



A GUIDE FOR
PUBLIC SECTOR ORGANIZATIONS

Get started with generative AI on AWS





Generative AI can open doors to endless possibilities increasing creativity, productivity, and progress within public sector organizations. With AWS, you can build and scale generative AI applications with security, privacy, and responsible AI built in from day one.

Jeff Kratz

AWS Vice President for Worldwide Public Sector Industry Sales

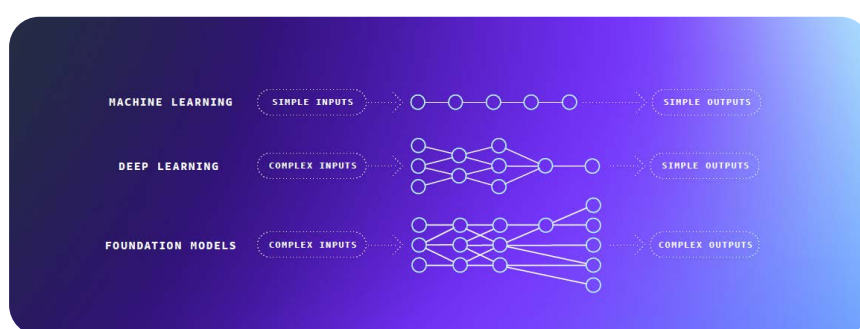
CONTENTS

| | | | |
|---|----|--|----|
| Foreword How the public sector can benefit from generative AI | 01 | Chapter 5 Managing AI model security, privacy, compliance, and bias | 24 |
| Chapter 1 Generative AI's transformative potential | 04 | Chapter 6 Building an AI-ready public sector workforce | 29 |
| Chapter 2 Building an effective AI strategy in the public sector | 09 | Chapter 7 A new era of building in the cloud with generative AI on AWS | 34 |
| Chapter 3 Understanding key AI model selection considerations | 14 | Next steps Beginning your generative AI journey with confidence | 39 |
| Chapter 4 Effective generative AI data management model inputs, prompt engineering, and output validation | 19 | Contributors | 40 |
| | | Glossary | 41 |

FOREWORD

Before your organization can fully unlock the business value of generative artificial intelligence (AI), it's important to have a fundamental understanding of how the technology works.

Generative AI is a term used to describe algorithms that can create new content and ideas, including conversations, stories, images, videos, and music. Generative AI is powered by large machine learning (ML) models that are pretrained on vast amounts of data. These are commonly known as foundation models (FMs).



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

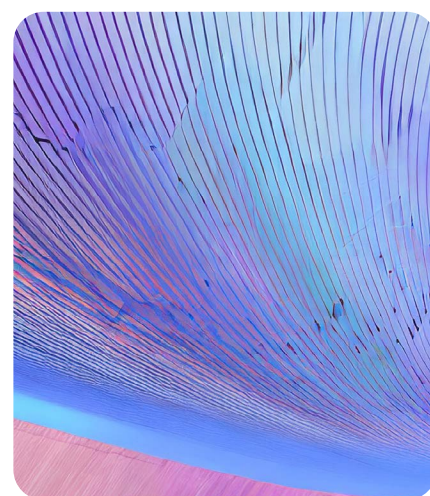
Contributors

Glossary

AI has recently captured widespread attention, and for good reason. This emerging technology has the potential to transform customer experiences, create new applications never seen before, help people reach new levels of productivity, and much more. In fact, according to [research from Goldman Sachs](#), generative AI could drive a seven percent (or almost \$7 trillion) increase in global GDP, and lift worker productivity by 1.5 percentage points over a 10-year period.

Applications and user experiences are poised to be reinvented with generative AI, and the public sector is no exception. Governments, education institutions, nonprofits, and health systems must constantly adapt and innovate to meet the changing needs of their constituents, students, beneficiaries, and patients. And new, simple-to-use generative AI tools such as Amazon Bedrock are democratizing the use of generative AI, making it accessible to anyone.

We have created this eBook to help guide public sector decision-makers as they navigate this new landscape. There are many areas for excitement with this rapidly-evolving technology, and many ways in which generative AI technologies can present new considerations for organizations in domains like regulatory compliance, data management, and access to technical expertise.





As this technology gains steam, we want to help public sector leaders use generative AI responsibly to its full potential. To get your thinking started, here are four key areas where we believe generative AI can make public sector organizations more efficient, responsive, and effective in achieving their missions.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



Empowering decision-makers with better data

Large language and foundation models can process and analyze vast amounts of data, oftentimes much faster and more efficiently than a human could. For example, government agencies could use generative AI to help query and analyze large amounts of information - such as public health data, economic indicators, and crime statistics - to identify patterns, trends, and correlations. This kind of analysis provides government officials with a comprehensive view of a particular topic, which in turn allows them to be more proactive in carrying out mission activities.



Personalized, convenient services

In today's digital world, citizens expect the same world-class technology experience from government that they expect when they log onto into their favorite streaming service, or shop online. Generative AI can help meet these expectations by tailoring services to individual citizens' needs across a breadth of platforms.



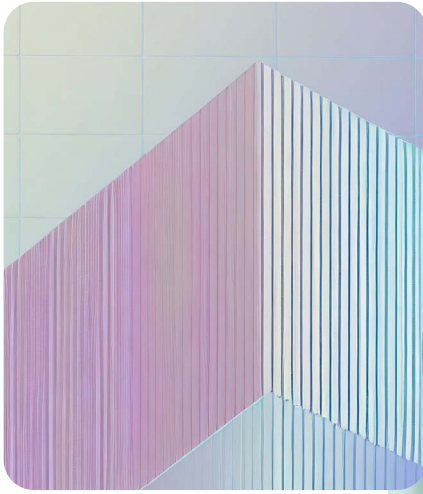
Accelerated research and discovery

Generative AI will impact how researchers work, helping them find information faster and freeing them up to pursue new experimental designs. It can help researchers accelerate their work by creating highly complex levels of analysis that would traditionally take days, months, or even years. Generative AI can also act as an idea generator, often surfacing overlooked connections that will help researchers consider more alternatives.



Improved productivity

According to [Accenture research](#), up to 40 percent of all working hours will soon be supported or augmented by language-based AI. Generative AI simplifies routine tasks, allowing humans to concentrate on more important mission goals. By automating some processes and decreasing reliance on manual labor, it not only frees up human effort for crucial missions but also results in substantial cost savings over time.



Looking forward: Understanding possibilities and requirements for generative AI in the public sector



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

While still in the early days, we believe that generative AI has the potential to greatly impact how public sector organizations operate, implement and manage services, and serve citizens and end-users. As part of Amazon's commitment to responsible AI usage, we have initiated a \$4 billion dollar strategic partnership with [Anthropic](#), one of the world's leading foundation model providers and a leading advocate for the responsible deployment of generative AI.

It's important to note that as generative AI evolves, public sector organizations' (or alternatively 'the public sector') should seek [to acknowledge and address](#) the many ethical considerations, data privacy issues, and potential for misuse by implementing strong governance frameworks and controls. When [used responsibly](#), this technology can revolutionize the way public sector organizations deliver on their missions, helping them make data-driven decisions faster, streamline processes, and enhance the experience of those they serve.

It is our hope that this eBook will help guide your organization towards effective, responsible generative AI applications that maximize mission value. As you progress through the chapters, we will help you build a strategic roadmap for considering generative AI applications for your organization. You'll explore essential quality metrics in generative AI, learn how to assess crucial technologies, establish compliance-conscious processes, and empower your workforce for success in the AI era.

Generative AI's transformative potential for the public sector

The advent of generative artificial intelligence (AI) can provide public sector organizations, and the businesses that serve them, with new opportunities to grow and deliver on their missions.

Teachers want to provide individualized learning opportunities to help students excel in the classroom and beyond. Healthcare providers are managing substantial workloads that include paperwork, diagnostic procedures, and a steady flow of patients seeking care. Judicial and law enforcement representatives aim to improve transparency in legal processes, to uphold their commitment to serving communities responsibly and equitably.

Resources to advance mission impact and meet the needs of stakeholders—time, finances, and personnel—remain consistently constrained. Generative AI technologies provide an opportunity to streamline processes and support decision-making so public sector leaders can scale to meet growing needs.

Generative AI can be used to automate administrative tasks or complement human inputs to accelerate processes. It can also enhance accessibility and support analysis to aid decision-making processes—letting leaders and staff focus on scaling their impact by dedicating their time to more mission-critical activities.

As public sector organizations weigh adopting AI tools, they should first consider:

- What are the core areas where generative AI can transform operations and services to fulfill stakeholder needs?
- How can organizations responsibly create value by selecting the use cases and tools that align with both their mission goals and the compliance, and constituent obligations placed on them?
- How can we integrate generative AI while considering the unique opportunities and obligations of each organization?

Use cases for generative AI in the public sector

Generative AI has a wide range of applications that can be tailored to address different needs within an organization.

It improves human-driven tasks like software development by providing top-notch suggestions and enabling collaborative interactions.

This accelerates workflows and boosts team productivity, as demonstrated by Accenture's [increased developer efficiency with Amazon CodeWhisperer](#).

Intelligent document processing is another common use case, where information is extracted from documents or used to understand content in a more contextually intelligent manner. It can be employed in tasks such as natural language understanding, summarization, language translation, and text generation, allowing for more sophisticated and context-aware document processing.

Use cases for generative AI in the public sector can go far beyond software development facilitation and intelligent document processing. Generative AI can serve as a transformative force through a myriad of use cases, enabling efficient, accessible communication, faster data processing, and richer context for decision-makers. These transformative impacts can be felt across all lines of business, including engineering, customer service, finance, and more.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



Generative AI can provide communications that are highly efficient and accessible. Diverse content formats, real-time transcription and translation services, and customized AI-generated emails and chatbots are all possibilities to elevate communications.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

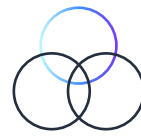
Contributors

Glossary



More efficient, accessible communication

An illustrative use case for enhancing accessibility involves indexing an FAQ or similar document and developing a conversational interface, such as a chatbot, enabling end users to inquire about the content. Generative AI can improve a team's ability to respond promptly by helping prioritize communications and offering suggested replies.



Richer context and support for human decision-makers

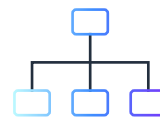
Generative AI supports pattern discoveries and aids in decision-making processes by performing summarization and creating a narrative from data points. For example, an application may use a conversational interface answering natural language questions using generative AI to write queries for data lakes and retrieve data answering the user's original question.



Faster data processing and analysis

Generative AI can help automate administrative tasks like data entry, data management, and information processing to streamline processes and optimize resources.

Generative AI can also enable organizations to process vast amounts of data and can assist with video analysis, document management, summarizing meetings, and extracting agenda items—ultimately expediting documentation, assisting with human-driven data analysis and decision-making processes, and helping your teams get more work done.



More efficient resource allocation and prediction

Generative AI can support planning resource allocations and identifying problem spots with features like summarization to reduce the time it takes teams to orient. This helps organizations like emergency responders make more effective use of resources and even predict and prevent failures before they happen.

