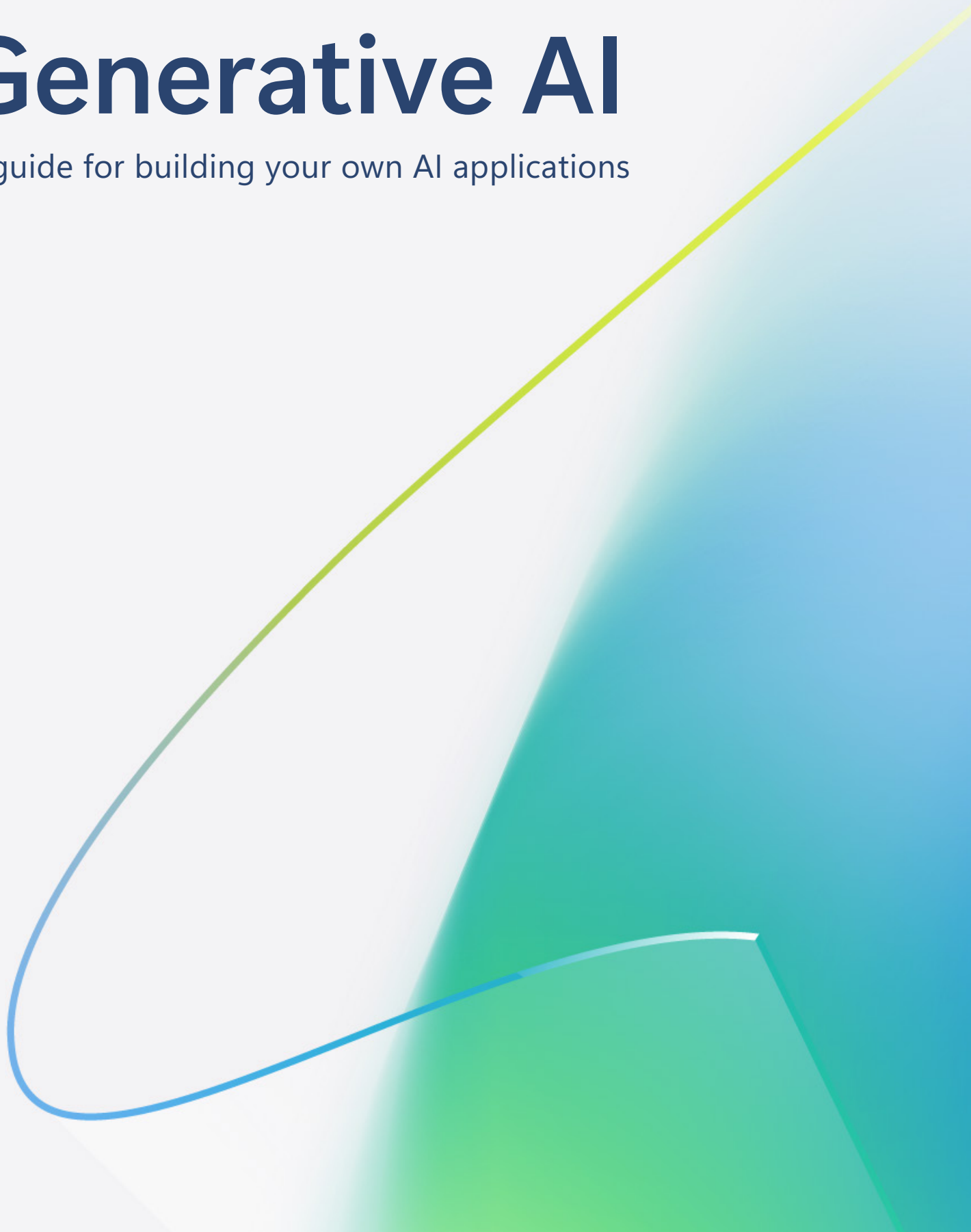
Microsoft Azure

Generative AI

A guide for building your own AI applications



Contents

Introduction

Executive Summary

Generative AI is transforming organizations and the people who drive them. It allows software to be more intuitive and helpful, raising employee productivity and user satisfaction. At the same time, generative AI poses a new set of challenges and risks.

This e-book delves into the strategic considerations for those who want to build their own AI applications using foundation and generative AI models, or those that want to add generative AI to their existing applications. It covers details on generative AI, offering guidance tailored for ITDMs on harnessing this technology to fulfill specific business needs and achieve competitive advantages.

Intended audience

This e-book is specifically geared towards IT decision makers (ITDMs), such as Chief Technology Officers (CTOs) or Chief Information Officers (CIOs) at companies of all sizes, including Independent Software Vendors (ISVs), who are interested in building an AI application using a foundation model.

The approximate reading time for this e-book is **30-45 minutes**.

Key topics covered in this e-book include:

1. Understanding generative AI

Differentiating generative AI from traditional software approaches, highlighting the ability to create dynamic and contextually relevant responses beyond predetermined outputs.

2. Building a generative AI team

Detailing the roles within a generative AI development team to optimize the deployment processes.

3. Developing with responsible AI practices

Learn about the tools and best practices that mitigate risk and support AI safety, quality, and compliance.

