-a-generative-ai-model-from-idea-to-deployment/.

Rosala, Maria. 2020. *The Discovery Phase in UX Projects.* March 15. Accessed April 9, 2024. https://www.nngroup.com/articles/discovery-phase/.

Saltz, Jeff. 2024. *Data Science Alliance.* February 2. https://www.datascience-pm.com/the-genai-life-cycle/#:~:text=While%20the%20tasks%20are%20very,development%2Fmodeling%2C%20and%20evaluation.

—. 2023. *Data Science Alliance.* https://www.datascience-pm.com/ai-lifecycle/#:~:text=As%20shown%20below%2C%20the%206,%3B%20and%20(6)%20MLOps.

Sanseviero, Omar. 2024. "LLM Evals and Benchmarking." *github, Omar Sanseviero.* 03 10. Accessed 2024. https://osanseviero.github.io/hackerllama/blog/posts/llm_evals/.

Shneiderman, Ben. 2022. "Human-Centered AI." In *Human-Centered AI*, by Ben Shneiderman, 7-13. Oxford: Oxford University Press.

Smith, L. W. 2000. *Stakeholder analysis: a pivotal practice of successful projects.* Accessed April 9, 2024. https://www.pmi.org/learning/library/stakeholder-analysis-pivotal-practice-projects-8905.

Tam, Adrian. 2023. *Machine Learning mastery.* July 20. https://machinelearningmastery.com/what-are-large-language-models/.

n.d. *The User Experience Research Field Guide.* Accessed April 9, 2024. https://www.userinterviews.com/ux-research-field-guide.

Tompkins, Sarah. 2024. *Demystifying AI: Navigating Myths, Harnessing Innovations at CMS.* March 4. Accessed April 8, 2024. https://planetoit.cms.gov/articles/demystifying-ai-navigating-myths-harnessing-innovations-cms.

Turing, Alan. 1950. *Computing Machinery and Intelligence.* https://medium.com/@jetnew/a-summary-of-alan-m-turings-computing-machinery-and-intelligence-fd714d187c0b.

United States Government. 2023. *Explore Government Uses of AI.* 09 01. Accessed 03 2024. https://ai.gov/ai-use-cases/.

Unwin, Antony. 2020. "Why Is Data Visualization Important? What Is Important in Data Visualization?" *Harvard Data Science Review.* January 31. https://hdsr.mitpress.mit.edu/pub/zok97i7p/release/4.

User Interviews. n.d. *Discovery Methods.* Accessed April 9, 2024. https://www.userinterviews.com/ux-research-field-guide-module/discovery-methods.

—. n.d. *Evaluative Methods.* Accessed April 9, 2024. https://www.userinterviews.com/ux-research-field-guide-module/evaluative-methods.

—. n.d. *How to Create a User Research Plan.* Accessed April 9, 2024. https://www.userinterviews.com/ux-research-field-guide-chapter/create-user-research-plan.

Vigliarolo, Brandon. 2024. "Copilot can't stop emitting violent, sexual images, says Microsoft whistleblower." *The Register.* 03 06. Accessed 2024. https://www.theregister.com/2024/03/06/microsoft_copilots_images/.

Violino, Bob. 2021. https://www.techtarget.com/searchenterpriseai/feature/Designing-and-building-artificial-intelligence-infrastructure.

WCAG 2.1. 2023. *Web Content Accessibility Guidelines 2.1.* September 21. https://www.w3.org/TR/WCAG21/.

Whitfield, Jayla. 2023. *How Health Tech Leaders Use AI to Combat Fraud.* May 22. Accessed April 8, 2024. https://govciomedia.com/how-health-tech-leaders-use-ai-to-combat-fraud-2/.

Young, Shalanda D. 2024. "Memorandum for the Heads of Executive Departments and Agencies: Advancing Governance, Innovation, and Risk Management for Agency Use of ." *White House.* March 28. https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf.

