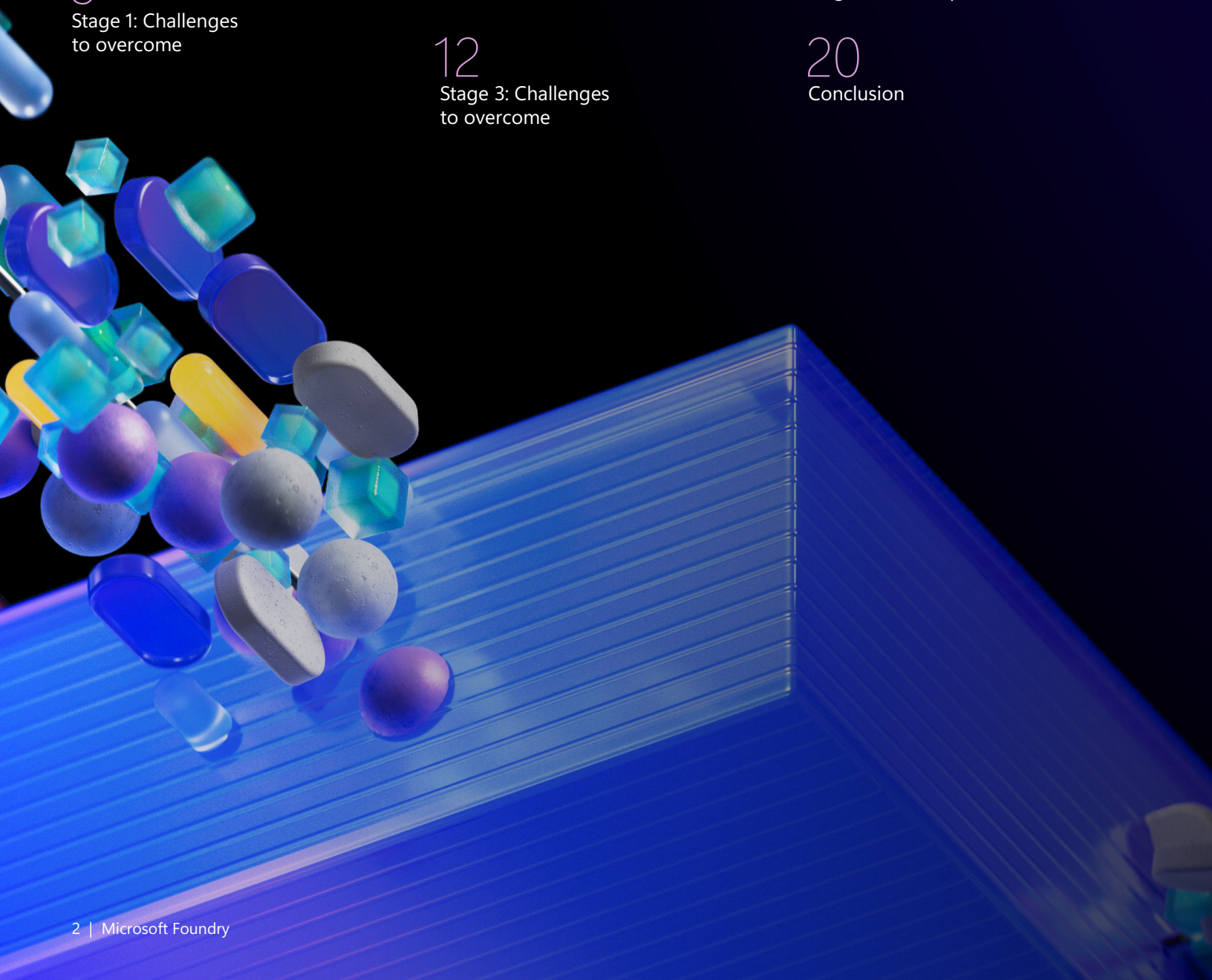
Microsoft Azure

The AI adoption journey:

Moving from AI pilots to transformation at scale with Microsoft Foundry

Contents



How can organizations bridge the gap?