4. Utilizing a comprehensive AI development platform

Guidelines for selecting generative AI platforms and tools that align with specific business objectives. This includes learning about **Microsoft Foundry**, an essential tool for developing, deploying, and managing generative AI applications. The platform supports
the entire AI lifecycle with advanced tools for model selection, data integration, and enterprise-grade production at scale.

This e-book serves as an introductory guide to navigating generative AI and facilitating informed decisions that drive efficiency and innovation.

Learn the basics of generative AI with Microsoft Azure AI Fundamentals: Generative AI.

What's the difference between custom machine learning (ML) models and generative AI models?

Custom ML models

→ Built for a specific purpose

→ You're probably the model builder

→ Useful to make predictions on future outcomes based on pre-existing data (past outcomes)

Generative AI models

→ Built as "general-purpose"

→ You're probably not the model builder

→ Useful when you want to create new content or data that resembles/mimics patterns from pre-existing data

Microsoft provides a variety of courses designed to help data and development professionals enhance their AI skills.

These can be found in our AI learning hub.

Comparing approaches: Generative AI-enabled software development vs. standard software development

Integrating generative AI into applications, when compared to standard software development, is different in three key ways.

1. Ownership and data control

Standard software development:

In standard software development, data control and ownership are clearly defined. Developers maintain complete control over their data, which remains within their own data estate. This ownership provides straightforward management and compliance with data regulations, ensuring security and privacy without the need for third-party involvement.

Generative AI development:

Alternatively, generative AI development involves less control over data, as it often requires sending data outside one's own data estate to interact with AI models. This could potentially raise security and regulatory concerns. Although the data still legally belongs to the developer, the dependency on external AI models and services managed by third parties introduces complexity in ensuring data privacy and compliance.

Microsoft Azure Security

Regulated industries still require risk managers and regulators to demonstrate that data remains private throughout the development process. This is why security and compliance are so important when choosing a platform for the responsible development and deployment of generative AI applications at scale.

With Microsoft Azure Security, gain multi-layered security across physical datacenters, infrastructure, and operations. Azure cloud is built with customized hardware with integrated security controls and firmware components, as well as added protections against threats such as Distributed Denial-of-Service (DDoS) attacks.

Learn more about Azure Security.

2. Development process and problem solving

Standard software development:

Developers write explicit instructions to solve explicit problems, leading to predictable and consistent outputs.

Generative AI development:

In contrast, generative AI operates on probabilistic principles, using patterns and contextual data to generate outputs that may vary with the same inputs. This uncertainty makes the development process more complex and less predictable. It requires extensive testing and iteration to achieve a production-ready state. The variability in outputs often necessitates a significant involvement from domain-specific experts who assess the relevance and accuracy of the responses from AI.

3. Evaluation

Standard software development:

The developer uses unit tests to determine if the application is returning the correct answer and working as intended.

Generative AI development:

Generative AI requires a deeper, continuous involvement of domain subject matter experts (SMEs) throughout the development process considering the multiple answer outputs that need to be evaluated if they are all correct answers.

These experts not only contribute to defining the scope and specifications of applications, but they also play a decisive role in continually assessing the AI-generated outputs for metrics such as accuracy and relevance. The iterative nature of this process, along with the need for frequent evaluations and adjustments, underscores the importance of the human-in-the-loop development process.

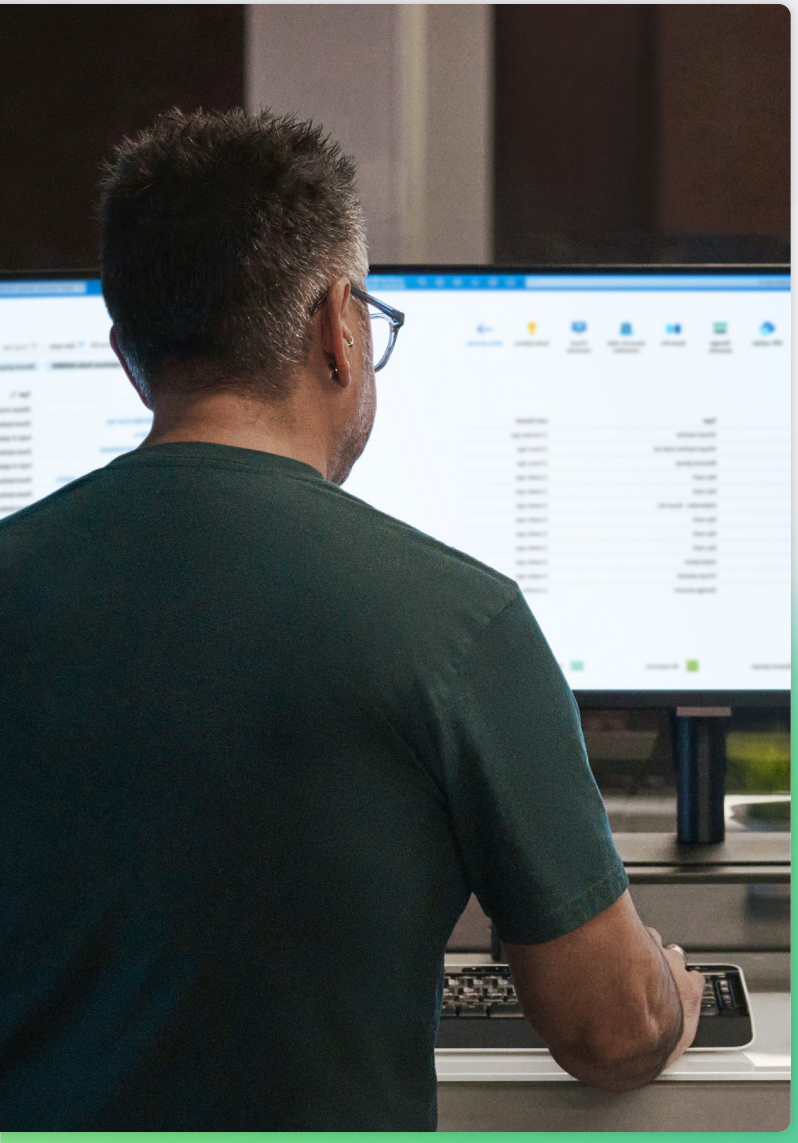
One way to evaluate applications is for human experts to rate individual answers. However, this is time-consuming, error-prone, and not feasible at scale or for ongoing monitoring of applications in production.

