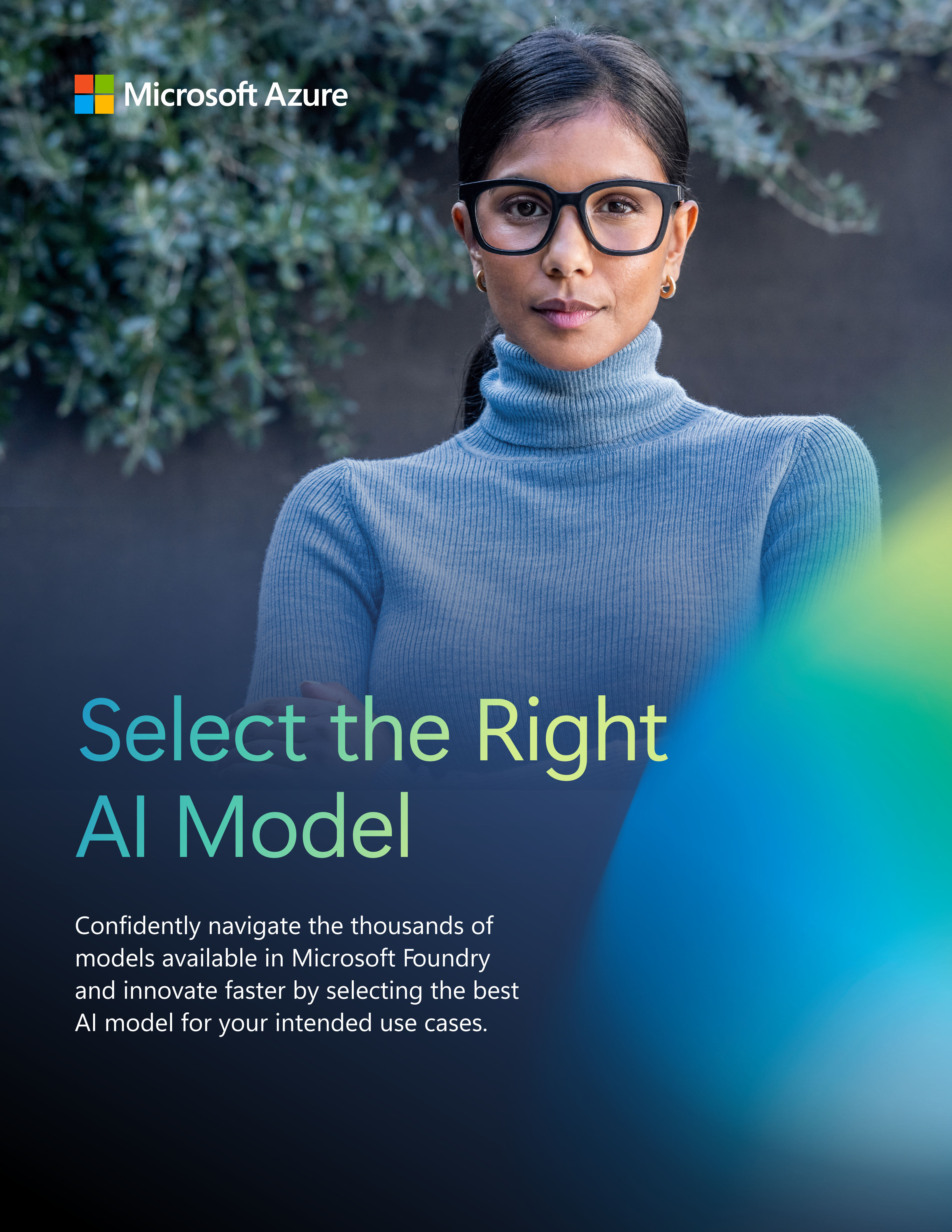
Microsoft Azure



Select the Right AI Model

Confidently navigate the thousands of models available in Microsoft Foundry and innovate faster by selecting the best AI model for your intended use cases.