Nearly every organization is experimenting with AI. In fact, around 80% of companies have piloted generative AI or language model projects – yet only about 5% have integrated these into production at scale.[1]

The gap between promising proofs of concept and fully scaled solutions is the crux of the AI adoption challenge. Yet, for organizations that successfully navigate this transition, the rewards can be substantial.

- Organizations with advanced genAI readiness outperform their peers by 47–64% across operational efficiency, customer experience, innovation speed, workforce productivity, and revenue growth.[2]
- While only 17.7% of organizations are considered genAI leaders, these leaders generate 56% more value from AI than their peers.[2]

Accelerate AI adoption by progressing through four key stages

Microsoft Foundry is designed to help organizations bridge the gap, with most organizations typically progressing through four key stages on their AI adoption journey:

Stage 1: Experimenting and early pilots

Technologies in focus: Foundry Models and Foundry Tools

To validate use cases, teams typically start their AI journey with generative models (e.g., Azure OpenAI in Foundry Models) and Foundry Tools like speech-to-text in Azure Speech or ORC in Azure Vision.

Stage 2: Grounding AI with enterprise data

Technologies in focus: Foundry IQ and Azure Machine Learning

Organizations can enhance initial pilots by connecting AI to internal knowledge bases and documents via retrieval-augmented generation (RAG) for more relevant results. They can also leverage Azure Machine Learning to build, train, and fine-tune custom models.

Stage 3: Building intelligent agents and workflows

Technologies in focus: Foundry Agent Service

Most businesses don't want just chatbots – they want automation that's faster and with fewer errors. That's why organizations can leverage Foundry Agent Service, which connects the core pieces of Microsoft Foundry, to evolve simple chatbots into multistep agents that can take action and integrate business processes and systems.

Stage 4: Deploying and scaling in production

Technologies in focus: Microsoft Azure

At this final stage, organizations are empowered to deploy AI solutions at enterprise scale with robust observability. Having used Azure Machine Learning to train custom models, they can deploy solutions to Azure App Services, Azure Container Apps, and Azure Kubernetes Services to achieve a variety of scenarios. Organizations are also encouraged to leverage Foundry Control Plane to enable continuous monitoring and governance.

Each stage of the AI adoption journey comes with distinct goals, challenges, and best practices. In the chapters ahead, we will explore each stage, incorporating direct insights from customer interviews and highlighting exemplary use cases. We'll also provide strategic recommendations throughout to help technical decision-makers deploy and provision the right models and agents, making it easier for organizations to effectively implement pilots and scale solutions to production.

STAGE 1

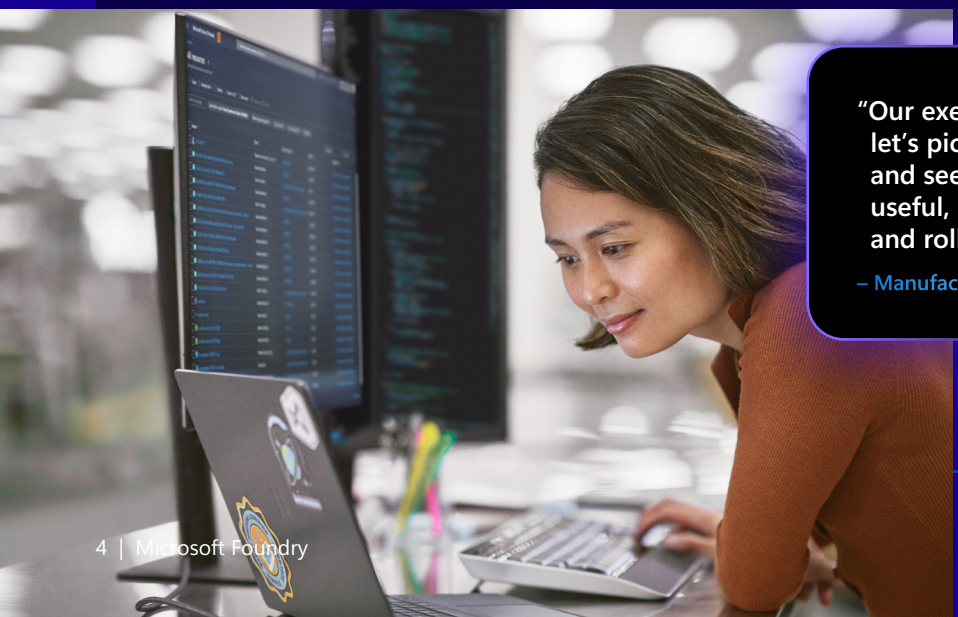
Experimenting and early pilots

Organizations begin their AI adoption journey by running small experiments with generative AI models and tools to see what's possible. Organizations might be startup developers or pilot teams in larger enterprises with high enthusiasm, but limited scope beyond an initial use case.