To help expedite and remediate this process, Microsoft researchers are developing tools for automating assessment by using models to evaluate the output of other models.

Human experts remain necessary even in this automated assessment to provide some ground-truth answers, which can be a significant investment in time.

Five common applications of building with generative AI

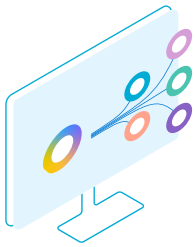
Generative AI-enabled applications can be developed from scratch, but often a faster way to gain generative AI benefits is integrating it into existing applications. This can allow users to interact with applications in new, more intuitive ways, contextualize assistance," and provide more relevant information. Generative AI is, as a result, improving applications across the value chain from customer contact to financial and operational analysis.



Below are five types of applications organizations are building with generative AI. They combine intuitive interfaces, links to proprietary knowledge bases, and enterprise systems:

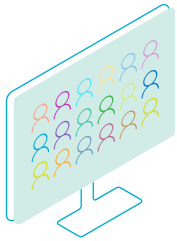
Chat with your own data

Many organizations possess a wealth of proprietary knowledge in documents and websites. To unlock this value, they are developing chat applications that allow employees and customers to query this data in plain language and receive direct answers instead of just links to source documents. Generative AI tools should be equipped with built-in responsible AI capabilities so that responses are derived solely from internal data, safeguarding against off-brand experiences and any exposure to malware through untrusted data sources. In doing so, information remains secure and compliant with company standards.



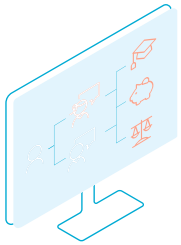
Create personalized AI agents

Software developers use generative AI to integrate advanced helpers into their systems, aimed at reducing the monotony of repetitive tasks and minimizing disruptions that hinder professional focus. These AI agents can offer enhanced in-context support, automatically draft responses tailored to the requester's identity and the context of their inquiry and compile comprehensive end-of-day or weekly summaries detailing activities and statuses. This automation aids professionals in maintaining productivity without sacrificing accuracy or attention to detail.



Add hyper-personalization

Tailor interactions to individual users at scale. Hyper-personalization can range from customizing marketing messages and adjusting website interfaces in real-time based on user actions, to personalizing product recommendations at an individual level and enhancing customer engagement and satisfaction.



Build customer service avatars

Generative AI and speech AI are allowing enterprises to create sophisticated, multilingual brand ambassador avatars where they can engage with the public in the brand's look and tone via website, kiosk, or smartphone. While early versions provided general information, avatars are now increasingly able to interact with business systems to provide specific information, make informed recommendations, check inventory, and add items to an e-commerce cart, making them an exciting new way to create high-value branded experiences.

Access insights

When important information is scattered across incompatible systems, analysts can spend hours just assembling the data for routine reports, such as end-of-quarter performance and prospect profiles. With new agent architecture, organizations are now using generative AI to pull information from multiple systems, assemble it in standard visualizations, and draft a narrative. This enables analysts to spend their time building and adding value since the computer has already done the rote work.

Select the right model for your use case

The number and diversity of generative AI models is growing rapidly, creating confusion for would-be users. Here are some considerations for selecting the right model for your application:

Model power

Evaluate the model's computational power and sophistication. This determines its capability to handle complex datasets and produce nuanced outputs. Higher power might be required for intricate tasks, impacting operational demands and associated costs.

Time efficiency/latency

The model's speed is vital, especially for tasks that need fast data processing or instant decision-making. Efficiency not only affects performance but also influences user satisfaction and operational efficiency.

Cost effectiveness

Analyze all associated costs, including initial setup, ongoing operations, and maintenance. It's essential to strike a balance between the model's capabilities and your budget.

Small Language Models (SLMs)

With SLMs, which are trained on smaller amounts of data with fewer parameters, you get the benefit of a model that can work with limited computing resources, while allowing for increased specificity. SLMs can be run locally—an advantage for regulated industries and in cases where lag time is extremely important—and can be fine-tuned to specific tasks and work with narrow context. This can be an excellent choice for specific industries looking to complete a specific task. For instance, a financial services organization can benefit from SLMs when processing claims but wants to determine the precise training to use.