Contents

Unlock the potential of today's transformative AI models

Innovators everywhere are using AI to shape the future. They're prototyping and designing AI applications and agents faster than ever, transforming customer engagement, employee experiences, and business processes. They're building smart copilots and combining AI agents that not only plan and execute steps in a workflow but also adapt to changes while keeping humans in the loop.

Behind these breakthroughs are AI models trained on massive datasets. The rising tide of clever generative AI content has made models like DeepSeek and GPT-5 familiar names. But there are thousands more, offering a multitude of options for use cases based on performance and cost requirements. It's hard to choose. Models can be general-purpose or task-specific, proprietary or open source, large or small. They come from prominent providers and startups.

The pace of innovation poses its own challenges. One lesson we've learned from our work with customers is that the right model can unlock enormous potential, but one size does not fit all.

Trends and opportunities

Providers continue to rapidly upgrade models and release new ones, pushing the frontier and creating new opportunities. Today, popular foundational models are offered by OpenAI, DeepSeek, HuggingFace, Mistral, xAI, Cohere, Meta, Microsoft, NVIDIA, and Black Forest Labs. Many of our customers choose to integrate these pretrained models into their workflows, fine-tuning them with their own data to hone answers, improve accuracy, and apply domain-specific knowledge.

Another trend is the use of right-sized AI models that balance performance and cost-efficiency, including task-specific and industry-specific small language models (SLMs). Device-specific models are gaining traction in mobile and edge scenarios, as are ultra lightweight nanomodels that can run on microcontrollers.

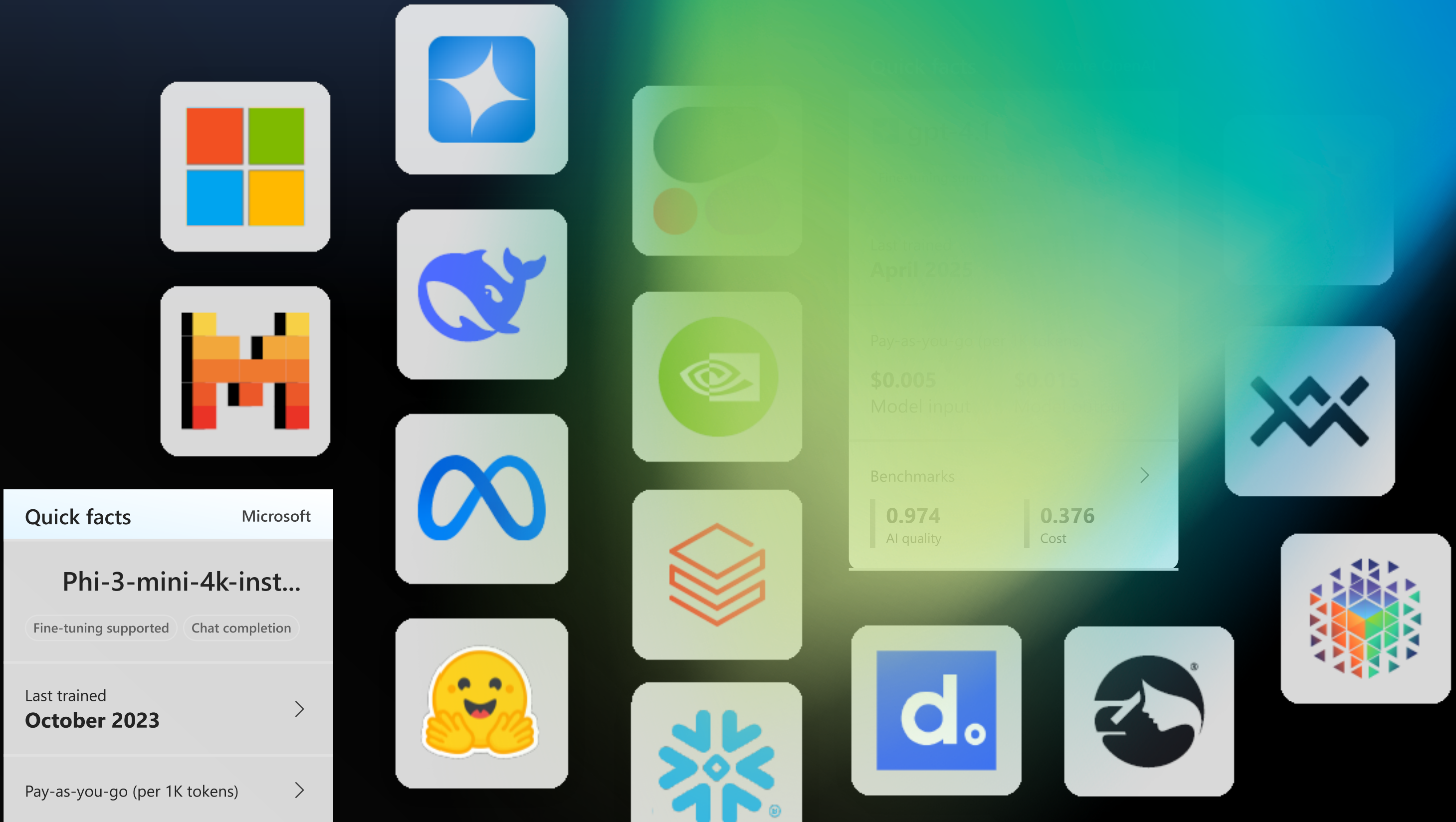
AI models keep progressing—and so do the ways our customers apply them. Until recently, the focus was on implementing single AI models capable of text, audio, and visual processing. Now organizations are turning to multimodal models that support richer analyses and more immersive user experiences. They're orchestrating systems that can learn and adapt continuously, deploying multiple models and combining agents to autonomously handle workflows.