Appendix C. Glossary of Key Terms

Table 35. Glossary of Key Terms

Term	Details
Accountability in AI	The principle that AI systems' creators should be responsible for the outcomes of AI systems, including making amends for any harm caused.
AI	Artificial intelligence, referring to systems or machines that mimic human intelligence to perform tasks and can iteratively improve themselves based on the information they collect.
AI Ethics	The branch of ethics that examines the moral implications and societal impacts of artificial intelligence.
AI Playbook	A guide designed to support the implementation of AI by documenting guidelines and best practices, specifically within CMS in this context.
Bias in AI	The introduction of prejudiced assumptions and preferences into AI algorithms and data sets, which can lead to unfair outcomes or decisions.
CMS	Centers for Medicare & Medicaid Services, a federal agency that provides health coverage to over 160 million Americans through Medicare, Medicaid, the Child Health Insurance Program, and the Health Insurance Marketplace.
Data Governance	The process of managing the availability, usability, integrity, and security of the data in enterprise systems, based on internal data standards and policies that also control data usage.
Data Privacy	The aspect of information technology that deals with an organization's or individual's ability to determine what data in a computer system can be shared with third parties.
Data-Driven AI	AI that emphasizes the importance of data in enhancing technology's ability to learn from and augment human intelligence. It involves effective data understanding, governance, and a mindset that extends the value of data toward augmenting business processes through AI.
Drift	The change in a model's performance over time, as the data it was trained on no longer represents the current environment or conditions.
Exploratory Data Analysis (EDA)	An approach to analyzing data sets to summarize their main characteristics, often with visual methods, before making further assumptions or testing hypotheses.
Foundation Models	Models that are pre-trained on large datasets, used as a starting point for specific AI tasks. They form the basis for various AI applications.
Generative AI	AI techniques that can generate new content or data that is similar but not identical to the data on which it was trained, including text, images, and more.
Hallucination in AI	A phenomenon where AI models generate false or misleading information despite being presented with accurate data.
Human-Centric AI (HCAI)	AI that emphasizes the impact of AI technologies on individuals and society, prioritizing human well-being, needs, and goals.
Key Performance Indicator (KPI)	A measurable value that demonstrates how effectively an organization is achieving key business objectives.

Term	Details
Large Language Models (LLM)	A type of AI model designed to understand, generate, and interact using human language, capable of performing tasks like translation, question-answering, and content creation.
Machine Learning (ML)	A subset of AI where algorithms improve automatically through experience and the use of data.
Metrics	Quantitative measures used to track and assess the status of specific processes, projects, or activities.
Model Deployment	The process of integrating a machine learning model into an existing production environment to make practical and actionable predictions.
Natural Language Processing (NLP)	A branch of AI that enables computers to understand, interpret, and generate human language, allowing for the analysis and understanding of vast amounts of natural language data.
Predictive Analytics	The use of data, statistical algorithms, and ML techniques to identify the likelihood of future outcomes based on historical data.
Privacy in AI	The protection of personal data and information in the development and application of AI systems, ensuring data is used ethically and with consent.
Proof of Concept (PoC)	An early stage of project development that demonstrates the feasibility of an idea or technology to prove its potential application in solving a particular problem.
Reading Comprehension and Generation (RAG)	An AI technique used to enhance the understanding and generation of text by providing a data pool for reference, aiming to avoid issues like hallucination in language models.
Reliability in AI	The ability of AI systems to operate consistently under specific conditions, delivering accurate and dependable outcomes.
Responsible AI (RAI)	AI practices that uphold society's moral values, ensuring AI systems function fairly, as intended, and are accountable for their results. This includes adherence to principles like fairness, transparency, accountability, safety, privacy, and reliability.
Robustness in AI	The strength of an AI system to maintain its performance in the face of changing conditions or when dealing with unexpected or adversarial inputs.
Scalable and Interoperable AI	AI that ensures adoption within an organization is efficient, adaptable, and harmonious with existing workstreams, enabling AI-based solutions to grow and operate in sync with the agency's goals.
Stakeholders	Individuals or groups that have an interest in any decision or activity of an organization, including employees, customers, investors, and the community.
Transparency and Explainability in AI	The ability of AI systems to be understood and the processes and outcomes explained in human terms.
User Experience (UX) Design	The process of designing products, systems, or services with a focus on the quality and efficiency of the user's interaction with and experience of the product.
User Research	Research conducted to understand the behaviors, needs, and motivations of users through observation techniques, task analysis, and other feedback methodologies.

Appendix D. Acronyms

Table 36. Acronyms

Term	Full Form
AI	Artificial Intelligence
AIC	Akaike Information Criterion
AIE	AI Explorers
ARB	AI Review Board
ARC	AI Review Committee
ATO	Authorization to Operate
AUC	Area Under Curve
AWS	Amazon Web Service
CAIO	Chief AI Officer
CCI	Cross-Cutting Initiative
CDN	Content Delivery Network
CMAs	Computer Matching Agreements
CMS	Centers for Medicare & Medicaid Services
CO	Contracting Officer
COR	Contracting Officer Representative
CPI	Center for Program Integrity
CPU	Central Processing Unit
DSACMS	Digital Service at CMS
DevSecOps	Development Security Operations
DOE	Department of Energy
DOL	Department of Labor
DSACMS	Digital Service at CMS
DSH	Disproportionate Share Hospital
EDA	Exploratory Data Analysis
EQ	Emotional Intelligence
ETL	Extracting, transforming, and loading
FAIR	Facebook AI Research