Learn more about Microsoft's recent advancements with SLMs.

Fine-tuneability and extensibility

Take into consideration the ability to tweak the model to your specific data, as this can be important for the extensibility and modifications necessary for your model and the intended purposes. However, if your use case requires playing with the weights, then choosing an open model is more appropriate. Keep in mind that not all open models will have permissive license to build commercial applications.

Licensing and availability of LLMs

To use LLMs for commercial purposes, consider the license of a particular model. Remember that availability is not always straightforward, considering that some models are closed source.

Another availability consideration is that models are dependent on the data center and not all models are available in all data centers, creating challenges around restrictions with data crossing certain boundaries.

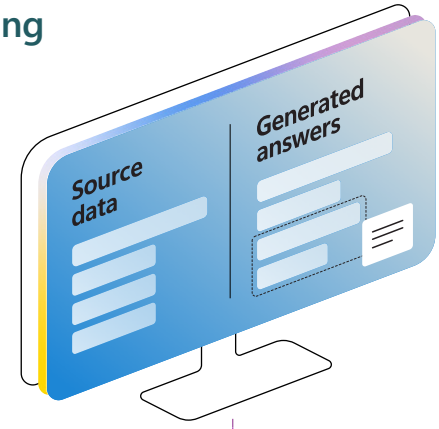
Generality vs. specificity

Choose between general-purpose and specialized models based on the breadth or specificity of your needs. General models offer flexibility, while specialized models provide high efficiency in specific contexts. For example, some models specialize in specific tasks, such as chat completion or summarization, or are purpose-built for specific data types such as code, images, video, or text, or a specific industry such as healthcare.

Microsoft evaluation and monitoring metrics for generative AI

Automated metrics for output evaluation

To enable evaluation at scale, Microsoft is developing tools to have LLMs evaluate the output of generative AI applications.



Groundedness

Azure AI Content Safety-based groundedness

The model's generated answers are evaluated based on their alignment with information from the source data, such as retrieved documents in RAG, question and answering, or documents used for summarization. The evaluation process flags output that lacks grounding.

Prompt-only-based groundedness

Measures how well the model's generated answers align with information from the source data (user-defined context).

Relevance

The extent to which the model's generated responses are pertinent and directly related to the given questions.

Coherence

How well the language model can produce output that flows smoothly, reads naturally, and resembles human-like language.

Fluency

How well written and easy to understand the answer is.

GPT-similarity

How close the answer is to a user-provided ground truth and only applies when you've supplied ground truth and are using a generative AI model to compare them.

By incorporating these considerations, organizations can more effectively align generative AI capabilities with strategic goals. The nuanced approach required for selecting and managing generative AI highlights the importance of expert involvement and iterative testing, ensuring the technology not only performs well but also integrates seamlessly into organizational processes to meet business needs.

Further details on the roles of domain SMEs and other team members will be explored later in this e-book, providing deeper insights into the collaborative and dynamic nature of generative AI development.

Building a generative AI development team

Developing with generative AI requires a hybrid approach—standard software development blended with AI expertise. Roles like the AI engineer have emerged to provide this connection between software development and AI.

At the same time, there is a partnership between technical and business teams, with business needs driving development. Non-technical business leaders and other stakeholders play focal roles in determining the utility and trustworthiness of AI applications. Their decisions are vital in deciding whether to continue using and investing in the AI system.

Preparing your data for generative AI development

Organizations can effectively prepare for generative AI projects by making sure their data is ready. Generative AI runs on data, so it's essential to ensure its quality so the output is more accurate.