About this e-book

Based on best practices, this e-book shows you the key factors that go into selecting the right AI model for your use case and recommends an iterative approach based on Foundry. As an enterprise-ready platform, Foundry helps you design, customize, and manage your AI apps and agents at scale based on responsible AI principles—whether you're starting your AI journey or a seasoned pro.

Foundry offers more than 11,000 of the latest AI models, including foundation, frontier, and open source—or bring your own. You can discover and filter models by task, provider, and capability, easily compare model and service costs, then benchmark and try out the top contenders. You can even switch to a new model without starting from scratch or writing code.

With so many AI models to choose from, how do you find the right one for your project?



NBA: Real-time experiences with Azure Open AI

> ”
>
> At the heart of NBA Insights is Azure OpenAI, processing immense amounts of data quickly. Fans now are receiving detailed information about their favorite players and teams, making it easier to stay updated on developments during live games.”

Sydney Sarachek
AI Technical Lead, NBA

Use case:
Enhance fans' experience of National Association of Basketball (NBA) games using data to provide real-time insights behind all the great plays.

Models:
Azure Open AI - GPT-4o series models

↑ time to market

↑ user experience

Read the full story →

1.
Start with your use case and explore

Evaluate, test, and monitor models in one place to make an informed selection

Your use case influences every decision going forward—from model type and data requirements to resource allocation and evaluation metrics. As you begin considering AI models, it's common to filter your choices by prioritizing cost, throughput, or performance. For example:

- **High-stakes applications** such as healthcare diagnostics and fraud detection require models with accuracy, interpretability, and data availability.
- **Real-time conversations** require fast AI models with acceptable precision that run affordably and integrate smoothly.
- **Enterprise predictive analytics** require accurate, scalable AI models with cost-effective performance.

Model selection is an iterative process of weighing the trade-offs as you prototype, optimize, and operationalize your AI apps and agents. The most popular AI models support a broad range of capabilities and multimodal inputs with low cost and latency. Both large and lightweight models provide reasoning, problem-solving, and compliance with responsible AI principles.

A typical approach is to adapt a powerful foundational model to your use case, fine-tuning it with your own data or using retrieval-augmented generation (RAG) to train it on your knowledge base. When you start with a model from a flagship provider, you get the most advanced version with the best intelligence. After prototyping and testing, you may want to choose something else, including one of the next-generation frontier models. However, flagship models give you a strong, stable starting point to evaluate feasibility.