Term	Full Form
FAR	Federal Acquisition Regulation
FAS	Feedback Analysis Solution
FDA	Food and Drug Administration
FISMA	Federal Information Security Management Act
GSA	General Services Administration
GATs	Generative AI Tools
GPU	Graphical Processing Unit
HAX	Human-AI Experience
HCAI	Human-Centric AI
HCD	Human-Centered Design
HHS	Department of Health and Human Services
IAM	Identity and Access Management
IDEs	Integrated Development Environments
IDR	Integrated Data Repository
ISPG	Information Security and Privacy Group
KMP	Knowledge Management Platform
KPIs	Key Performance Indicators
LLM	Large Language Models
MACFin	Medicaid And CHIP Financial
ML	Machine Learning
MLTRL	Machine Learning Technology Readiness Level
NAIIA	National Artificial Intelligence Initiative Act
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OAGM	Office of Acquisition and Grants Management
OCAIO	Office of the Chief Artificial Intelligence Officer

Term	Full Form
OHC	Office of Human Capital
OIT	Office of Information Technology
OSPO	Open Source Program Office
OSS	Open-Source Software
PI	Predictive Intelligence
PIAs	Privacy Impact Assessments
PII	Personally Identifiable Information
PoC	Proof of Concept
QSC	Quality Service Center
RAG	Reading Comprehension and Generation
RAI	Responsible AI
ROC	Receiving Operating Curve
SIA	Security Impact Analysis
SMEs	Subject Matter Experts
SORNs	System of Record Notices
TEC	Transparency, explainability and contestability
TF-IDF	Term frequency/inverse document frequency
TPU	Tensor Processing Unit
TRA	Technical Reference Architecture
TRB	Technical Review Board
TRL	Technology Readiness Level
UI	User interface
UX	User experience
VPC	Virtual Private Cloud
WCAG	Web Content Accessibility Guidelines
XAI	Explainable AI

Appendix E. Full Example of Mapped Governance Tasks

Table 37. Governance Task Mapping

	Research and Design	Model Development	Implementation and Scaling
Fairness and Impartiality	• T32. Data ontologies, inferences, and proxies • T40. AI system harms and impacts pre-assessment • T41. Algorithm risk assessment, • T43. AI system non-discrimination assurance • T44. AI system impact minimization	• T45. AI system impact metrics design • T46. AI system impact monitoring design	• T47. AI system impact monitoring, • T48. AI system impact health check
Transparency and Explainability	• T3. AI system use case • T19. Algorithm use case design • T44. AI system impact minimization, • T49. Transparency, explainability and contestability (TEC) expectation canvassing • T50. TEC Design	• T51. TEC monitoring design	• T52. TEC monitoring • T53. TEC health checks
Accountability and Compliance	• T17. Algorithm ID • T23. Algorithm version control design • T31. Data sourcing T54. Head of AI • T55. AI system owner • T56. Algorithm owner T60. Regulatory canvassing, • T61. Regulatory risks, constraints, design parameter analysis • T62. Regulatory design review	• T63. Compliance monitoring design • T64. Compliance health check design • T65 Compliance assessment	• T28. Algorithm version control • T66. Compliance monitoring • T67. Compliance health checks
Safety and Security	• T5. AI system operating environment • T20. Algorithm technical environment design • T21. Algorithm deployment metrics design • T40. AI system harms and impacts pre-assessment • T42. AI system health, safety and fundamental rights impact assessment • T44. AI system impact minimization	• T26. Algorithm verification and validation • T27. Algorithm approval • T59. AI governance integration [AI system governance should be integrated and aligned with other organizational governance processes]	• T29. Algorithm performance monitoring • T30. Algorithm health checks
Privacy	• T4. AI system user • T40. AI system harms and impacts pre-assessment • T44. AI system impact minimization	• T37. Data health check design • T57 AI development [design and implement appropriate workflows and processes for its AI-related data acquisition, permitting, and analytics operations, including approvals and signoffs]	• T39. Data health checks

	Research and Design	Model Development	Implementation and Scaling
Reliability and Robustness	• T18. Algorithm pre-design • T21. Algorithm deployment metrics design • T22. Algorithm operational metrics design • T23. Algorithm version control design • T33. Data pre-processing • T34. Data quality assurance • T40. AI system harms and impacts pre-assessment • T44. AI system impact minimization	• T24. Algorithm performance monitoring design • T25. Algorithm health checks design • T26. Algorithm verification and validation • T27. Algorithm approval • T35. Data quality metrics • T36. Data quality monitoring design • T37. Data health check design • T58. AI system operations	• T28. Algorithm version control • T38. Data quality monitoring • T39. Data health checks