Prepare your relevant data

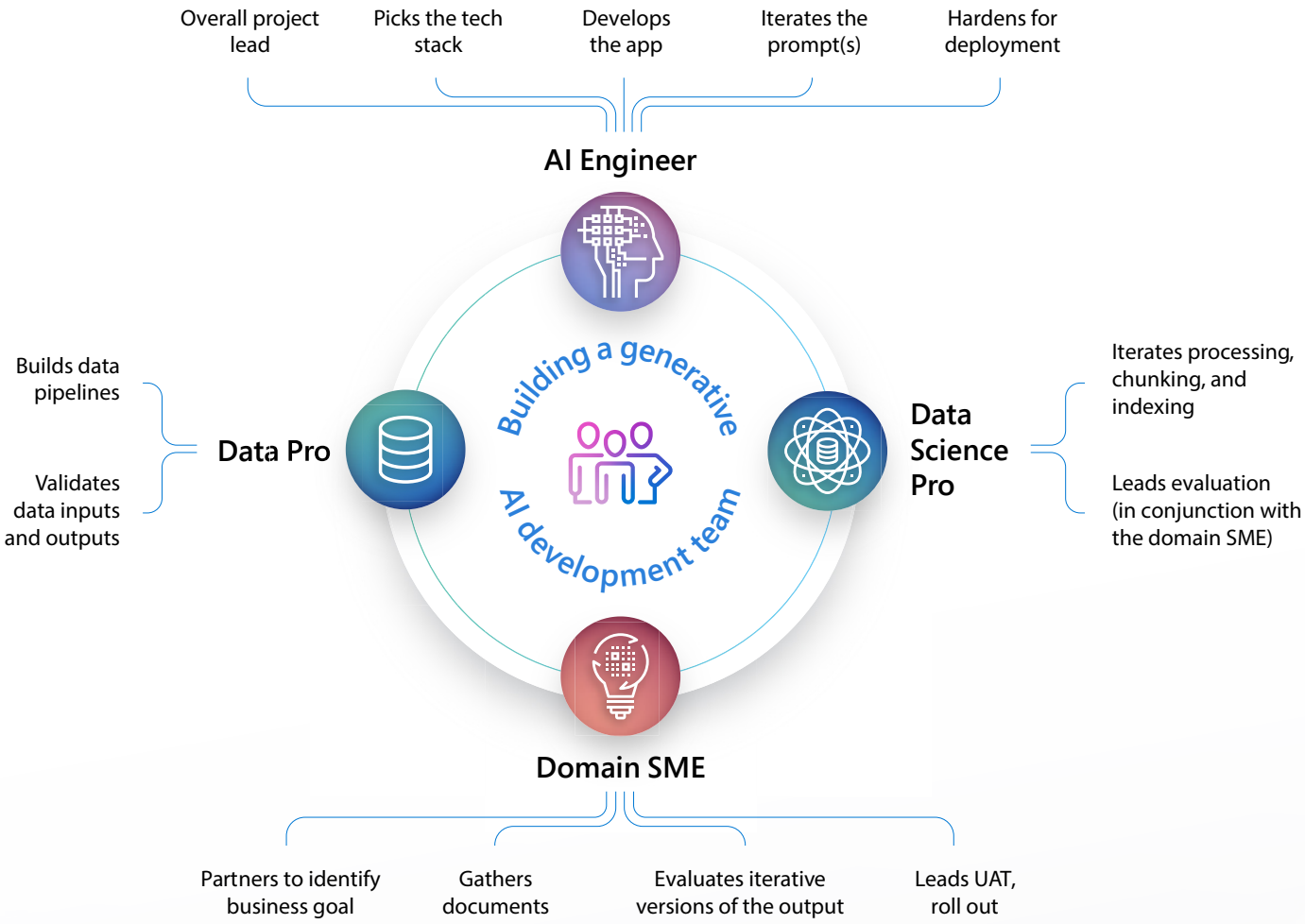
Your organization needs robust data infrastructure to handle large volumes of data and high-performance computing tasks. Data must be accessible, well organized, and secure to facilitate effective AI training and operations.

Ensure data quality

Clean and prepare your data to ensure that it is trusted and suitable for using prebuilt foundation models. In addition to data cleaning, consider if data is normalized and if biases are mitigated. Quality is essential, as it affects how precise and dependable the AI models' results are.

Building a generative AI development team

From the visual below, we can see that the AI engineer leads the development team, while time investment may be greater for the domain SMEs.



AI engineer

AI engineers are typically central, orchestrating the development initiative. Often, they are software developers who have been informally upskilled to fill this role on the team. Consequently, the AI engineer assumes the dual role of a leader and an integrator, bridging the gap between AI capabilities and business needs without the necessity to delve deeply into data science, which is typically handled by data science professionals. The AI engineer also takes information from the domain SME and adjusts the application based on their testing.

AI engineer learning resources

Looking for specific AI engineer educational resources? Get started with:

→ AI engineer certification path

→ Azure AI Fundamentals

→ Introduction to Foundry

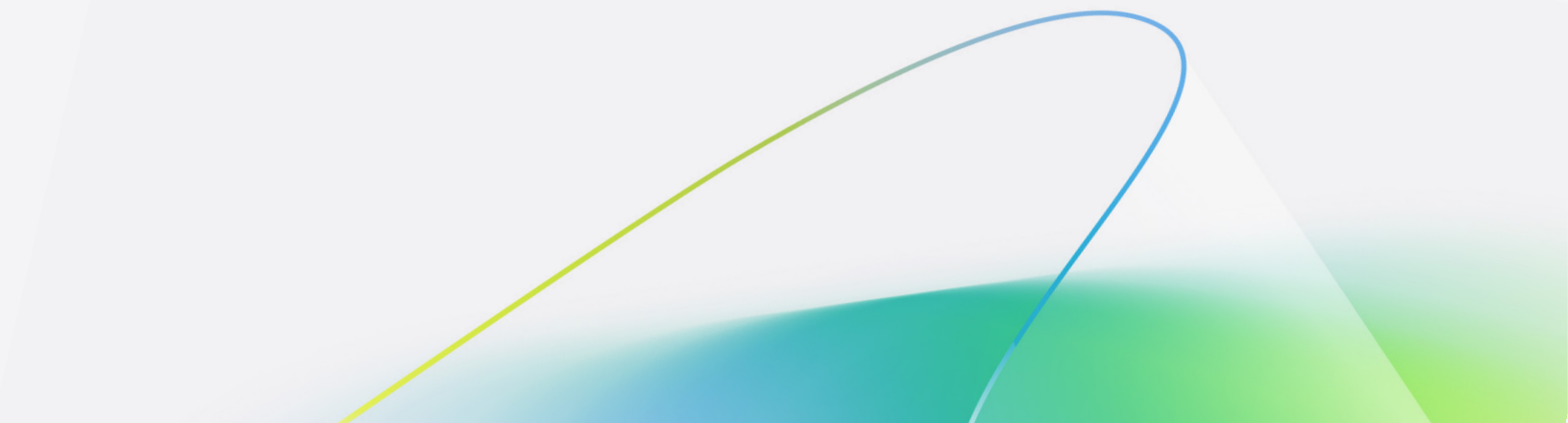
Data professional

A critical component of building with generative AI is pulling data from business systems and getting it to the model. The data professional is the one who builds the pipelines that make this happen safely and efficiently. Therefore, expertise in the nuances of data sources can be critical to create quality generative AI applications. The data professional must also navigate the complexities of data privacy regulations to ensure all data collection and usage comply with legal standards.