To help you get started, Foundry Model Catalog brings the latest open-source and foundation models under one roof. You can experiment with any of the major model families leading the market today.

By the numbers[1]

8 months

Average genAI deployment time

13 months

Average time to realize value

10x

Average return for every $1 a company invests in AI

2. Map tasks to models

Can AI solve your use case?

Find any model you need—from OpenAI to xAI

ai.azure.com/explore/models

Can AI solve your use case? To find out, you must frame the question in terms of the capabilities you need. Try asking the following:

- As input, does the model need to handle natural language processing (NLP), audio, computer vision, multimodal input, or something else? What output will it generate—text summaries, decisions, transcriptions, predictions?
- How complex does the reasoning or interaction need to be? Do you need fine-tuning to match the required precision and domain knowledge, or do you need high performance on general knowledge across diverse fields?
- What are your inference speed requirements?

With this framing, you can focus only on models that are functionally aligned from the start, before evaluating other factors, like resource consumption and compliance.

MARS

Mars Science & Diagnostics: Better animal health outcomes with Mistral

”

The Azure AI catalog provides access to a wide range of prebuilt models such as Mistral to help restructure data and enhance our accuracy. We know it’s accurate because nothing goes into our production systems without being validated and signed off by radiologists.”

Mark Parkinson

Sr. Director of AI Development, Mars Science & Diagnostics

Use case:

Automate and accelerate veterinary diagnostic processes

Models:

Mistral, Foundry Models

38%

improvement in precision

96.9%

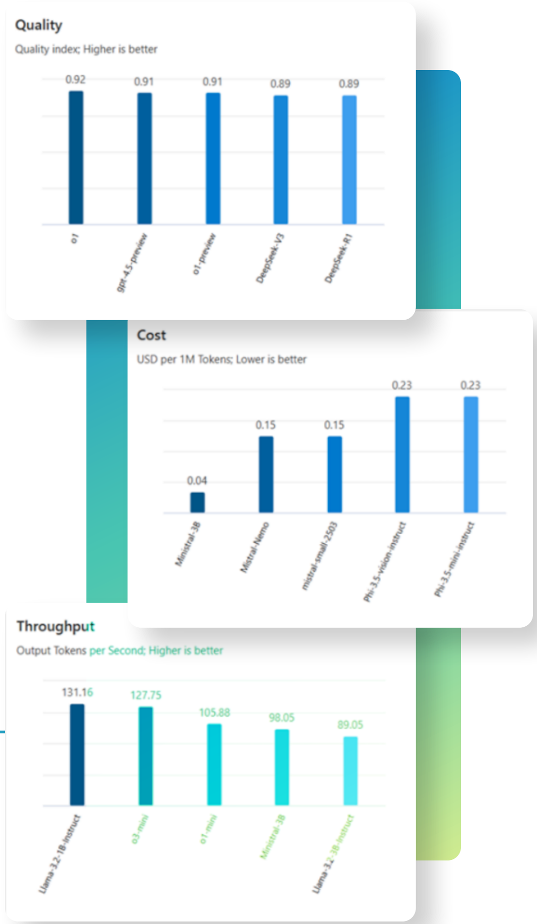
accuracy on test datasets

Read the full story →

3. Benchmark, prototype, and optimize models

These days, it’s easy to use benchmarks to compare AI models and identify potential candidates. Standardized industry benchmarks help you assess an AI model’s general capabilities like intelligence, reasoning, or code generation. Application-specific benchmarks can tell you how well a model performs in a real-world context, such as healthcare, finance, ecommerce, or security.

However, raw benchmarks aren’t everything. Two models may perform well against your benchmarks but behave differently in your app. The right model balances your priorities and constraints with your budget.



Leaderboards based on public datasets help you narrow down the choices. Quality is the most common criterion for model selection, followed by safety, cost, and performance. You can also use your own data to evaluate models in Foundry.