Many journeys begin with top-down enthusiasm (driven by AI buzz), leading to a manageable pilot project to test what's possible. Choosing the right first use case is critical – it should be something impactful enough to demonstrate value, yet feasible enough to implement quickly. Early successes often include an internal knowledge chatbot to query company policies, an AI tool to summarize reports or support tickets, or a simple workflow helper to assist with tasks like generating draft emails.

Common use cases to explore

- **Experimenting with Foundry Models:** Teams often begin by calling an OpenAI model (e.g., GPT-4.1, GPT-5 in Foundry Models) to build a chatbot prototype or text generation demo. For many, this means playing with prompts in a sandbox. An internal Microsoft study shows that most development teams using Azure OpenAI for the last 1.5 years typically start with a small internal project. In addition to models from Azure OpenAI, Foundry Models offer the widest selection of models on any cloud, with more than 11,000 foundational, open, reasoning, multimodal, and industry-specific models.[3]
- **Trialing Foundry Tools:** Early trials often include simple uses of Foundry Tools, like form recognition and translation, as standalone tests. For example, one automotive firm first tried Azure Translator in Foundry Tools to handle multilanguage report data. According to the organization, "The first time they used an Azure OpenAI endpoint, they found it easy to integrate with Power BI." This enabled them to quickly process data in different languages.[4] Additionally, a financial services team stood up an internal chat on their intranet for document Q&A and reported that "users were excited to not have to do monotonous tasks like summarizing PDFs" once they tried it.[3]
- **Piloting low-risk internal scenarios:** Nearly all initial AI solutions are internal-facing. A recurring theme is starting with an employee-facing Q&A bot or document summarization tool. One customer noted they chose an internal pilot because "it gave us time to learn and establish best practices and evaluation procedures," rather than immediately exposing AI to customers.[3] For example, a general-purpose internal chatbot deployed in Microsoft Teams (often replicating ChatGPT for company FAQs) is a popular Stage 1 project – especially in manufacturing firms.



"Our executives, during discussions, thought, 'Okay, let's pick one particular area, one small use case, and see how it works out for us. See if it's really useful, see what the ROI is …' We quickly built it and rolled it out. That's how we started."

– Manufacturing ISV, Enterprise (on launching their first AI pilot)

STAGE 1

Challenges to overcome

At this exploratory phase, teams frequently face a steep learning curve. They must understand model prompting, evaluate output quality, and avoid pitfalls (like inadvertent data leakage or unexpected API costs). Interviewed customers admit they "have a lot to learn" and sometimes are unsure if they're leveraging AI tools correctly. However, many experiments stall because they weren't tied to a real business need.[3]

Strategic recommendations

- **Start with a targeted, low-risk use case:** Identify a genuine business problem that a small generative AI solution could help (e.g., internal knowledge lookup, automating part of a workflow). Start with a targeted use case – successful implementation of one use case often sets the foundation for expansion to others. Keep the scope focused to deliver a quick win.
- **Leverage the quickstart AI solution templates in Microsoft Foundry:** To accelerate development, Foundry offers customizable, out-of-the-box code samples, pre-integrated Azure services, and GitHub-hosted quick-start guides. Development teams can skip setup and focus on outcomes (like Getting started with AI chat).
- **Engage with Microsoft resources:** Early-stage users benefit from documentation, guidance, and community support. Microsoft's account teams, engineers, and the community forums on GitHub and Discord can offer pointers to ensure an AI project is set up for success.