Domain SME

The domain subject matter expert (SME) collaborates closely with the AI engineer to ensure precise data collection. They also continually monitor key performance metrics set by the AI engineer. In generative AI development, domain SMEs tend to play a bigger role in than in standard software or machine learning because the system outputs are not easily validated.

Domain SMEs take on the role of evaluating the correctness of responses, which can be time-consuming. Their role is pivotal in bridging the technical and business aspects of AI implementations. To streamline their input, domain SMEs can work with the rest of the team to create test prompts and ground truth answers. AI engineers can then use Microsoft's automated evaluation metrics to assess the application's response quality over time and compare performance across different models.



Data scientist/machine learning professional

Not every generative AI team has a data scientist, but they are often present on complex and business-critical projects. The data scientist plays an advisory role, responsible for developing formalized test plans and setting evaluation criteria, which alleviates some of the domain SME's workload.

A common approach is Retrieval Augmented Generation (RAG), which combines the reasoning capabilities of models with custom data to generate responses. This involves testing various document chunking and annotation techniques, assessing retrieval quality, and evaluating responses across different prompts.

Looking for specific AI data science educational resources?

Get started with:

→ Microsoft Certified: Azure Data Scientist Associate

As the need for generative AI within organizations grows, so does the demand for a centralized platform to develop and deploy these applications responsibly. This underscores the increasing significance of the data science professional's role in ensuring these technologies are implemented effectively and ethically.

Foundry: A comprehensive platform

Organizations need tools that simplify AI development, allowing them more time to focus on big-picture business needs. **Foundry**, Microsoft's generative AI platform, is designed to democratize the AI development process for developers, bringing together the models, tools, services, and integrations necessary to begin quickly and efficiently developing your own AI applications.

87%

of organizations believe AI will give them a competitive edge[1]

66%

of surveyed board members report pressure to accelerate AI adoption[2]

McDonald's China transforms its operations, elevates service levels with Foundry

What was their goal?

McDonald's China needed to uplevel their customer service, quality, and operations to accommodate the growing number of locations and the rapid pace of innovation. They needed a way to adhere to their brand mission through increased digital transformation.

How did they achieve it?

With the support of Microsoft, McDonald's China established the AI Lab, allowing them to integrate AI into their existing operations, predominately by leveraging large language models (LLMs).

Through the AI Lab, AI has proliferated their entire operations, from supply chain operations to even supporting marketing campaigns.as it affects how precise and dependable the AI models' results are.

Learn more >

[1] Expanding AI's Impact With Organizational Learning
[2] CEO decision-making in the age of AI

As a comprehensive platform for developing and deploying generative AI applications, Foundry emerges as a game-changing tool for both seasoned developers and those new to AI. With drag-and-drop functionality, visual programming environments, and prebuilt templates, Foundry makes it easier for users to prototype, build, and refine AI applications without deep technical knowledge of the underlying algorithms.

This accessibility accelerates the development process and helps users quickly translate their creative and business ideas into fully operational AI solutions.

Siemens and Microsoft partner to drive cross-industry AI adoption

What was their goal?

Multinational technology conglomerate **Siemens** needed to simplify virtual collaboration of design engineers, frontline workers, and other teams across key business functions. They were looking to utilize generative AI to rapidly generate, optimize, and debug complex automation code.

How did they achieve it?

With the introduction of Siemens Industrial Copilot, a generative AI-powered assistant, Siemens enhanced human-machine collaboration and boosted productivity. They also significantly shortened simulation times, reducing task time from weeks to minutes. Maintenance staff could now use the power of natural language and gain assistance with detailed repair instructions and provide engineers with quick access to simulation tools.

Learn more >

Key features of Foundry that provide substantial value to users include:

Responsible AI tools and best practices:

Foundry empowers developers to safely, securely, and responsibly innovate and shape the future with AI. The comprehensive platform accelerates the development of production-ready agents to support enterprise chat, content generation, data analysis, and more. Developers use their protected data to build custom models and solutions with collaborative, responsible AI tools and best practices.

API and model choice:

Users can access and deploy the latest models as APIs, facilitating rapid, serverless, and fine-tuned model deployment. This reduces development time and resource costs, accelerating time to market for developers and providing end-users with state-of-the-art AI capabilities quickly. The Assistants API makes it easier for developers to create automated applications with sophisticated, agent-like experiences that sift through data, suggest solutions, and automate tasks using advanced tools like code.