This may translate to using generative AI to improve the real-world incident responses of public safety agencies by generating structured data from free-form notes gathered from citizens, dispatchers, and front-line responders during emergencies. This structured data can be subsequently used to provide critical insights and can, alongside human decision-makers, inform new policies and best practices.



Enhanced digital customization and personalization

Aspects of content creation and content customization can be automated using generative AI for intelligent document processing. Being able to deliver enhanced personalization of outputs can help support public sector workers with tasks like creating customized health recommendations, education plans, or social assistance programs.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



AI can provide recommendations and automate tasks, but human expertise is indispensable for considering nuances, ethics, and consequences—considerations that are particularly paramount in the public sector.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Next steps: Building strategic frameworks for generative AI projects in the public sector

When contemplating potential use cases and next steps, it's vital to recognize that the success of generative AI hinges on a collaborative approach that appropriately integrates human elements into AI-driven processes.

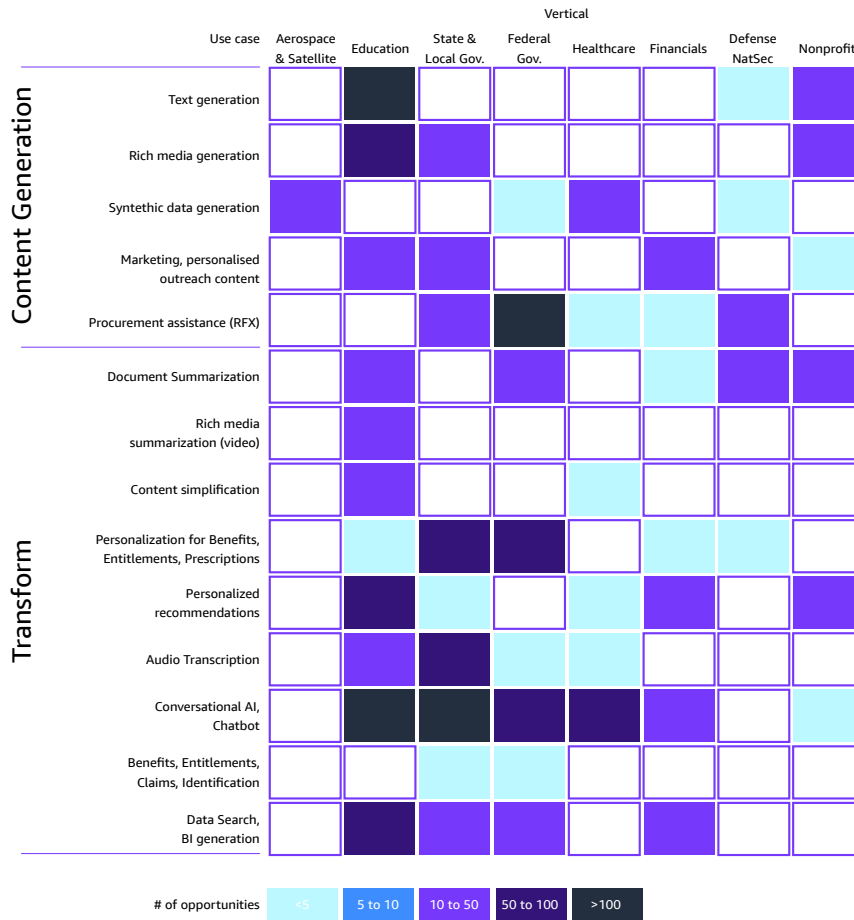
This is often referred to as a human-in-the-loop process and positions key users as responsible for verifying model outputs, making adjustments as needed, providing feedback, and retaining what is most useful while monitoring for bias. This concept is fundamental to many aspects of responsible AI usage, and so you will see it referenced throughout the eBook, with particular emphasis in Chapter 5 on addressing toxicity and bias, Chapter Four on data management practices, and Chapter 6 on building an AI-empowered workforce.

Beyond planning for potential bias in generative AI and how to incorporate the human element in AI strategies, it is important for public sector use cases in particular to build a roadmap that carefully considers any data compliance requirements and other relevant legal frameworks.

With all of this in mind, how can public sector organizations get started on their first generative AI project? And for leaders, what key considerations and compliance requirements should take the focus when overseeing generative AI projects?

In the next chapter we dive into what key considerations should be top priorities when starting out.

Generative AI use cases



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Generative AI customer story: BriBooks

BriBooks is the world's leading children's creative writing platform, assisting children of all ages in learning creative writing. They leveraged Amazon SageMaker to build and host a generative AI-powered assistant, BriBoo, meticulously designed to assist budding writers, regardless of age, in overcoming hurdles encountered while writing. BriBooks is customizing GPT-J 6 Billion parameters, an open-source large language model (LLM). Amazon SageMaker is used for fine-tuning the model using BriBooks' data. This helped the model provide focused recommendations based on an author's age group, genre of book, and style of writing on their platform. Using the auto-scaling feature of Amazon SageMaker inference, BriBooks has been able to handle requests from thousands of children. "Our generative AI-based writing assistant was trained on 1 billion words using Amazon SageMaker. It is trained so well that it has made writing fun, easy, and intuitive for kids."

- Rahul Ranjan, BriBooks Founder and Chief Business Officer

Important considerations for building an effective AI strategy in the public sector

Generative AI can effectively address a range of issues, automating time-consuming tasks and generating innovative solutions to maximize resources and support core public sector missions. Building an effective AI strategy in the public sector requires careful management of many considerations.

Questions to consider:

- What are your priority use-cases? What kinds of results would justify further investment?
- What is the user experience you want to create? What data will your application interact with?
- What regulations around security, compliance, and privacy obligations does your organization need to adhere to?
- What are your selection criteria for models?
- How can you architect for the agility to evolve your application with emerging technologies?

Understanding your generative AI use case

- How will AI technologies help your organization better deliver its mission?
- How can AI be used to elevate the power of human decision-makers, not replace them?

Public sector organizations are unlocking new abilities to scale their impacts with generative AI. There are text-based applications such as chatbots for connecting with stakeholders, and multi-modal applications like creating meeting summaries from video recordings or enhancing medical diagnostic abilities with patient chart processing. While it is exciting to think about the most advanced use cases, even simple AI tools can have big impacts on an organization.

When considering how and where to apply generative AI, there are a range of techniques to customize the outputs, which vary in cost and complexity.

Applications that interact with your data will require more engineering and compliance attention. These projects require appropriate guardrails to mitigate risks like toxicity, bias, and inaccuracy in outputs. If your models are being further customized with new data, machine learning operations (MLOps) workflows become essential for model maintenance and monitoring.

Addressing these considerations can be complex and involve multiple stakeholders. For public sector organizations, this makes it especially important to right-size AI projects at the beginning to deliver high-quality tools within budget.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

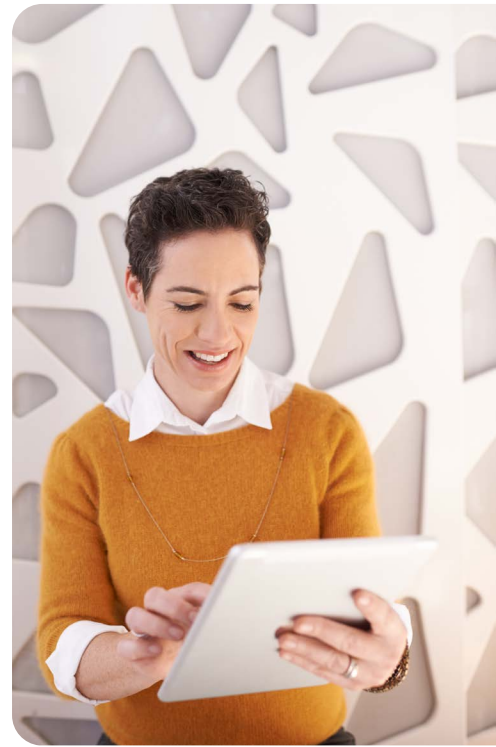
Glossary



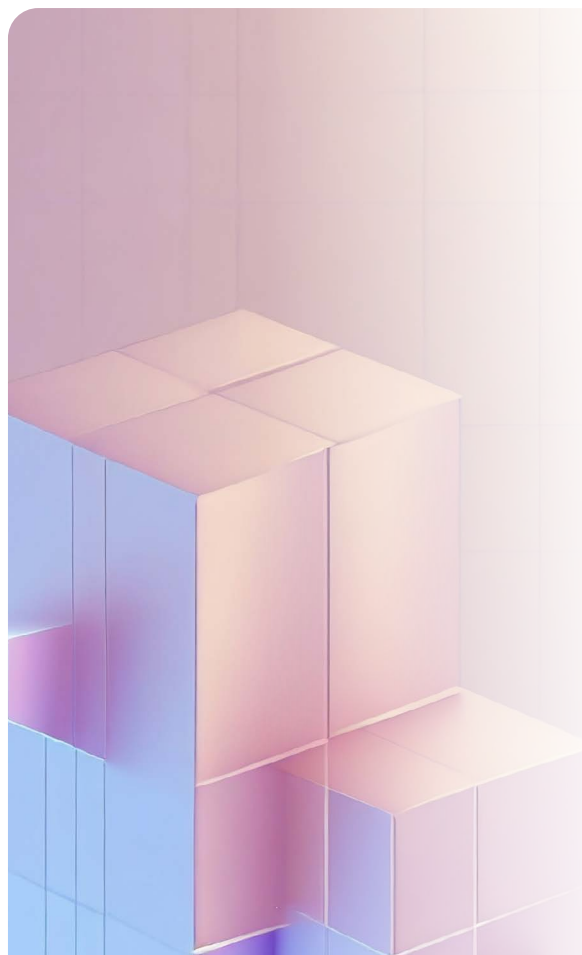
Data sources, quality, and bias

- What specific knowledge does your AI system require? Is the data your application requires available for consumption?
- How do you mitigate biases in the data used to train AI models?

To make the most effective use of generative AI requires a clear understanding of your organization's data sources and desired outputs. Define the collection of knowledge your AI system requires, such as specific data points for a chatbot, event summaries, or content alignment with learning standards.



The quality of data collected, which is used to train and fine-tune machine learning (ML) models, is one of the strongest factors contributing to project success.



"Training models with poor data quality will lead to poor results," says Werner Vogels, VP & CTO at Amazon.com. "You will need to filter out bias, hate speech, and toxicity. You'll need to make sure that the data is free of personally identifiable information (PII) or sensitive data...and make sure your data is deduplicated, balanced, and doesn't lead to oversampling."

Ensuring data is relevant, complete, and free from bias can be a time-consuming process—but it is vital in the public sector, where AI is used to serve the public.

To aid in this process, Amazon SageMaker has features that can help complete each step of the data preparation workflow, including data selection, cleansing, exploration, bias detection, and visualization from a single visual interface in minutes. You can read more about addressing biases and other issues in Chapter 5, including learning about specific tools to help with bias detection in data.