Early successes with Microsoft Foundry often translate directly into developer productivity gains. As one CTO in the software sector shared, "Our engineering teams use Microsoft Azure AI Foundry [now Microsoft Foundry] with [Microsoft] Copilot for coding assistance, and it saves them 20–30% of their time."[5]

Another leader noted, "Our developers are saving anywhere from 10% to 40% of their time with Microsoft Azure AI Foundry [now Microsoft Foundry]."[4]

In another case, Assembly Software used Foundry with Azure AI Search to modernize its legal case management solution, enabling lawyers to retrieve case facts, filings, and precedents instantly, cutting hours of research time per case.[6]

Read Assembly Software's success story

By the end of Stage 1, a successful team should be able to create a working prototype or pilot demonstrating AI value.[3] Equally important, the team gains confidence and organizational buy-in to move forward.

Typically, however, the pilot reveals that to be truly useful, AI solutions need more domain-specific knowledge and better integration. This naturally leads to Stage 2.

STAGE 2

Grounding AI in enterprise data

After the initial trial proves the concept of generative AI, the next step is making AI responses relevant and accurate for the business. This means feeding the models with the right context and data. This is where Foundry IQ (powered by Azure AI Search) comes into play. Stage 2 of the AI adoption journey focuses on connecting AI to the organization's knowledge – documents, databases, and domain information – so that the system can move from prototypes to solving real business needs.

This stage corresponds to customers who use a mix of AI tools. These organizations have typically moved past experimentation and have an early production use case in mind, often augmenting an existing product or internal workflow with AI. For instance, a midsize ISV might have added an AI-powered search feature to its software, or an enterprise IT team might use Foundry Models plus Foundry IQ to power an internal knowledge base.[3]

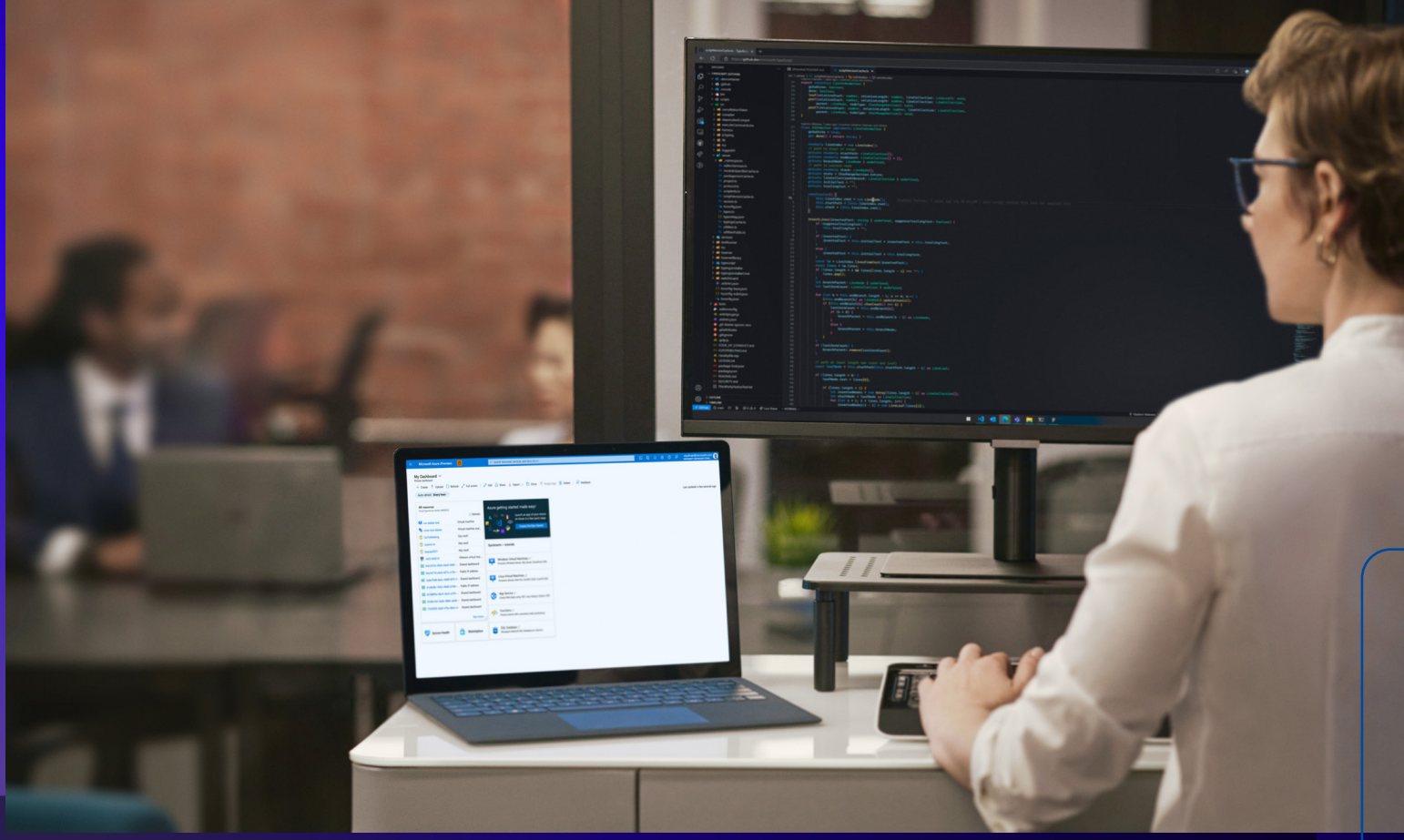
The hallmark of Stage 2 is securely attaching enterprise data to the model, typically through a retrieval-augmented generation (RAG) pattern. At this stage, organizations can also leverage Azure Machine Learning to build, train, and fine-tune custom models using their own internal datasets. Azure ML supports the full machine learning lifecycle – including data preparation, experimentation, and deployment – helping operationalize AI projects securely and at scale. Integrating Azure ML with knowledge-driven applications further enhances the relevance and impact of AI across the business.