Performance and accuracy

Measures of performance help you make the most of a model's unique architecture and training methods. To align a model's strengths with the needs of your use case, you need both quantitative measures of performance and qualitative metrics of context handling and safety.

For genAI models in particular, *evaluators* are essential. These tools assess the quality, safety, and reliability of AI responses. For example, you must assess a model's *ground truth*—how well it gives the correct answer. Foundry supports general-purpose, RAG, agent, and many other evaluators to help you assess a model's fit and trustworthiness.

Best practices

Define your acceptable thresholds for performance metrics, such as latency, throughput, accuracy, precision, and recall.

Compare the top contenders based on your key metrics for quality, cost, safety, and throughput. Try browsing model leaderboards in Foundry to quickly surface leaders and run side-by-side comparisons.

To assess the correctness of a model's output, use exact match (EM) scores, F1 scores, and recall and precision metrics.

Measure the reliability of answers using risk and safety evaluators that detect bias, hallucinations, and fairness.

To compare the semantic quality of models, use textual similarity evaluators—vital for text generation, summarization, and translation tasks.

Perform load testing to make sure your AI model can handle the volume and variety of production-level data. You can use Foundry to simulate real-world workloads, stress-test your model, and make sure performance and accuracy are maintained.

Use A/B testing to compare different model versions using your scenarios and data. Continuously test to see which models provide the right balance of impact, risk, and cost.

Use Foundry model router to select and maintain the best models over time. For example, use prompts to query Foundry model router and see which models it uses to respond based on factors like query complexity, cost, and performance.

Resource management

AI models use virtual and physical resources whether they run on an edge or local device, on-premises servers, or in the cloud. The constraints of your deployment environment determine the type and size of model you can run.

Cloud-based deployments can use large foundational models, while smaller nanomodels suit on-device deployments and ensure that no data leaves the device.

Best practices

Keep costs and performance in balance by verifying a model's compute requirements, memory footprint, and loading times.

Make sure your infrastructure can support increases in demand without racking up costs or wasting resources. Foundry can help you monitor and optimize the computational resources required for training and deploying your models.

Factor in any plug-ins or multiple-agent setups needed for models that reason, plan, or use tools.

If you're a GitHub developer doing early testing, start with GitHub models for quick, low-cost experimentation, then move to Foundry for scalable, secure deployment when you're ready to go to production.

For edge or offline scenarios, start with Foundry Local models like the Microsoft Phi family for on-device inference, then scale to Foundry for hybrid enterprise deployments with Azure Arc when you need centralized management and cloud integration.

Costs and benefits

Budget always plays a role in assessing whether an AI model is the right fit for your needs. Freely available open-source models are popular, but a model's resource use and compute instances dramatically affect the overall runtime and maintenance costs—in addition to any other services you deploy. Plus, training a model using your own data usually means a bigger budget and infrastructure.

Here's a common-sense approach:

$ **Use open source.** You can experiment and prototype a minimum viable product (MVP) with less risk. As your project evolves, you can switch to another model.

$$ **Try a foundational model from a leading provider.** Pretrained models offer cost-effective operations at scale and reliable results across a wide range of applications.

$$$ **Opt for largest and most powerful models** available and run them on hosted hardware or using top-tier API services for the best performance and capabilities.

Best practices

Compare models on usage-based, per-token pricing but look further. Providers offer different pricing structures based on use case, deployment mode, and model type. For example, Foundry supports token-based Pay-As-You-Go pricing, provisioned throughput units (PTUs) for capacity reservation, third-party model pricing, region-specific pricing, and potential discounts for certain usage patterns.

In addition to inference costs per request, assess the cost per inference at scale should the model need to handle more users or data over time.

Include the costs of customizing or fine-tuning a model. Work with providers that offer a low-cost window for experimentation and fine-tuning, like the Developer Tier in Foundry.

To optimize overall costs, take advantage of serverless or autoscaling infrastructure in your applications and agents.

Compare hosting options. For example, Foundry features Standard, Global Standard, and Regional provisioned throughput, each with its own pricing structure that may include hourly rates or token-based charges.