Compliance, security, and privacy

- Which security, compliance, and privacy regulations does your organization need to follow regarding the data sources and types your application will gather?

[Many governments are in the early stages of regulating generative AI](#), and it is crucial to remain in compliance with anticipated regulations. Creating foundational organization-wide practices that ensure high levels of security, privacy, and anti-discrimination should be the focus.

Additionally, consider any regulatory, security, or compliance requirements associated with various data sources, especially if you're handling sensitive information like personally identifiable information (PII) or healthcare data. Architects should work closely with their legal teams to ensure non-technical requirements, such as data lineage, are considered as well.

United States federal laws, such as HIPAA for healthcare, FERPA for education, and FedRAMP for federal information, dictate specific requirements that must be met in the US, and for any application it should be assumed that generative AI needs to adhere to these rules like any other project.

Managing compliance, security, and privacy needs in the public sector will require significant resource investment, but it is worthwhile because of a growing concern about the risk of data leakage through AI models. AWS is committed to never sharing your data with third-party model providers, nor using it to train first party models on Amazon Bedrock.

Model selection, cost drivers, and change readiness

- How can public sector organizations determine the true cost of AI models, and what methods can be used to optimize cost-effectiveness?
- What factors should public sector organizations consider when deciding between first party-hosted and third party-hosted foundation models?

On a high level, your organization will need to decide on whether to host models using a managed hosting service like Amazon SageMaker, or consume models as API calls from a fully managed service like Amazon Bedrock, or a combination of approaches. Consider that you may need to use multiple models to address a range of different use-cases across your organization.

Teams can choose from commercial foundational models like those available from AI21, Anthropic, Cohere, Meta, and Stability.

When making your selections, consider both the upfront and ongoing costs as you select the models that will power your application. If your application requires a customized model, the availability and selection of tools for customization may vary from model to model. This can impact the level of effort required of your team, so it is an important part of your model selection criteria.

Have your organization's experts determine if any of the available foundation models meet your requirements as-is, and review factors such as cost, quality, and the length of context windows. Multiple approaches can be used to integrate data into your AI system, including prompting, fine-tuning, and retrieval augmented generation (RAG).

The decision to self-host a model or consume it as an API determines the cost to host and query your models. When you use a fully managed service like Amazon Bedrock, you can pay for the tokens that you send and receive as inputs and outputs to the model or purchase provisioned capacity for your customized models or production workloads.

Your organization should also consider tuning prompts with related components like prompt templates and output processors, and evaluate effects on response speed, hosting expenses, and output format suitability—all of which can evolve as you experiment and iterate with different models.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7


A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



While it may be tempting to view AI as a one-time project, it's important to recognize that it's an ongoing journey that requires change-ready solutions. Plan strategically for ongoing investment in responsible AI needs such as compliance, security, privacy, accuracy, and bias mitigation.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

If your organization needs to customize the model outputs, this may add new continuous improvement workflows around your models, which can affect hosting and staffing requirements. Combining the model with a data source, will bring additional data engineering requirements as well. You can learn more about model customization approaches in Chapter 5, and their impacts on team requirements in Chapter 6.

As generative AI continues to advance, more complex services will become available. An example is Amazon HealthScribe, a HIPAA-eligible service that empowers healthcare software vendors to create clinical applications. It automatically generates clinical notes from patient information. Another example is Amazon CodeWhisperer, an AI developer companion that suggests new code based on natural language prompts.

The ability to quickly train new ML models themselves is also rapidly advancing thanks to new technological developments. The Amazon Trainium AI Accelerator efficiently trains models for natural language processing, computer vision, and recommender models. It's optimized for various applications like text summarization, code generation,

question answering, image and video generation, recommendation, and fraud detection. For hosting models, the AWS Inferentia accelerator is optimized to deploy increasingly complex models, such as large language models (LLM) and vision transformers, at scale.

Ultimately, the generative AI landscape is evolving rapidly, particularly in the open-source community. Change should be anticipated throughout an organization's AI journey, and it is crucial to engineer systems that can adapt.

Use a common library or build your own abstractions so that teams can evolve your application to adopt new models and other technologies that bring new and improved capabilities. Your team should plan to integrate new models and techniques, as generative AI is constantly evolving.

Developing services and tools powered by generative AI comes with both initial development costs and ongoing support requirements.

Generative AI customer story: SciSpace, formerly Typeset

SciSpace, formerly Typeset is an EdTech company with a focus on simplifying access and consumption of research papers. They focus on innovation via AI/ML and have built several generative AI models for summarizing, paraphrasing, and generating Q&A content against the research paper content. All of these are large language models requiring GPU-based machines for inferencing.

Next steps: Understanding generative AI model selection considerations for the public sector

Once you have developed a strong understanding of the opportunities and obligations shaping your organization's approach to generative AI and determined priority use cases, you're ready to explore more technical questions. One of the key areas of technical consideration for generative AI projects is what machine learning model(s) to use.

In the next chapter we look at the fundamentals of foundation models, large language models (LLM), and how they can be evaluated. For decision makers, we will also uncover key cost drivers, compliance impacts, and strategic decision points.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



Understanding key AI model selection considerations for the public sector

Navigating the AI landscape can be difficult and complex for decision-makers in the public sector. Considerations such as data source management, organizational needs, and compliance requirements are fundamental in building an effective AI strategy.



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Technical decisions, such as selecting and validating foundation AI models, are made with this context and provide the next step in your journey toward defining a productive AI landscape for your organization.

You should start by identifying organizational or business-level requirements like licensing and determining if the data used to train the model meets your compliance requirements. Once you know the non-technical requirements, you must define evaluation criteria supporting objective comparison of model outputs.

You should consider cost drivers, like hosting requirements, to choose a model that meets your needs at an optimal cost. Finally, architect for change. You may want to update to a different model as new models become available.



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Understanding foundation AI

The realm of AI is experiencing a pivotal shift, thanks to the advent of foundation models (FM) — these behemoths of ML have emerged as versatile tools, learning from vast amounts of data to cater to a diverse array of downstream tasks such as document summarization, extracting structured data from free-form text, or providing conversational interfaces to data.

At their core, foundation models are machine learning models designed to ingest and learn from colossal datasets. Their training mechanism is rooted in self-supervision, enabling them to grasp general representations across various modalities—text, visuals, or voice.

models for the public sector

The capability of these models to grasp and internalize universal constructs about domains like language or visual cues paves the way for them to adapt quickly to fresh challenges. Instead of starting the learning process from the ground up, they can take advantage of their preexisting knowledge, dramatically reducing training durations and extensive data prerequisites.

Foundation models such as generative pre-trained transformers (GPT) have radically changed how organizations use AI capabilities now and how they imagine using them in the future. Hosting services for these models, like Amazon Bedrock and Amazon SageMaker on AWS, simplify the technical knowledge previously required to interact with AI by providing a managed service that makes foundation models available with an API call.

Pre-built foundation models reduce the need for and sometimes eliminate model training and preparation. Amazon Bedrock provides a fully managed hosting and customization solution for foundation models to accelerate development, while Amazon SageMaker JumpStart enables you to easily deploy and customize foundation models. This simplicity and acceleration make AI solutions more accessible and rapidly deployable than ever before.

Use cases in the public sector are everywhere, and many sectors are embracing AI capabilities in foundation models to help improve education, revolutionize healthcare practices, save taxpayers money, and drive innovation and creativity across all industries.

AI models are advancing rapidly. Foundation models represent a huge advance in accelerating the development of AI solutions, and you should follow best practices that allow new models and functionality to easily integrate with your AI solutions.

On AWS, there is a vast range of possibilities for using existing models or tailoring them to public sector needs. You can use fully managed foundation models from AWS and leading AI companies on Amazon Bedrock, where you can privately customize those FMs with your own data through a visual interface without writing any code. Amazon SageMaker lets you deploy published models yourself using managed notebooks and a point-and-click UI. You can also easily use most open-source models available on HuggingFace using Amazon SageMaker JumpStart or the SageMaker SDK HuggingFace integration. Both Amazon Bedrock and SageMaker also offer several built-in capabilities to support security and privacy concerns. To further help organizations meet compliance requirements, Amazon Bedrock has achieved Health Insurance Portability and Accountability Act (HIPAA) eligibility and general data protection regulation (GDPR) compliance.

Evaluating and validating model suitability in the public sector

Evaluation criteria should be determined by your use case. For example, an application that presents educational information to students needs to be factually accurate and free from harmful bias. On the other hand, a second application might create fictional content where ingenuity and variety in output are a welcome source of creativity. Clear evaluation criteria needs to be defined to fit the use case your application seeks to solve with generative AI.

When evaluating AI models, understanding model capability, performance, and cost are vital to consider in deciding which model to select. These factors are relative to the nature of the task at hand. Additionally, more widely applicable factors should also be considered, including:

- **Authenticity** – Models must deliver appropriate levels of factual correctness and inference accuracy for the task they perform.
- **Efficiency** – While some use cases can afford slower models, some tasks, such as those with more interactive needs, demand faster model performance. This can also have cost implications with a trade-off between model complexity and output quality.
- **Lack of bias** – Models should account for bias and provide fair and equitable results within the context of their task, especially concerning factors around gender, race, and other protected classes and groups.
- **Backtracking** – Understanding where the model's conclusions come from allows for better validation and accountability.
- **Understanding context** – When users consult models, models must offer answers tailored to the specific context in which the question is asked.
- **Context window** – A model's context window determines the prompt size (the defined value of variables like text, numbers, characters) a model can handle. The context window must allow for a suitable prompt size to support your tasks.

If a customized model is needed, training costs will vary depending on the amount of training needed, the training method used—managed or unmanaged—and the associated pricing model. Data size and model choice will also influence cost.

Operating costs for a model will vary with platform—Amazon Bedrock ([pricing](#)), Amazon SageMaker ([pricing](#)), or self-hosted. Instance size and model will directly affect operating costs. For more complex scenarios, you should consider requirements for batch jobs or real-time endpoints, as these features are priced per hour depending on the selected instance size.

In addition to the model performance metrics, you should consider how quickly the model returns a response, hosting cost, and whether the model can be prompted to output in a format that suits the assessed application. Each of these factors may change as you adjust the model, try new models, and continue to iterate.