Common use cases to explore

- **Enterprise Q&A and knowledge bases:** Companies can integrate their internal documents into the AI solution. Using Foundry IQ, they can index sources like policy PDFs, SharePoint files, wikis, product manuals, databases, and more. Users can then ask natural language questions and get answers grounded in those sources. A healthcare provider, for example, built a centralized repository of provider data and implemented a search-driven chatbot for queries like "Show me all diagnostics providers within 15 miles of ZIP 12345." Essentially, they turned a static database into an interactive Q&A assistant.[3] This drastically improved how employees retrieve information. For example, an executive vice president of a financial services institution described, "Our focus with Microsoft Azure AI Foundry [now Microsoft Foundry] has been to develop tools to help our sales team find answers to common questions. Now, salespeople can save 20–30 minutes a day searching for information and can spend that time in client conversations. If we give salespeople back a few hours a week, they can potentially cover three, four, five more clients."[4]
- **Retrieval-augmented generation (RAG) in apps:** Many started with a simple chatbot in Stage 1, and then evolved it by injecting relevant context from business data. A financial services team described that once they stood up an internal chat, users immediately wanted it to reference more internal content. By connecting their models to an Azure AI Search index of intranet documents, they enabled the bot to provide company-specific answers rather than generic ones. This pattern – combining a model with a vector search over internal data – is a core scenario and one of the best use cases of this stage.

"We created a couple of webpages that incorporated our SharePoint site for easy search… Once it was up, users started asking IT to reference more things – they were excited not to have to do monotonous tasks like summarizing PDFs."

– Financial Services customer (on evolving an internal chatbot with RAG)[1]



- **Domain-specific knowledge assistants:** Beyond general FAQs, organizations build assistants tailored to specific domains using their internal data. For example, an engineering firm created an AI assistant on top of their employee expertise database stored in Azure Cosmos DB. By indexing resumes and past project data, they enabled queries like, "Who in our company has experience in cable bridge design in California?" – which helped staff project teams more quickly.[3] This use case shows how attaching the right corpus, in this case employee skill profiles, can unlock a novel solution for a business.
- **Enhanced document processing:** Some Stage 2 efforts involve pairing models with enterprise documents for deeper analysis. One consulting company built a system to summarize lengthy accident reports, measuring more than 150 pages containing sensitive data. They used Foundry IQ to ingest the report content and Azure OpenAI to generate summaries, taking care to do this on Azure for privacy, since the data contained private health data. The result was an internal tool that dramatically reduced the time lawyers spent extracting key points from reports.[1]