Complete AI toolchain:

Foundry offers tools to ground models on specific data, orchestrate complex AI workflows, and evaluate model outputs for quality and safety, ensuring robust end-to-end management. Developers find it easier to integrate and manage AI projects, enhancing productivity and operational efficiency, while users experience more reliable and effective AI applications.

Enterprise-grade production:

The platform facilitates scalable deployment of models, flows, and apps, incorporating continuous monitoring and fine-tuning capabilities within a secure and governed environment. Organizations can scale their AI solutions as needed without compromising on security or performance, providing developers with a flexible and robust infrastructure and ensuring that users enjoy consistent, reliable AI services.

The centralized nature of the Foundry platform provides a collaborative development environment so generative AI teams can efficiently work together and stay in sync. Foundry helps users build AI solutions faster with prebuilt capabilities and templates to ultimately accelerate solution development.

Deploy responsible AI with Foundry

Responsible generative AI refers to the development and deployment of generative AI systems in a manner that is safe, transparent, and accountable. Microsoft offers numerous tools and controls that help with the responsible deployment of generative AI.

Foundry Tools: Developers can utilize pre-built and customizable APIs and models to rapidly build cutting-edge, responsible applications. These services provide detailed Transparency Notes and fairness assessments, such as Face and Speech, to support customer choice and transparency.

Responsible use of AI with Foundry Tools

Azure AI Content Safety: Developers can get support in detecting and mitigating risky content, including prompt attacks, and the generation of ungrounded or copyright material. By using Azure AI Content Safety as a built-in safety system, develops can build more trustworthy applications.

Build AI applications responsibly with Azure AI Content Safety

ASOS uses Foundry to surprise and delight young fashion lovers

What was their goal?

With the prevalence of generative AI platforms, **ASOS**, a United Kingdom-based fashion and cosmetic retailer, saw an opportunity to expand its business model, enrich its technology infrastructure, and meet customers' modern tech expectations.

How did they achieve it?

ASOS used Azure OpenAI and prompt flow, part of Foundry, to quickly streamline their development and testing cycles, helping the customer and the solution interact effectively. ASOS helped developers onboard quickly with built-in prompt flow resources, limiting the need for custom code.

Learn more >

Evaluation and Monitoring in Foundry: In addition to mitigating problematic inputs and outputs, developers can continuously measure the effectiveness of their mitigations during development and in production environments using pre-built evaluation and monitoring metrics for risks and safety. As data and user behavior change over time, these measurement tools help developers and domain SMEs understand their application's behavior and intervene quickly when performance degrades, or an end user is attempting to manipulate the application to behave outside its prescribed purpose.

Evaluation and monitoring metrics for generative AI

Security: Microsoft provides the multilayered security backbone of your applications, using built-in security controls and unique threat intelligence to help identify and secure against evolving threats.

A focal point of Microsoft security is Microsoft Defender for Cloud—Microsoft's cloud-native application protection platform (CNAPP)—which provides comprehensive security to proactively mitigate risks and identify threats to AI applications from code to cloud.

Learn more about **Microsoft Defender for Cloud**

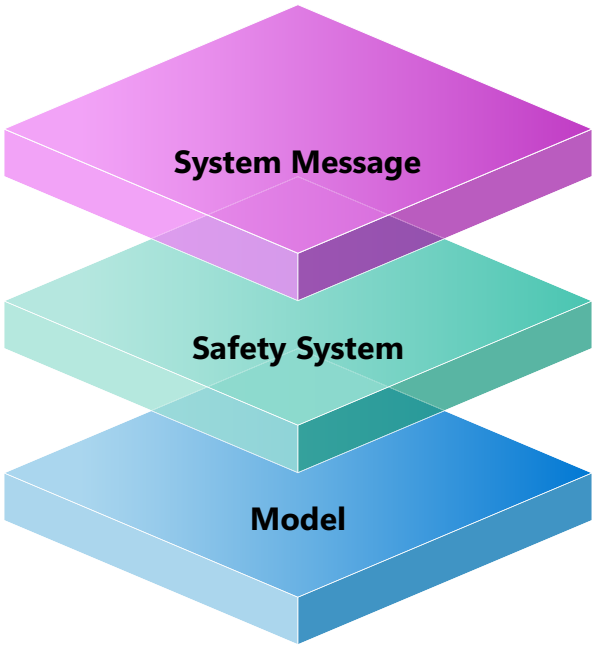
Mitigation layers in Foundry

Using a layered mitigation plan can help developers identify and remediate potential risks earlier in the development process. Here is a brief overview of Foundry's robust safety mechanisms, designed to secure and optimize your generative AI deployments:

System message layer: This layer provides hidden instructions to your model with every user prompt, so you can guide the model's behavior and data retrieval to generate higher quality responses by default.

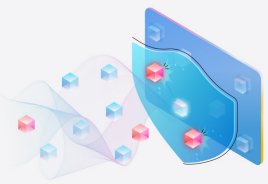
Safety system layer: This added layer goes beyond the basic safety finetuning that is part of the model. Azure AI Content Safety provides this extra layer by running both the prompt and completion of your model through classification models. These models are designed to detect and prevent the output of harmful content across a range of categories and severity levels.

Model layer: With Foundry's model catalog, you can explore benchmarks and model cards for Azure OpenAI, Meta, Hugging Face, and other model developers, all organized by collection and task.