Generative AI customer story: Bedfordshire Police Department

Bedfordshire Police Department is using generative AI to help redact documents for use in court. An AWS partner created a platform that helps the department quickly redact documents. By doing this, their police department is able to have five more police officers in the field instead of behind desks, making the community safer.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



Model behavior with your data or use case is a critical aspect of evaluation. How a model's outputs can be guided with prompting and available options to further customize the model are often key selection criteria, especially for tasks that involve sensitive data or critical decision support.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

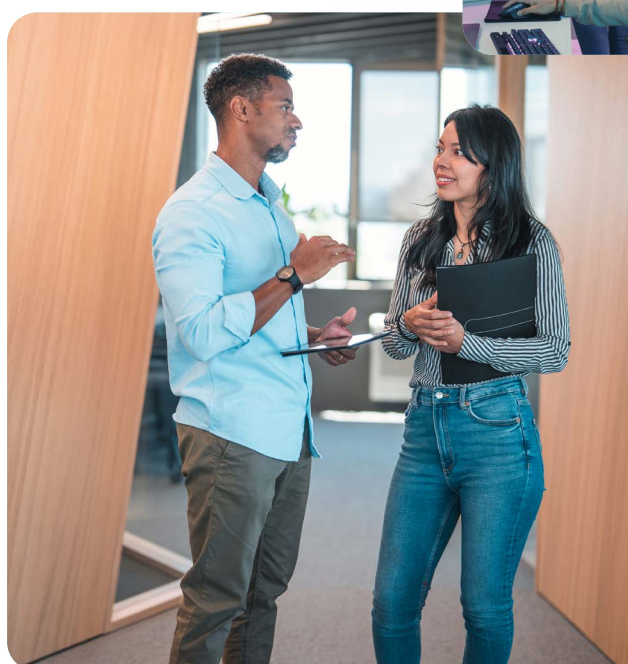
A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Modifying usage and customization

Prompt engineering involves selecting appropriate words, providing context, and shaping the prompts so the model produces the desired response. Finding the best prompts is often an iterative process, and prompts will vary from model to model. Prompt engineering is the easiest entry point to customizing the model behavior to your use case, and you will see the practice referred to in later chapters on data management, and AI-enabling your workforce.

Retrieval augmented generation (RAG) is a technique where you retrieve data from outside of the model and use it in your prompt to generate a response. You can use RAG to improve response quality by grounding the model on external sources of knowledge to enhance the model's internal representation of information. Implementing RAG ensures that the model uses the most current and reliable facts. It also provides access to the model's sources, ensuring that its claims can be checked for accuracy and ultimately trusted. The implications of this approach are explored in more depth when we look at data management in Chapter 4.

Fine-tuning can be used when your application requires domain specificity that cannot be provided through approaches like RAG. Fine-tuning is used to customize a pre-trained model further using a smaller, task-specific dataset. This process adjusts the model's parameters, enhancing its performance for a specific task across all prompts. A combination of approaches may be used to select data for additional context, or even customize your model. These approaches vary in cost and complexity.

Model fine-tuning allows you to make the best use of pretrained foundation models and adapt them to specific tasks using data limited to a particular subject area. Fine-tuning is especially relevant to the public sector, where models may prioritize specific contexts and interest areas, such as K-12 instruction in education, clinical diagnosis in healthcare, or constituency voting in government, at the expense of general model applicability.

Foundation models represent a huge leap in off-the-shelf functionality for AI, because you can use techniques like RAG and fine-tuning to quickly use them with your data. Outputs vary and their suitability for each use case needs to be evaluated on a case-by-case basis. A consistent evaluation process with a clear understanding of your use cases will help you quickly determine the suitability of any model.

Next steps: Building effective AI data management practices for public sector organizations

As you continue developing an AI strategy for your organization, an important next step is learning more about how data management requirements intersect with generative AI needs for model training, output validation, and prompt engineering.

In the next chapter, we look at these questions with a particular focus on how considerations like changing regulatory landscapes will affect generative AI usage by public sector organizations. In the process, we examine how managing data effectively with generative AI is not a static set of decisions, but an ongoing process of strategic iteration and improvement.



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

CHAPTER 4

Effective generative AI data management in the public sector: Model inputs, prompt engineering, and output validation

As the public sector looks to the transformative potential of generative AI, effectively harnessing these tools will require a keen focus on effective data management.

Effective data management fundamentals like classification, storage, movement, protection, and governance can prepare organizations to act effectively as the technical and regulatory landscape evolves. Generative AI tools bring new aspects to data management, highlighting the importance of addressing human impacts from errors or biases in outputs. This requires enhanced collaboration among teams, including data science, legal, and information security, to tackle complex issues.

Establishing what “good” and “bad” outputs mean for a particular generative AI project is an essential baseline for evaluating the suitability of different approaches to managing AI data flows. Developing metrics and tests for output quality are important steps, allowing teams across an organization to apply an aligned philosophy of continuous improvement in terms of quality, compliance, and resources.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

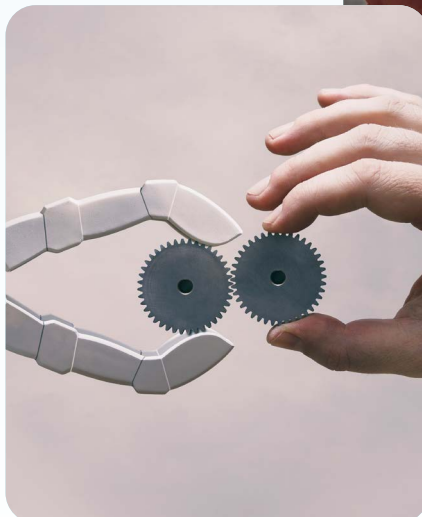
A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary





Generative AI customer story: AWS Cloud Innovation Center (CIC) at Cal Poly

The AWS Cloud Innovation Center (CIC) at California Polytechnic State University (Cal Poly) enabled students and staff to build a secure, cloud-based tool that makes cybersecurity recommendations for public sector organizations. Working with the San Diego Cyber Center of Excellence and Regional Cyber Lab, a “My e-CISO” generative AI-enabled chatbot was developed which provides recommendations after the user asks a series of open-ended questions. Inputs and chat responses are securely stored and accessible to only those who have been granted access.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Navigating the regulatory landscape: Preparing for AI model and data source changes

A common question public sector leaders have is, “How do I work with generative AI tools responsibly?” The question is especially imperative for leaders of organizations that work with sensitive data like personal health information (PHI) or are subject to compliance regulations like the Federal Risk and Authorization Management Program (FedRAMP). There are various technologies and services available to help organizations support their compliance needs, and we examine the topic from different angles throughout this eBook—but leaders should consider that what compliance for generative AI tools means today is likely to change in the coming months and years.

Governments are creating wide-ranging laws governing AI tools and the data that powers them. The European Union has recently released its landmark AI regulation action, and in the United States, the White House has released an Executive Order on AI regulation.

This is expected to inspire and inform other countries' approaches as well—like how regulations for PII compliance changed after the introduction of the General Data Protection Regulations (GDPR). The introduction of AI-specific regulations may have wide-ranging impacts, such as making previously legal sources of data no longer permissible, which might require retraining AI models that rely on noncompliant data. Explainability is also a topic with increasing scrutiny, with growing consideration of a model's ability to provide an explanation of how it arrived at its outputs, such as linking to a source legal document.

Understandably, organizations may experience analysis paralysis as they consider this future. But leaders who design their generative AI projects with a plan for constant improvement and iteration will be well positioned to support evolving compliance needs and technological advances.

So how do public sector leaders move forward with the generative AI projects that can serve them today?

First, leaders need to define measures of success—and failure—and build observability of these metrics into the entire lifecycle of a generative AI application.

Defining “good” AI outputs: Humans in the loop and building guardrails for data quality

Defining what constitutes “good” in AI-generated outputs varies significantly depending on the desired outcomes, which may range from increased accuracy in estimating resource needs, enhanced creativity or faster response times for human decision-makers. The definition of “bad” is similarly variable but needs to be driven by a consideration of the human impact of potential issues in outputs such as hallucinations, toxicity, and bias.

In public sector work, the human impact of potential negative outputs like bias or inaccuracy is often quite high. For instance, in a medical scenario where the objective of the generative AI tool is to identify medical issues from clinical notes, the risk of misdiagnosis is significant. In such a case, letting an AI model make autonomous decisions itself could lead to catastrophic errors. For this use case, the focus must be on AI supporting and enabling human reviewers in decision-making.

Public sector use cases are also often particularly sensitive to toxic patterns in outputs, such as inflammatory language, which can necessitate proactive monitoring and analysis of model outputs for drift in quality over time. We delve into the concepts of toxicity and bias in more detail in Chapter Five.

The pressures of avoiding bias and toxicity bring into focus the necessity of aligning AI output metrics with your mission imperatives. It is crucial to carefully define good and bad outputs before making key technological decisions, such as selecting the model(s) and customization approaches that will power your generative AI application. A universal software concept applies here: write tests first!



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



Customizing foundation models and integrating your data

Achieving the desired level of output quality from AI models usually requires implementing one or more methods of inputting new data into the system. As desired model skillsets or output formats get more complex, the data inputs and management needs of your organization at various stages will become more advanced and in-depth.

Correctly scoping the needs of an AI application will help with identifying the appropriate level of customization necessary for the outputs. Employing simpler customization approaches when possible can significantly reduce initial and ongoing data management needs, especially in terms of governance, stewardship, and compliance.

The simplest and most familiar data input approach is prompt engineering, which is the use of templated and user-generated data prompts at runtime to enable in-context learning, allowing an AI model to temporarily learn from prompt data without affecting the model itself. Prompt engineering can be employed by non-technical users through a dialog interface or by developers through basic API integration, if appropriate foundation models are selected.

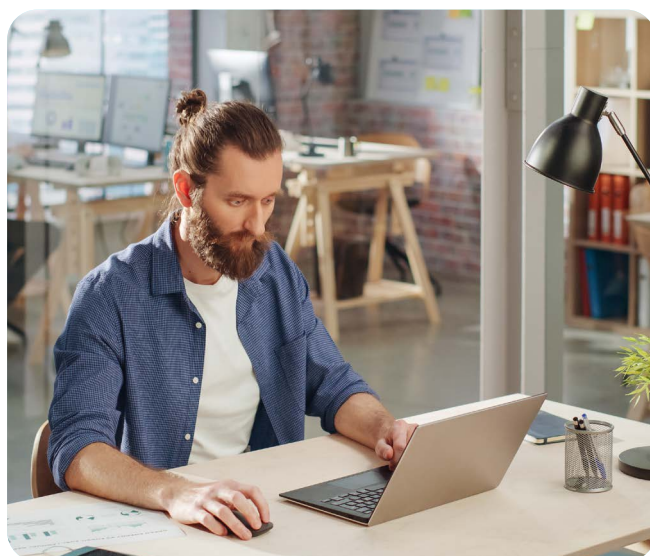
[Retrieval augmented generation \(RAG\)](#) is a customization approach that utilizes ongoing retrieval of information from data sources you provide, allowing your foundation models to generate responses based on the curated, validated, and

customized solutions while maintaining data relevance and optimizing costs. Successfully delivering RAG solutions requires organizations to have strong development capabilities and mature engineering practices, which we'll cover in the next chapter.

AWS customers can leverage fully managed [AI agents on Amazon Bedrock](#) to streamline complex business operations. With automatic prompt creation, users can input their desired task or query, and Bedrock automatically generates the necessary prompts and templates. Agents are able to securely access company data sources for RAG, orchestrate interactions between foundation models, data sources, software applications, and user conversations, and automatically call APIs to take actions. Developers can automate workflows and optimize processes efficiently by integrating agents, accelerating delivery of generative AI applications and saving weeks of development effort.