> **"This first project has really been a learning experience for us… We built a foundational knowledge retrieval system with Foundry Models and tools (Foundry IQ) to make an AI solution that could be applied to different use cases around the firm, internally and externally."**
>
> – Engineering ISV, Enterprise (on their RAG knowledge system)[1]

These quotes highlight two key outcomes of Stage 2: better answers through grounding and a growing vision for AI's role. Once the AI is connected to trusted data, its responses become far more useful and accurate for the business. Employees quickly gain confidence in the tool ("it knows our stuff now"), which drives demand for broader usage. In our interviews, many customers at this stage discovered additional opportunities once they attached knowledge to one use case – planning to reuse their "knowledge retrieval system" for other scenarios and even for customer-facing apps.

STAGE 2

Challenges to overcome

Connecting models with enterprise data isn't without hurdles. Companies often struggle with data preparation – like cleaning and ingesting documents, dealing with permissioning (ensuring the AI only shows data the user is allowed to see), and handling knowledge updates. There can also be a learning curve in optimizing the RAG pattern: setting the right chunk size for indexing, managing vector embedding costs, and avoiding irrelevant results. Many users are still refining these details, often in a trial-and-error fashion.

Beyond these foundational steps, model fine-tuning and distillation are emerging as critical activities in Stage 2. Fine-tuning involves adapting pre-trained models to your organization's proprietary data, such as Q&A logs, domain-specific documents, or customer interactions. This process can yield significant accuracy improvements – one customer saw a 30% accuracy jump after fine-tuning GPT on their internal data.[7] Distillation further optimizes models for production, reducing resource requirements while maintaining performance, making it easier to deploy AI solutions at scale.

Strategic recommendations

- **Bring your data early:** One best practice is to incorporate proprietary data into your AI project as soon as possible. Microsoft's experts note that grounding AI in internal knowledge can yield significant accuracy gains. (One customer saw a 30% accuracy jump after fine-tuning GPT on their Q&A logs.)[3] Use Foundry IQ to connect to data sources (documents, wikis, databases) and keep that index updated. This helps to ensure your pilot doesn't stagnate with generic answers. Leverage Foundry tools to fine-tune models with your organization's data. As one CEO explained, "We have seen both accuracy improvements and performance improvements when we fine-tune models with Microsoft Azure AI Foundry [now Microsoft Foundry]." Another technology leader added, "We definitely see an improvement in model accuracy with Microsoft Azure AI Foundry [now Microsoft Foundry] with the proper data. It's an easy way to do the fine-tuning work, but you need to know how to fine-tune and have the right data scientists."[4]

- **Data governance and security:** As you ingest enterprise content, leverage Azure's security – set up role-based access so the AI only serves data that the user is permitted to see. Foundry IQ integrates with Microsoft Purview and Entra ID, which customers in this stage found essential for clearing compliance checks. Also, use content filters to avoid exposing any sensitive text inadvertently.

- **Experiment and iterate:** Successful model customization depends on robust data preparation and ongoing experimentation. Teams should test different chunk sizes, embedding strategies, and evaluation metrics to optimize both RAG and fine-tuned models for their specific use cases.

- **Iterate on relevance tuning:** RAG solutions improve with feedback. Monitor what users ask and whether the answers satisfy them. If the AI is pulling wrong docs or missing info, try adjusting your Foundry IQ settings (e.g., add custom synonyms, tune scoring) or enriching your index with additional metadata. Some organizations set up a regular review of AI query logs to continuously refine the knowledge base and prompts, which is a smart habit to develop in Stage 2. Azure Machine Learning could be leveraged to build, train, and fine-tune custom models using their own internal datasets. By combining Azure ML with enterprise data sources, teams can create more accurate, domain-specific AI solutions that go beyond out-of-the-box capabilities.

- **Show quick wins with a specific scenario:** To encourage adoption of the knowledge-powered system, evangelize a particularly successful use case. For example, highlight how a salesperson used the internal assistant to find an answer in seconds that previously took hours digging through files.[3] These stories build confidence and entice other teams to leverage the new knowledge-driven AI tool.