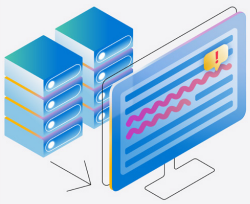
Safety feature highlights in Foundry

Azure AI Content Filters



Prompt shields:

Detect and block prompt injection attacks, including direct and indirect prompt attacks.



Groundedness detection:

Detect "hallucinations" in model outputs, for better accuracy and reliability of the content generated by AI.

Configurable harmful content filters:

Detect four categories of harmful content (violence, hate, sexual, and self-harm) at four severity levels respectively (safe, low, medium, and high), and optional binary classifiers for detecting jailbreak risk, existing text, and code in public repositories.

Safety system message: Steer your model's behavior toward safe, responsible outputs using system message templates developed by Microsoft Research.

Safety evaluations: Assess an application's vulnerability to jailbreak attacks and content risks using your own test dataset or a test dataset generated with AI assistance.

Risk and Safety Monitoring: Understand what model inputs, outputs, and end-users are triggering content filters to inform mitigations.

→ **System message and grounding layer:** This is where retrieval augmented generation (RAG) comes in. Instead of using the model as a source of information, the model serves as a reasoning engine over data sources that are relevant to the query. The system message helps guide the model to use grounding data effectively and can help steer overall behavior for more predictable, responsible model outputs.

→ **User experience layer:** At this level, there is a litany of user-centered interventions, guidance, and best practices that can be provided to users to ensure the system is used as intended.

These enhancements to Foundry's safety features demonstrate Microsoft's commitment to responsible AI, ensuring that your applications are not only effective but also align with the highest standards of data integrity and security.

Another key component of responsible AI is the evaluation and monitoring of your application. We've discussed the generation of quality metrics that Foundry provides, but there are three overarching topics for consideration:

Manual evaluations:

This is the form of evaluation we recommend starting with before moving to automated evaluations. This can be a useful way of tracking progress ona small set of priority issues.

Automated evaluations:

On the other hand, automated evaluations can be more useful for measuring quality and safety on a larger scale and can gain more comprehensive results.

Monitoring:

Monitoring models that are deployed in production is an essential part of the generative AI application lifecycle. Changes in data and consumer behavior can change your applications performance over time. Foundry makes it easy for you to monitor your applications in production for ongoing safety and quality.

Learn more: **Evaluation of generative AI applications**

Trailblazing AI answer engine Perplexity. AI doubles throughput, cuts cost with Foundry

What was their goal?

Startup **Perplexity.AI** is the creator of Perplexity Ask, a revolutionary AI-based conversational answer engine that combines large language models with a robust semantic search engine. To better support Perplexity Ask, they needed a platform that would support faster time to market, serve as a force multiplier for their lean staff, scale to support millions of users, and deliver security and reliability at a cost-effective price.

How did they achieve it?

Perplexity.AI used Foundry to develop their first prototype in just hours. They were able to try out large language models available with Azure OpenAI, getting going "with just a few clicks." Overall, this resulted in running twice as many experiments, in parallel, before they adopted Azure, enabling them to retrain the new version of the model twice as fast.

Learn more >

Integrated data security and privacy

To help ensure data security and privacy, Microsoft makes several commitments to its customers. For one, Microsoft Azure is built on security, privacy, and compliance. When data is stored in Azure, your data remains yours—it is never used for marketing, advertising, or foundation model training purposes without your permission. The prompts and the completions back are still your data.

Microsoft Fabric, Microsoft's unified AI-powered analytics platform, can help reshape and improve how you access, manage, and act on data. On a single platform, Fabric unites data and services to simplify data integration into a single, multicloud data lake for your organization to work from the same data across analytics engines and languages.

In addition, Foundry contains enterprise security configurations, including:

→ Azure Policy integrations

→ Role-based access control (RBAC)

→ Network isolation and security

→ Data protection and encryption

→ Vulnerability management

To tie it all together, when managing data security and privacy, Fabric has a native integration with Microsoft Purview, which enables organizations to govern, protect, and manage their data across their entire data estate. When integrated together, Fabric and Purview enhances AI capabilities through secure data integration.

Microsoft is obligated to defend customers against certain third-party intellectual property claims relating to Output Content. For Azure OpenAI in Foundry Models and any Configurable Generative AI Service, Customer also must implement all mitigations required by the Azure OpenAI documentation in the offering that delivered the Output Content that is the subject of the claim.

Learn more: Introducing the Microsoft Copilot Copyright Commitment

Key insights and next steps

As the demand for generative AI within organizations grows, the need for a centralized platform to develop and deploy these technologies responsibly and effectively becomes critical.

Foundry provides a robust solution, offering a comprehensive platform that caters to developers at all levels, including those without data science expertise. This platform not only simplifies the development process by providing access to a comprehensive model catalog from leading providers but also supports the full lifecycle of AI development with tools for advanced data integration, workflow orchestration, and interactive app

Get started with Foundry →

- → Microsoft Certified: Azure AI Engineer Associate
- → Microsoft Certified: Azure AI Fundamentals
- → GitHub repo for building a RAG application
- → Foundry model catalog
- → Foundry Overview
- → Introduction to Foundry—Training | Microsoft Learn