Performing well on tasks with a high degree of domain specificity may require fine-tuning foundation models, using techniques like reinforcement learning from human feedback (RLHF) to provide specific data that modifies the model itself, resulting in your own custom version. This requires developers skilled in training and tuning machine learning models, with strong data and DevOps capabilities supporting them with ongoing deployment, interfacing, and management of hosts and services.

While we anticipate a very low percentage of customers will train foundation models from scratch, Amazon SageMaker provides an end-to-end machine learning platform to build, train and deploy machine learning models. Training new models requires access to large amounts of training data, strong ML expertise, and extensive development and data engineering resources. The needs of many generative AI applications can be met without going to this level of customization by using foundation models on [Amazon Bedrock](#) or those available on [Amazon SageMaker JumpStart](#).



Supporting validity in model outputs: Tools and techniques for public sector applications

Maintaining the quality of AI outputs and supporting compliance with regulations requires vigilance. Tools like [Amazon SageMaker Model Monitor](#) can aid in monitoring inputs and outputs, providing insights into potential issues such as toxicity and bias. Regularly sampling AI outputs helps detect unwanted developments and responses or patterns indicating adversarial user behavior. Amazon SageMaker Clarify also provides a set of tools that can be used to enhance the explainability of outputs, helping identify the sources of any issues that do arise.

Continuous improvement and aligning to compliance needs should be at the forefront of your AI strategy. This involves not only monitoring outputs but also managing inputs as you iterate on model selection, prompt templates, and other specific customizations. It requires keeping the data your application uses up to date, improving your model with new data from user feedback, applying new techniques, and using new models. As you iterate, you need to continuously guard against new bias with continuous monitoring and measurement.

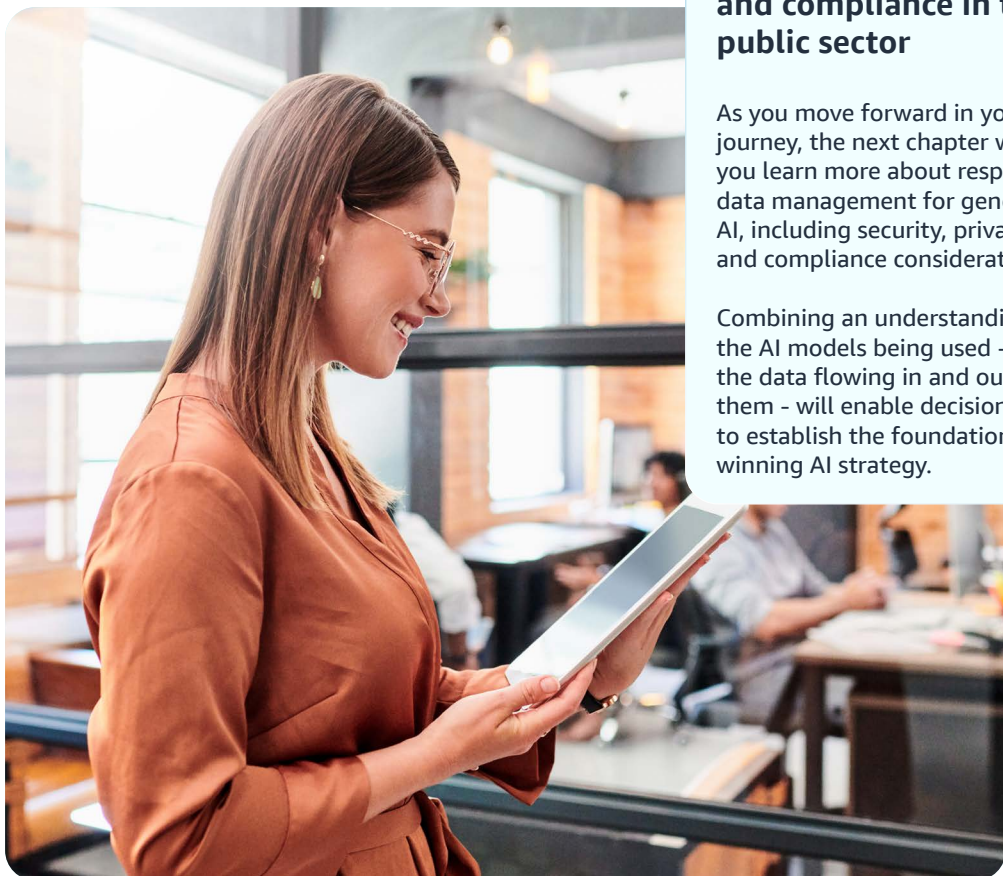
It's important to match the intensity of your efforts to the potential damage of input/output errors, particularly in public sector applications where your application may have a broad and sweeping impact on populations and carry a risk. It's important to implement positive controls to manage the risk of accuracy, bias, and toxicity.

It's wise to create human touch points with your AI applications. Human help in training, prompting, and reviewing outputs is a crucial component of creating a relevant, "good" output. To get the most out of AI applications, consider that AI models can help teams make better decisions by providing better support, context, and connections. You can use generative AI to augment your human workforce with better support and context so they can make connections faster and reduce undifferentiated heavy lifting in their roles.

Next steps: Managing AI security, privacy, bias, and compliance in the public sector

As you move forward in your AI journey, the next chapter will help you learn more about responsible data management for generative AI, including security, privacy, bias, and compliance considerations.

Combining an understanding of the AI models being used - and the data flowing in and out of them - will enable decision makers to establish the foundation for a winning AI strategy.



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

CHAPTER 5

AI model security, privacy, compliance, and bias in the public sector

At the core of any generative AI solution is the machine learning (ML) model, which drives the generative process, quality of outputs and prompts, and retrieved data.



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Your choice to customize a foundation model or use it as-is along with your decisions about how you will host and maintain this model influences many of the development and operating cost considerations. It also significantly impacts requirements around data security, privacy, compliance, and bias avoidance.

Considering these elements when assessing and selecting a model that satisfies requirements is especially pronounced in the public sector, where technology decisions can have far-reaching societal implications.



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

To adopt AI responsibly and ethically, many teams across your organization must continuously work together to ensure that security, privacy, accuracy, and compliance obligations are being met. Everyone shares responsibility in this process, from data science and DevOps teams to security, legal, and communications departments, and outwards to any vendors and contractors involved.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



Examining the AWS Shared Responsibility Model for security and privacy

Given the computing requirements for generative AI, most applications will involve working with infrastructure providers of some sort. AWS services are built with security and privacy at the core, helping you build a safe, compliant AI environment. The [AWS Shared Responsibility Model](#) defines how responsibility for compliance involving AWS components is divided between AWS as a provider and you as the customer.

AWS is responsible for the “security of the cloud”—securing the infrastructure that runs all the services offered to its customers. Third-party auditors regularly test the effectiveness of our security as part of the [AWS compliance programs](#). This includes the hardware infrastructure and extends to service functionality and model security for AI products like Amazon SageMaker and Amazon Bedrock. For example, SageMaker protects notebook instances, model-building data, and model artifacts with at-rest encryption.

The customer is responsible for “security in the cloud”—securing the data and applications you run on top of that infrastructure. This includes managing data classification and access control, your organization’s requirements, and applicable laws and regulatory compliance. Implementing access controls, protection, and policies for your data and applications in AWS is a critical and mandatory step in ensuring a protected environment is established to support your AI solutions.

Building for trust and compliance

AI running at scale must do no harm. An algorithm that runs unchecked can cause systematic damage, which undoes the very thing it was designed for: serving humanity. In the public sector, where the focus is often on educating, protecting, and caring for the population, it is essential that results can be trusted and are compliant with all applicable regulations.

AWS is working to follow evolving compliance standards worldwide. Work is being done in the European Union on defining and regulating AI capabilities to mitigate risk and ensure the safe use of AI that observes the rights of the human beings that use it, or are affected by it.

Generative AI is meant to provide support so that experts can focus on what they do best. Doctors may use the tools to summarize patient notes, while teachers might align content to learning standards. When the tools work correctly, the doctor has more time to care for their patients, and the teacher has more time to “see light bulbs turn on” in the eyes of their students. There is a shared, human-driven process for developers and users to verify the model’s output, adjust it as needed, provide feedback, and retain what is useful.

Generative AI customer story: Swindon Borough Council

Swindon Borough Council, a local authority in England, is using Amazon Bedrock to make government information more accessible and comprehensible for the learning disability community. They created “Simply Readable,” an innovative solution that converts complex documents into an accessible easy read format—enabling a community where no one is excluded or disadvantaged. Swindon have just made this solution available open-source, license-free, worldwide.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Understanding model accuracy and bias

Generative AI is designed to produce human-like content and interactions, and we have seen radical advancements in capability and efficiency. However, the simulation of human capabilities will always come with concerns about accuracy and reliability. Addressing these concerns motivates a strategy to mitigate risks of inaccurate, misleading, or even false content.

Models trained on vast datasets can learn and predict patterns using statistics. They are not necessarily able to analytically “understand” content like a human would.

AI models learn to produce patterns of communication and responses based on training input and available data. Because of these factors, models can easily create content that seems coherent but is factually incorrect. For example, the model might generate a fake historical event with realistic-sounding details that never occurred. From the model's perspective, it simply predicts words that fit the pattern without the benefit of human “fact-checking” ability.

The massive volume of training data enables models to fabricate details that sound convincing to humans. They learn to produce sentences, facts, and stories that seem believable to us and project confidence, while communicating facts that may or may not be accurate.

Bias can occur when an AI model's underlying data does not reflect the model's intended learnings or accurately represent the appropriate population of interest. This could result from training data that does not accurately represent the target audience.

Toxicity within the data is also a concern. If the training data contains inappropriate content, including content that's racist, misogynistic, or otherwise discriminatory, output from the model may also contain toxic content.

While such data might reflect aspects of currently-held human viewpoints, it is not information that a public sector organization wishes to propagate out into the world through an AI model. Bias and toxicity often occur due to the underlying data collected and used to train the model. This can go unnoticed in a dataset, especially if minority or marginalized groups of people are left out of or misrepresented in the data.

Addressing AI accuracy and bias responsibly

Bias and inaccuracy, when combined with automation, can have catastrophic effects in the public sector. To properly evaluate these models, you need testing methods that directly measure performance of the task, not just generic benchmarks. The key is rigorously testing models to stress their capabilities and reveal inaccuracy or bias. Researchers are developing new techniques and evaluation practices to better measure and improve accuracy and bias detection, but processes can be put in place to mitigate the risk, including model measuring, human-in-the-loop review, and measuring dataset bias. AWS tools such as ML Governance and Model Monitor for Amazon SageMaker, and Amazon Augmented AI can provide support for these processes.

Bias in AI is not just a technological challenge; it is a systemic and societal one. At AWS, we are committed to providing tools that not only detect but also prevent bias from taking root. Amazon SageMaker Clarify, for example, offers a suite of functionalities to detect biases. But it is the combination of our tools and the expertise of IT teams in deploying them that truly makes a difference, as you will see in the following chapter on building AI-powered organizations.