For predictable workloads, try reserving capacity in advance. Foundry supports 1-month and 1-year reservations. Some Foundry customers have saved up to 70% with annual PTU reservations.[2]

Monitor costs to help identify spending trends and potential overspending. For example, use the cost analysis tools in Microsoft Cost Management, available to anyone with access to a billing account, subscription, resource group, or management group.

Flexibility and adaptability

Models that you can train, fine-tune, and distill give you the flexibility to customize outputs for your target audience, use case, and industry, and even improve performance. For example, you can refine a model's training data based on your use case to boost accuracy and deliver results faster. You can then use reinforcement fine-tuning (RFT) to improve a pretrained model's reasoning abilities.

Many models support lightweight customization through RAG, low-rank adaption (LoRA), adapters, and prompt tuning. Some models include an API, command-line interface, or user interface for further fine-tuning. You can use Foundry to determine the level of customization a model supports and safely test models in evaluation playgrounds.

Best practices

To enhance precision in a specialized domain, choose a model that can be fine-tuned to match the specific knowledge needed.

To improve output using RAG, look for models that you can connect to external data sources or that support knowledge retrieval.

To improve performance, choose a model that can be quantized, pruned, or distilled without a noticeable drop in accuracy.

Check how much expertise is required for tuning. Some models include simple interfaces that don't require a data scientist to use.

Test whether your customization and fine-tuning efforts are, in fact, better. For example, use the Azure AI Evaluation SDK to evaluate model outputs using model-based graders, custom rubrics, and structured scoring—all from code.

For development ease, use models that can be deployed as API endpoints.

Make sure the model is compatible with other tools and systems in your stack.

Safety and compliance

Before shipping any solution, you need to ensure that your model adheres to compliance and security standards at a larger scale. Safety includes protecting user data, preventing unauthorized access, and complying with industry regulations.

Foundry offers built-in security features and compliance certifications that help you follow regulatory and data protection standards, such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA).

Best practices

Choose a model that handles data appropriately given your regional and regulatory requirements.

To preserve data privacy and sovereignty, choose open or licensed models that support local inferencing.

Use Foundry and Responsible AI tools to assess and configure how AI models handle sensitive or personal data, ensuring compliance with privacy and ethical standards.

Extend secure deployment using a private infrastructure through compatible services like Azure Arc or Azure Kubernetes Service (AKS).

Implement end-to-end security, observability, and governance with controls and checkpoints at all lifecycle stages. Foundry provides a unified AI development toolchain that helps you continuously monitor and optimize AI performance.

Consistency and innovation

Selecting a model backed by robust research and a strong foundation ensures that your AI apps and agents can take advantage of the latest advancements as they emerge. It's a way of future-proofing your agents and apps, because you get the advantage of the model provider's ongoing improvements.

However, you may need a model with long-term consistency to meet regulatory requirements or to support product stability. For example, customer support scripts, compliance tasks, and clinical decision-making rely on predictable behavior, so you don't want surprise changes in how a model works.

Best practices

Choose models that reflect responsible AI practices. Foundry provides tools and resources to help you apply the Microsoft Responsible AI standard for security, safety, privacy, fairness, transparency, and accountability.[3]

Determine how often a model is updated or retrained and the provider's version control and rollback policies.

Consider vendor stability. Is the model provider likely to be around three to five years from now?

For long-term consistency, use a model specifically designed for stability and reliability. Models sold directly by Azure typically offer enhanced integration, optimized performance, and direct support from Microsoft—factors that contribute to more predictable behavior.

To handle the unexpected, incorporate your model into a workflow that supports continuous integration and continuous deployment (CI/CD) pipelines. You can update and improve your models as requirements evolve.

DraftWise

DraftWise: Choosing Cohere for legal contracts

> Foundry Models is an absolute game-changer. It's ignited development for us, improving developer efficiency by 60% over traditional methods."