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Integrating generative AI into the public sector holds great promise but also raises important challenges around ethics, accuracy, and responsible use. While AI models exhibit impressive fluency, organizations cannot blindly trust their outputs. Awareness, proactive testing, risk assessment, and mitigation strategies are crucial for success.

Next steps: Building your AI-ready team

Organizations building a generative AI tool may find their staff need new or refined skills. In the next chapter, you'll learn how to nurture an AI-ready workforce, including the technical skills required to create an effective and iterative AI tool. We'll also cover how you can consider effectively outsourcing your needs to the right vendor so you can stay focused on your priority mission areas.

CHAPTER 6

Building an AI-ready public sector workforce

A successful strategy for generative AI has to include careful consideration of the people throughout an organization. In earlier chapters we discussed important considerations for building your AI-ready public sector workforce, including how human work and decision-making can be enhanced through generative AI tools, and what technical resources are necessary to support those tools. We also emphasized how local information regulations affect data processes, and how processes and people may need to change as regulations evolve.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Other key considerations for generative AI in public sector workforces include what roles each department and personnel have in delivering quality, privacy, security, and compliance of generative AI tools and data. And how clearly your organization communicates to staff the role that generative AI will play in delivering your mission and supporting their work.

Effectively and responsibly supporting public sector staff with generative AI tools

One consideration for integrating any AI tool that is especially important for public sector use cases is evaluating the impact from issues like bias and inaccuracy so that you can determine the types of strategies and resources needed for prevention.

Given the impacts from issues like bias and toxicity, which we examined in Chapter Five, public sector organizations are likely to face increasing pressure to demonstrate how AI tools came to support their decisions—a concept called “explainability.” This can be a complex requirement depending on outputs, but it sometimes can be as basic as making sure a chatbot not only provides the response but also attaches a link to the source document where the response was found.

Generative AI tools have incredible powers to elevate human decision-making by better presenting information, enhancing access to data, highlighting deeper context and connections, and providing faster processing of inputs. In situations where mistakes have human impacts, like misdiagnosis of patients in a medical setting or approval for assistance programs, it is particularly important to make sure that human decision-makers are being supported, not replaced.

Given the importance of being both accurate and free of toxicity in public sector work, it is necessary to invest in ongoing improvement activities like continuous training or management of the data that the application interacts with, in order to appropriately fine-tune outputs.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors**Glossary**

Architecting ways to collect new ground truth and feedback from your end-users to fuel your fine-tuning efforts can improve results and help win staff buy-in. It is important to clearly communicate the intensions of AI tools and value for staff when building such efforts.

Scoping generative AI development and technical resource needs

When delving into the world of AI, the level of customization you require of the AI tool in terms of specific skill sets and output formats will determine the extent of your investment in technical resources. To embark on this journey, you'll need a dedicated team of experts. Machine learning specialists, developers, data scientists, and DevOps teams will become invaluable collaborators.

Ensuring the quality of AI outputs and mitigating issues like toxicity and bias is an ongoing process. It necessitates data science work and investment in continuous improvement throughout the product's lifecycle. The investment in machine learning operations and data engineering staff who will maintain and improve your AI systems is as crucial as the initial development investment, especially when aiming for specific compliance goals.

Carefully understanding the goals of your AI project and the scope of your data engineering and model customization needs will help you effectively evaluate the level of depth and engagement needed in your staffing. If your needs can be met with foundation models requiring no model customization, then you may have simpler data engineering requirements.

One decision that can affect necessary skills is how much of your AI capabilities you want to build in-house versus relying on vendors. While subscriptions to pre-trained models with vendor APIs can reduce your technical workload, you still have to select the best model and integrate it with your application. It's essential to evaluate your specific needs and resources before deciding on the best approach.

Building an effective generative AI development team at every level

Once you have a good sense of the scope of your generative AI project and customization needs, you can estimate the size and depth of the technical teams you will need in place to accomplish your goals. In Chapter 3, we discussed the increasing levels of complexity in generative AI techniques: RAG, prompt engineering, foundation model fine-tuning, and training of new models for the most custom use cases. The types of skillsets your organization needs access to for execution increases as additional techniques are applied to produce the desired outputs.

Generative AI applications can be achieved using prompts and browser interfaces with foundation models, such as those found on [Amazon Bedrock Playground](#). Less technical staff, familiar with data querying, can handle this work after the initial setup, user access, and data management have been properly established by someone with the necessary technical expertise.

For use cases that require higher degrees of domain specificity or customization, fine-tuning foundation models with further training data may be necessary. At this level, your team might include ML specialists to support model evaluation, customization, and continuous improvement activities. You will need both mature development capabilities and strong data science staff, supported by a degree of DevOps and Machine Learning Operations (MLOps) resources for deploying and training models. DevOps teams can often expand scope to take on MLOps tasks.

Adding in RAG techniques to increase the currency or specificity of a model's context will increase your needs on both the data management and DevOps fronts. The data needs to be properly classified, and the access controls carefully considered. Meanwhile, that data needs to be indexed and available according to application access patterns.

If your organization is pursuing trail blazing new uses cases, you may need to train new models from scratch. Doing this successfully will require substantial investment in ML experts who are capable of training new models, as well as all of the domain-specific training data. Classifying, managing, and consuming the training data will also require substantial investments in data science teams, and MLOps staff will be necessary to help deploy the models and monitor them through training cycles.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Generative AI customer story: Technology Innovation Institute

The Technology Innovation Institution, a leading research center in the United Arab Emirates, launched two versions of the Falcon large language model (LLM) one with 40 billion and one with 180 billion parameters, both built on Amazon SageMaker. Falcon 40B is trained on 11 tokens (sequence of characters in a document this are grouped together as a useful semantic unit for processing) and was ranked the top open-source model for several weeks in the public Hugging Face Open LLM leaderboard.

Shared responsibility: Ensuring privacy, security, and compliance when working with public sector data

When working with generative AI in the public sector, particular attention should be paid to privacy, security, and compliance. Delivering those requirements is a responsibility that should be shared by staff throughout the organization. Building strong cross-functional connections within your organization, especially between teams like data science, security, and legal, is paramount

When building generative AI, underlying data may have personally identifiable information or protected health information which will need to be anonymized or safe-guarded. To navigate this complex landscape successfully, it's essential to establish strong access controls that let teams work with the data they need, while protecting all other access.

Effective collaboration between data science and legal teams is key to managing the legal aspects of handling sensitive data. Establishing security measures to protect data privacy is equally vital. It's a joint effort that demands constant vigilance and adaptation to evolving regulations and threats, and it's an ongoing commitment to safeguarding sensitive information and maintaining public trust.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

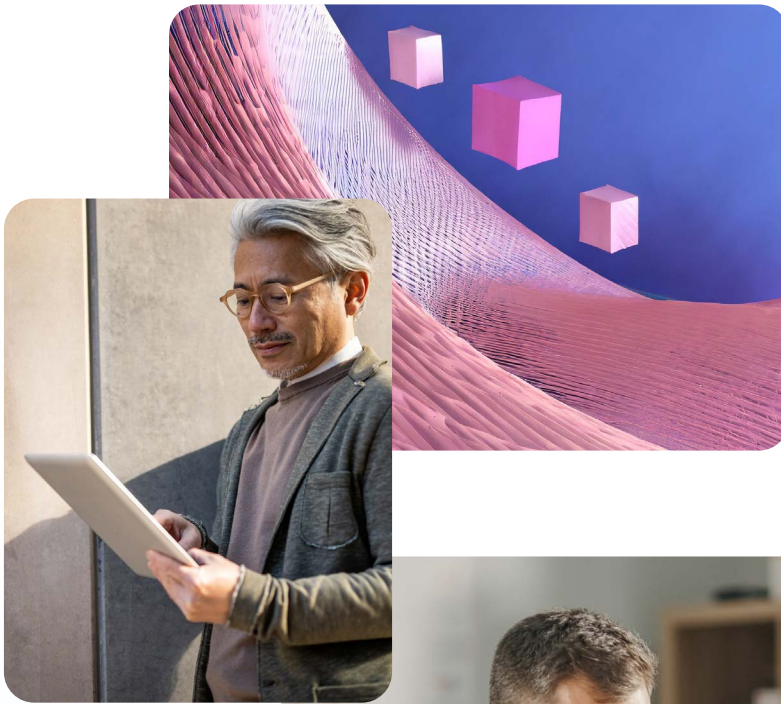
A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Communicating the role of AI in delivering your mission

Effective communication of your AI strategy and how it relates to mission function is essential for creating alignment and understanding within your organization. Leading organizations work to create a shared vision and then empower their teams to ideate and explore with experiments and proofs of concept around solutions that drive end-user value. Actively solicit ideas and concerns from your team, which will be invaluable as you work to identify and mitigate risks in your project.

It is also important to provide clearly defined positions on critical AI issues, such as privacy, toxicity, and bias. Your staff should know where your organization stands on these ethical and operational matters. This clarity helps guide their actions and decisions when working with AI systems. As part of AWS's commitment to developing fair and accurate AI and ML services and tools, [this toolkit has been produced as a starting point for these conversations.](#)

Building with generative AI requires an ongoing process of continual improvement. You will learn in the next chapter how participation and engagement from your staff are key to grow your organization's capacity to deliver its mission with generative AI.



CHAPTER 7

A new era of building in the cloud with generative AI on AWS

Generative AI has the potential over time to reinvent customer experiences across industries.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

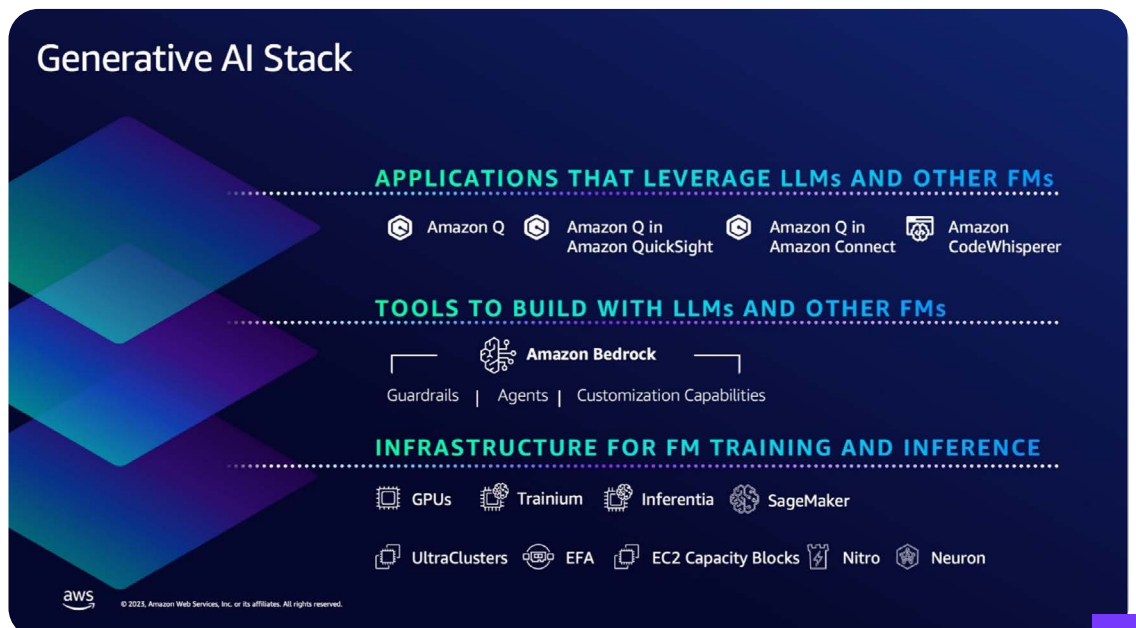
Glossary

Many organizations are already launching generative AI applications on Amazon Web Services (AWS), like LexisNexis Legal & Professional. Leading Foundation model providers like Anthropic have selected AWS as their primary cloud provider for mission-critical workloads and to train future models. Global services providers like Accenture are benefiting from customized generative AI applications as they empower developers with [Amazon CodeWhisperer](#).

These organizations choose AWS because of the focus on democratizing complex technology by reducing the up-front investment needed to get started and transform experiences and business for all. Amazon innovates across the generative AI stack to provide comprehensive capabilities.

At the bottom layer is the infrastructure to train their own large language models (LLMs) and foundation models (FMs) and run inferences. The middle layer provides easy access to pre-trained models and tools to build and scale apps with expected security and access controls. Amazon has also invested in top-layer applications like generative coding companion and enterprise ready intelligent chatbots. Organizations value the breadth, depth, data-first approach, and enterprise-grade security and privacy across all layers.

At the 2023 re:Invent AWS user conference, Amazon announced notable new capabilities across the stack to make generative AI more practical and accessible.





Bottom of the stack: the most advanced cloud infrastructure for generative AI

The bottom layer of the stack is the infrastructure—compute, networking, frameworks, services—that are required to train and run LLMs and other FMs. We continually invest in unique innovations that make AWS the best cloud to run GPUs, including the price-performance benefits of the advanced virtualization system [AWS Nitro System](#), powerful petabit-scale networking with [Elastic Fabric Adapter](#) (EFA), and hyperscale clustering with Amazon Elastic Compute Cloud (EC2) [UltraClusters](#). We also expand access to GPUs for generative AI through offerings like EC2 Capacity Reservations, a unique service for reserving GPU capacity for future short machine learning (ML) workloads.

To keep pushing the envelope on price performance, AWS committed several years ago to innovate all the way down to the hardware level. To accomplish that, AWS began investing in its own chips. AWS first developed Inferentia, a purpose-built inference chip. Now, second-generation Inferentia2 powers Inf2 instances with up to 4x higher throughput and 10x lower latency.

The AWS Trainium chip powers Trn1 training instances, distributing workloads connected by EFA networking for better price performance. As models grow exponentially, AWS continues to push the boundaries of performance and scale. The [recently announced AWS Trainium2](#) chip delivers even better price performance with up to 4x faster training than Trainium and up to 65 exaflops of aggregate compute in EC2 UltraClusters. This enables training a 300 billion parameter LLM in weeks instead of months. Anthropic is using Trainium2 to train future models, collaborating with AWS on Trainium and Inferentia innovation. Look for the first Trainium2 instances to be available to customers in 2024.

Major advancements have also been made with [AWS Neuron](#), an ML silicon software development kit (SDK), that maximizes performance from Trainium and Inferentia. Since 2019, AWS has invested in compiler and framework technologies. Now Neuron supports many of the most popular public models and integrates with frameworks like PyTorch and TensorFlow. Customers easily transition existing pipelines to our chips with Neuron.

With a combination of choice of the best ML chips, superfast networking, virtualization, and hyperscale clusters, it's not surprising that some of the most well-known generative AI startups like AI21 Labs, Anthropic, Hugging Face, Perplexity AI, Runway, and Stability AI run on AWS. But you need the right tools to effectively use this infrastructure to build, train, and run models efficiently.

Amazon SageMaker provides the right tools to effectively use this infrastructure to build, train and run models efficiently. Amazon SageMaker removes barriers in training and deploying large generative models for public sector organizations. Whether building proprietary models or using popular public models, training complex models at scale is difficult and expensive. Running them cost-effectively also poses challenges in data preparation, maintaining accelerators, distributing training, monitoring models, and resolving hardware issues.

Amazon SageMaker simplifies training and deployment by automating these tedious processes. Over the 6 years since its launch, SageMaker has grown in scope to support the rapid growth of models' size and complexity. As models rapidly grow in size and complexity, so does SageMaker's scope. Over 6 years, SageMaker has gained over 380 features such as automatic tuning, distributed training, flexible deployment, ML tools, data preparation, notebooks, and responsible AI.

Amazon SageMaker HyperPod simplifies high-scale distributed training by automating tedious processes, speeding up training by up to 40 percent. Customers such as Hugging Face and Hippocratic already use HyperPod to build, train, or evolve models. SageMaker also now helps customers deploy multiple models to the same instance so that they can share compute resources—reducing inference cost by 50 percent on average. By actively monitoring instances that are processing inference requests and intelligently routing requests based on which instances are available, SageMaker is able to achieve 20 percent lower inference latency on average.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary





Middle of the stack: easy access to generative AI models and tools

For public sector organizations to succeed at their mission, they need easy access to the middle layer of the generative AI stack: the models and tools needed to build and scale apps. To enable easy access to generative AI, AWS introduced Amazon Bedrock, a managed service that allows customers to easily build and deploy generative AI applications using leading models like Anthropic's Claude, Stability AI's, Stable Diffusion, Meta's LLAMA, and AWS's own Titan models. Customers can choose from different models optimized for specific use cases, customize models with their own data to improve accuracy, and take advantage of AWS's security, access controls, and infrastructure.

Many customers are excited to use Amazon Bedrock across diverse industries and use cases like summarization, question answering, and image generation.

AWS continues to rapidly innovate and iterate on Amazon Bedrock capabilities:

- Expanding model choice. Customers want the flexibility to experiment with different models for different use cases and features. [Amazon Titan](#) FMs are created and pre-trained by AWS to offer powerful capabilities. AWS has also added Anthropic's Claude 2.1, Meta's LLAMA 2 70B LLMs, AI21 and others.
- Increasing accuracy and relevance in search. A new embeddings model, Titan Multimodal Embeddings, uses images and text (or a combination of both) as inputs to power multimodal search, personalization, and recommendation experiences for users using images and text (or a combination of both) as inputs.
- New model choices from Amazon Titan for high-performing image, multi-modal, and text model, via a fully managed API. For example, Titan Text Lite is more expansive and can be used for a wider range of text-based tasks, such as open-ended text generation and conversational chat.
- Powering image generation. Amazon Titan Image Generator gives organizations across the public sector the ability to use a text input to generate realistic, studio-quality images in large volumes and at low cost.
- Adding new customization capabilities. Knowledge Bases for Amazon Bedrock lets organizations prompt FMs with contextual information from their own data sources for Retrieval Augmented Generation (RAG). Continued pre-training allows organizations to tailor models with proprietary data while maintaining security and privacy.
- Improving AI safety. [Guardrails for Amazon Bedrock](#) applies safeguards based on responsible AI policies to avoid toxic language, stay on topic, and redact PII.
- Enabling multistep tasks. [Agents for Amazon Bedrock](#) helps AI applications execute multistep tasks like answering customer questions or taking orders by integrating systems, data sources, and company knowledge.

The future of generative AI is choice. There won't be one dominant model, and organizations need to be able to select the generative AI tools that let them fulfill their mission. AWS offers unique customization capabilities and continues to innovate quickly to make building with different models easy.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Top of the stack: commitment to accessible innovation

At the top layer of the stack are applications that leverage LLMs and other FMs so that you can take advantage of generative AI at work. One area where generative AI is already making an impact is in coding. Amazon CodeWhisperer helps you build applications faster and more securely by generating code suggestions and recommendations in near real-time.

Amazon recently previewed a [new customization capability](#) in CodeWhisperer that securely leverages an organization's internal code base to provide more relevant and useful code recommendations. With this capability, CodeWhisperer is an expert on your code and provides recommendations that are more relevant to save even more time.

Persistent Systems, a global digital engineering and enterprise modernization company, has run several pilots and formal studies with CodeWhisperer and found that customizations help developers complete tasks up to 28 percent faster than with CodeWhisperer's general capabilities. Now a developer at a healthcare technology company can ask CodeWhisperer to "import MRI images associated with the customer ID and run them through the image classifier" to detect anomalies. Because CodeWhisperer has access to the code base, it can provide much more relevant suggestions that include the import locations of the MRI images and customer IDs. CodeWhisperer keeps customizations completely private, and the underlying FM does not use them for training, which protects organizations' valuable intellectual property.

Introducing Amazon Q, the generative AI-powered assistant tailored for work

Developers certainly aren't the only ones who are getting hands-on with generative AI—millions of people are using generative AI chat applications. What early providers have done in this space is exciting and useful for consumers, but applications need customization to be successful in a work environment.

Generative AI excels in its general knowledge and capabilities, but without knowledge of your organization and its operations, data, and users, its organizational use is limited. Generative AI tools are also limited by their lack of knowledge about individuals' roles within your organization—what work they do, who they work with, what information they use, and what they have access to.

These limitations are understandable because these assistants don't have access to your company's private information, and they weren't designed to meet the data privacy and security requirements companies need to give them this access. Security protocols work best when they are included as part of the original design rather than added as an afterthought. AWS tools give organizations a way to use generative AI safely in their day-to-day work.

[Amazon Q](#) is a new generative AI assistant designed specifically for the workplace. Amazon Q leverages a company's unique data, systems, and expertise to provide employees with fast, tailored answers and insights to help them work more efficiently.



Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

When you chat with Amazon Q, it provides immediate, relevant information and advice to help streamline tasks, speed decision-making, and help spark creativity and innovation at work. Amazon Q is secure and private, and it can understand and respect your existing identities, roles, and permissions and use this information to personalize its interactions. If a user doesn't have permission to access certain data, Amazon Q won't permit access. Amazon Q meets stringent requirements from day one—none of your organizational content is used to improve the underlying models.

Amazon Q is trained on 17 years of AWS knowledge and experience so it can help developers build solutions on AWS faster. Amazon Q can provide guidance on optimal AWS services for a use case, explain code, suggest optimizations, diagnose issues, and accelerate upgrades. For example, Amazon Q transformed 1,000 Java applications from Java 8 to Java 17 in just 2 days.

Amazon Q connects to a company's repositories and systems to understand its business. It answers questions using authorized data and generates content grounded to provided sources. Amazon Q helps employees by summarizing information, generating content, structuring meetings, and completing tasks. It streamlines communications and eliminates repetitive work.