James Ding
DraftWise founder and CEO

Use case:
Build a highly accurate legal platform, tailoring model behavior to legal tasks

Models:
Cohere family + RFT

60%
more efficient development

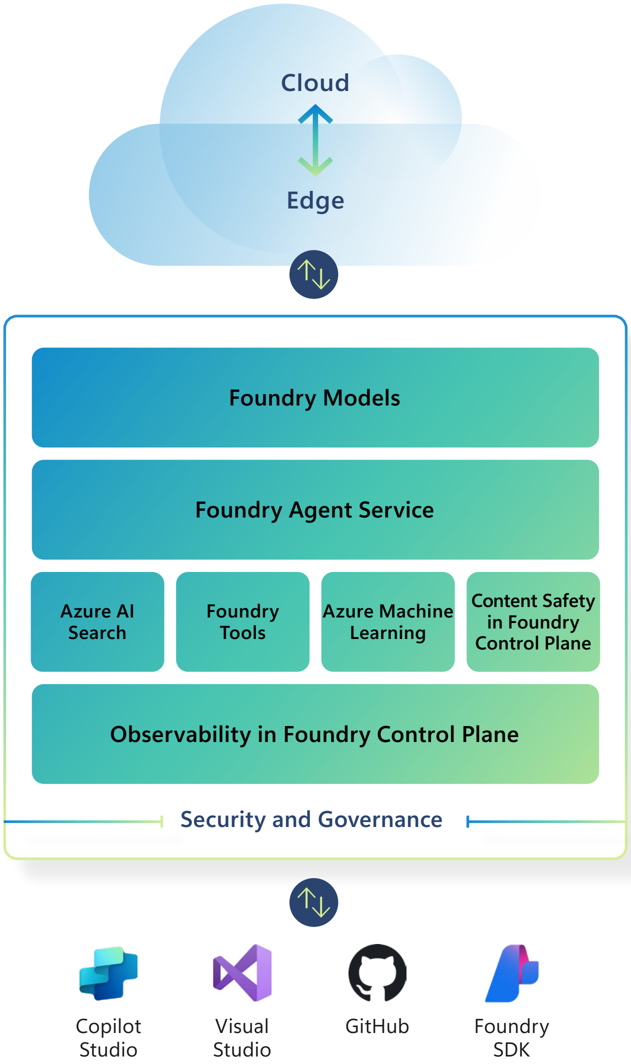
30%
better search results

Read the full story

4.
Design with the best models using Foundry

Foundry is your starting point

Whether you're building AI agents, automating industry workflows, or launching retrieval-augmented systems, Foundry provides end-to-end support across the development lifecycle. Built for modern enterprise needs, Foundry helps you find the right models at the best price with the least friction—so you can deliver AI apps and agents at scale.



Experiment in a controlled environment

As AI workloads grow more complex, you need self-service development tools that streamline the end-to-end lifecycle of AI agents and applications.

- **Work in on-demand, zero-setup** Foundry playgrounds to rapidly prototype, explore APIs, and validate feasibility.
- **Build what you want using the tools you prefer**—from low-code Microsoft Copilot Studio agents to pro-code with Foundry APIs and SDKs, GitHub, or Visual Studio Code.
- **Deploy a new model** to the same endpoints or applications with minimal changes—the Azure AI model inference API helps you quickly test and validate new models.
- **Use self-service tools** like the Foundry resource and the Foundry API to get apps and agents into production faster, at scale, and with central governance.

Available through a unified portal, SDK, and API, Foundry helps you go from prototype to production faster and develop responsibly with safety, security, and privacy. Integration with popular development tools and the Microsoft security ecosystem help you operationalize and monitor your workflows.

Next steps

Start building your AI solution today.

As a recognized leader in AI development, Microsoft supports responsible AI principles and provides the integrated services you need to build exceptional generative and agentic AI systems.

Create with Foundry

Get started with Foundry, and jump directly into Visual Studio Code

Download the Foundry SDK

Take the Foundry learn courses

Review the Foundry documentation

Keep the conversation going in GitHub and Discord

Sources

[1] The Business Opportunity of AI. IDC. November 2024.

[2] Azure reservations

[3] Responsible AI at Microsoft