Amazon Q is available in the business intelligence service Amazon QuickSight, the contact center service Amazon Connect, and will soon be available in AWS Supply Chain. [Amazon Q in QuickSight](#) helps create visualizations and explain changes in dashboards. [Amazon Q in Connect](#) assists customer service agents by suggesting responses and next steps. In Supply Chain, Amazon Q will help optimize inventory management by summarizing and highlighting potential stockout or overstock risks and can visualize scenarios to solve the problem.

Overall, generative AI promises to transform how we build applications and get work done. AWS is committed to giving customers choice in leading models while making adoption accessible and responsible. There is no one-size-fits-all approach, so Amazon Bedrock makes it easy to experiment, customize, and combine the best innovations as they emerge across models, techniques, and providers. With services like Amazon Q, Amazon Bedrock, CodeWhisperer, and more, AWS is rapidly innovating so all customers can benefit from and responsibly apply generative AI. Anyone can now reinvent what's possible with these groundbreaking new capabilities.

NEXT STEPS

Beginning your generative AI journey with confidence

With initial use-cases identified and a governance strategy in place, your team is ready to start innovating with generative AI. You can leverage the power of generative AI to improve document processing workflows, assist in constituent communications, assist in data query and analysis, and a number of emerging use-cases.

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

As your team begins to experiment with prompting, they will learn which models suit their target use, how to best deliver results, and what data to use in their prompts. Experimenting is critical in developing your understanding of the rough scope and scale for a project and its team. For example: including retrieval augmented generation (RAG) in your design might require additional data engineers on your team. Similarly, fine-tuning could motivate the hiring of additional data engineering resources as well as ML or MLOps resources.

To quickly experiment with scope for your AI project, you can use [PartyRock](#), a hands-on generative AI app-building Amazon Bedrock Playground, to quickly prototype features without any coding.

Architect your application for continuous improvement.

As advancements unfold in model releases and emerging techniques, it is essential to continually enhance your application for optimal performance and anticipate the need to replace models over time. Since each model adheres to unique prompting best practices, be prepared for adjustments in your prompts. Additionally, be mindful that the output format may vary slightly with the introduction of a new model. To navigate these nuances seamlessly, ensure your application is designed to adapt effortlessly to changes. Developing flexibility into your application architecture will enable you to manage evolving models and maintain peak efficiency.

You can host your models using fully managed hosting where you consume an API, or you can use managed tools to launch your own managed model endpoints. Your choice in hosting approach and choice in models may have some overlap because not all models are available on either hosting strategy. With prototype prompts and forecasted application usage, you can better understand hosting costs. If your requirements include a customized model, be sure to select one that includes the tools you need to perform the work.

All the usual data governance rules apply and you should be ready for change as new regulations are passed. You also need to manage the outputs of your application for new risks like inaccuracy, bias, monitoring, and a process for continuous improvement.

Your data and data governance requirements will be a driver in the size and complexity of your project. You'll need to manage the data your application uses, the data it collects, and any other data you use to improve outputs.

Now that you have a better understanding of workforce and project size considerations, you're well equipped to start exploring options. [Contact us to discuss your unique use case and determine how AI can work for your staff and your mission.](#)



Contributors

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

Abhilash Thallapally
Solutions Architect

Adam Tashman
Technology Advisor, Data Science

Americo Carvalho
Sr. Manager, AI/ML and Edge Services

Ashraf Osman
Sr. Solutions Architect, State and Local Government

Parnab Basak
Sr. Product Acceleration Lead Solutions Architect, Gov/Tech

Fred Azar
Principal AI/ML Business Development Lead, Healthcare

Henry Jia
Sr. AI/ML Specialist Solutions Architect

John Kitaoka
Solutions Architect, Gov Tech

Josh Famestad
Solutions Architect, GovTech

Mary Strain
AI/ML, Business Development Lead, Education

Prajwal Shetty
Solutions Architect

Pranusha Manchala
Sr. Solutions Architect, EdTech

Rahul Kulkarni
Solutions Architect

Ralph Gimash
Solutions Architect, Gov Tech

Sam Palani
Sr. Leader, AI/ML Specialist Solutions Architect

Sergio Ortega
AI/ML Business Development Lead, State and Local Government

Tom Romano
Sr. Product Acceleration Solutions Architect

Glossary

Accuracy: The degree to which AI models provide correct and reliable results according to expectations. (Chapters 2, 3, 4, 5)

AI Models': Foundation, Large Language, etc

Amazon Augmented AI: A tool for improving AI model accuracy through human review. (Chapter 5)

Amazon Bedrock: A fully managed generative AI cloud service that offers a choice of high-performing foundation models (FMs) from leading AI companies like AI21 Labs, Anthropic, Cohere, Meta, Stability AI, and Amazon from a single API, along with a broad set of capabilities to build generative AI applications, simplifying development while maintaining privacy and security. (Chapters 2, 3, 5)

Amazon CodeWhisperer: An AI coding companion that generates real-time, single-line or full-function code suggestions in your integrated development environment (IDE) to help you quickly build software. With CodeWhisperer, you can write a comment in natural language that outlines a specific task, such as, "Upload a file with server-side encryption." Based on this information, CodeWhisperer recommends one or more code snippets directly in the IDE that can accomplish the task. (Chapters 1, 2, 4, 6)

Amazon SageMaker: A cloud service for building and deploying machine learning models. With SageMaker, data scientists and developers can quickly build and train machine learning models, and then directly deploy them into a production-ready hosted environment. (Chapters 2, 3, 4, 5)

Amazon SageMaker Clarify: A set of tools for enhancing the explainability of AI outputs. (Chapters 4, 5)

Amazon SageMaker Model Monitor: A tool for monitoring AI models to detect issues like bias and inaccuracies. (Chapters 4, 5)

AWS: Amazon Web Services, a cloud computing platform.

Backtracking: The ability to explore and understand the rationale behind AI model decisions, which can provide better validation and accountability. (Chapter 3)

Bias: Unfair or discriminatory behavior or decisions made by AI models, often reflecting societal biases present in the training data. (Chapters 2, 3, 4, 5)

Chatbot: A program that uses artificial intelligence to engage in text or voice conversations with users, often assisting with tasks or providing information in a human-like manner. (Chapters 1, 2)

Compliance Certification: Official recognition that an AI model meets specific regulatory requirements. (Chapter 2)

Compliance Requirements: Regulations and rules that must be followed, especially in contexts involving sensitive data, such as personal health information (PHI) or government compliance programs like FedRAMP. (Chapters 2, 3, 4, 5)

Continuous Improvement: The ongoing process of enhancing systems and practices to achieve better results and outcomes. (Chapters 2, 3, 4, 5)

Context Window: The extent of contextual information an AI model can consider when generating responses. (Chapters 3, 4)

Data Governance: The framework and practices that ensure data quality, security, and compliance within an organization. (Chapters 4, 5)

Data Hygiene Practices: Procedures and measures for maintaining the quality and privacy of data. (Chapters 5, 6)

Data Lineage: A record of data's origins, changes, and transformations throughout its lifecycle. (Chapters 2, 5)

Development Costs: Expenses associated with customizing or modifying AI models. (Chapter 5)

DevOps Teams: Development and operations teams responsible for deploying and maintaining systems. (Chapters 4, 6)

Explainability: The ability of an AI model to provide clear explanations of its decision-making processes. (Chapters 2, 4, 6)

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary



Fine-Tuning: The process of adjusting pre-trained AI models using specific data to improve their performance for a particular task. (Chapters 2, 3, 4, 5)

Foundational AI Models (FMs): Large-scale neural network models that have been pre-trained on specific domains of knowledge and desired outputs, providing a basis for customizing and building more specific AI solutions. (Chapters 1, 2, 3, 4, 5)

Generative AI: Artificial Intelligence technologies that create and reshape content. Generative AI applications use foundation models to generate new text, images, sounds, or other mediums of output.

Human-in-the-Loop: A collaborative approach where humans are integrated into AI-driven processes to verify model outputs, make adjustments, provide feedback, and monitor for accuracy, bias, and other considerations. (Chapters 1, 2, 4, 5, 6)

Intelligent Document Processing: The application of generative AI to extract information from documents, understand content in context, and perform tasks such as natural language understanding, summarization, translation, and text generation. (Chapter 1)

Large Language Models (LLM): AI models that have been trained on large datasets of human-readable content, making them proficient in processing and generating natural language data.

Machine Learning (ML) Model: A neural network that has been pre-trained through machine learning techniques to acquire specific knowledge—the core component driving generative AI solutions.

Machine Learning Operations (MLOps): A set of software management practices for deploying and maintaining machine learning models. (Chapters 2, 4, 6)

Model Customization: The process of adapting AI models to suit specific needs or requirements. (Chapters 2, 3, 5)

Model Selection: The process of choosing the appropriate core AI model for a specific task or application.

Open-source Foundation Models: Pre-trained FM models available for use by the public or organizations. (Chapters 3, 4)

Output Validation / Output Grading: The assessment and verification of the results produced by AI models to ensure accuracy and compliance with requirements. (Chapters 3, 4, 5)

Personally Identifiable Information (PII): Information that can be used to identify individuals, such as names, addresses, and identification numbers. (Chapters 3, 4, 5)

Pre-Built Models: AI models that are readily available for use without extensive training or customization.

Prompt Engineering: The process of crafting suitable inputs or queries to elicit desired responses from AI models, based on desired quality metrics. (Chapters 3, 4, 5)

Prompt Templates: Predefined inputs or queries with specific structures that have been tested and validated, used as framing to interact with AI models. (Chapters 1, 3, 4, 5)

Protected Health Information (PHI): Sensitive medical information that is subject to strict privacy regulations. (Chapters 2, 4, 5, 6)

Reinforcement Learning from Human Feedback (RLHF): A technique that uses human feedback on outputs to refine AI models. (Chapter 4)

Regulatory Shifts: Changes in laws, rules, or regulations that impact how AI is used and governed in the public sector. (Chapter 6)

Retrieval Augmented Generation (RAG): An output customization approach that allows AI models to retrieve and incorporate information from external data sources and run-time to generate responses. (Chapters 3, 4, 5, 6)

Self-Supervision: A training mechanism for AI models where learning is done from large datasets without explicit human labeling. (Chapter 3)

Toxicity: Issues related to negative or harmful outputs in AI, such as discriminatory and offensive content. (Chapters 1, 2, 4, 5, 6)

Training Costs: Expenses associated with training AI models, influenced by factors like data size and model choice. (Chapter 3)

Vendor APIs: Application Programming Interfaces provided by AI vendors to access and use their pre-trained models. (Chapter 2, 6)

Foreword

How the public sector can benefit from generative AI

Chapter 1

Generative AI's transformative potential

Chapter 2

Building an effective AI strategy in the public sector

Chapter 3

Understanding key AI model selection considerations

Chapter 4

Effective generative AI data management model inputs, prompt engineering, and output validation

Chapter 5

Managing AI model security, privacy, compliance, and bias

Chapter 6

Building an AI-ready public sector workforce

Chapter 7

A new era of building in the cloud with Generative AI on AWS

Next steps

Beginning your generative AI journey with confidence

Contributors

Glossary