Cognizant developed specialized machine learning models with Azure Machine Learning to automate software delivery and testing pipelines, reducing cycle times and boosting quality across thousands of enterprise projects.[8]

Read more about Cognizant

In another case study, YoungWilliams used Azure Machine Learning and Azure Document Intelligence in Foundry Tools to automate public assistance case management, transforming manual workflows into intelligent pipelines. By training models to extract and classify structured data from thousands of forms and faxes, the organization now routes cases to the right departments automatically, accelerating response times and improving service delivery for families nationwide.[9]

Read more about YoungWillaims

By the end of Stage 2, the organization's AI project has typically evolved from a "cool pilot" to a truly useful assistant, thanks to the infusion of enterprise knowledge and model adaptation. Users trust it more ("it knows our policies, our data"), and usage often increases. At this point, many companies start thinking beyond Q&A to "What else can the AI do besides just answering questions?" The logical next step is to have the AI take action and handle more complex tasks, which leads to Stage 3: Building intelligent agents and workflows.

STAGE 3

Building intelligent agents and workflows

Having a knowledgeable AI that can converse is powerful, but organizations soon want to push further – can the AI do things on our behalf? Stage 3 is about transforming AI from a passive assistant into an active agent that can autonomously and with human oversight perform tasks, integrate into workflows, and orchestrate multiple steps or tools. This is enabled by Foundry Agent Service, which guides development teams to build multistep conversational agents with tool use, API calls, and logic, and connect them easily to other Azure services or custom APIs.

At this stage, companies are often transitioning to a more scaled stage, meaning AI is becoming a key part of their digital transformation and more departments or products are coming on board. These organizations typically have multiple concurrent AI projects. The focus shifts to expanding AI capabilities and more Azure services to deliver end-to-end solutions.

Common use cases to explore

- **Domain-specific AI agents:** Organizations develop agents tailored to specific workflows or roles. For example, a customer service department might build an "incident triage agent" that not only answers questions but can log tickets or send follow-up emails. In one case, a retail software provider wanted to automate email support: they created an email triage agent that reads incoming messages, consults order data (via Foundry IQ), and drafts appropriate responses or routes issues to the right team. This agent goes beyond Q&A to take action (classifying and responding to emails) that a human agent would normally handle.[3]

- **Tool integration and automation:** Foundry Agent Service is built for interoperability, enabling organizations to create useful agents that understand and act on real-time business data and events across their ecosystems. At the heart of this capability is support for Model Context Protocol (MCP) – an open standard that acts as a universal connector for agents. MCP standardizes how agents discover and invoke external tools, APIs, and data sources. This allows agents to dynamically call tools hosted on any MCP-compliant server – whether it's querying a database, invoking a workflow, or fetching contextual data – without writing brittle, custom integration code. Organizations can reuse MCP-enabled tools across projects and enforce security policies consistently, all while delivering richer, more integrated experiences. Developers can find, connect, and manage public and private MCP tools for agents from a single, secure interface within Foundry Tools. Foundry Tools also provides access to pre-built tools for over 1,400 business systems like SAP, Salesforce, and HubSpot; data sources such as Bing, SharePoint, and Microsoft Fabric; and out-of-the-box capabilities for transcription, translation, voice, and intelligent document processing to support rapid agent development.[10]

- **Multi-agent orchestration:** Some advanced scenarios coordinate multiple agents or a chain of reasoning steps. Foundry Agent Service allows one agent to break a complex job into parts or call specialized helper agents. For example, a domain-specific assistant might have one sub-agent that fetches data and another that interprets results – all coordinated under a parent agent. This capability is being explored especially by customers who are pushing toward autonomous AI workflows. Foundry has invested in open frameworks (like the emerging model orchestration standards) to support interoperability. Building on this, the Microsoft Agent Framework – an open-source SDK and runtime – supports advanced orchestration patterns from research (e.g., Magentic One) and shares a unified runtime with Agent Service. This alignment enables consistent execution across local and cloud environments, with built-in observability, compliance, and durability for enterprise-grade agentic solutions.

- **Early external-facing agents:** While most Stage 3 deployments remain internal (for caution and testing), some organizations begin to pilot customer-facing agents in limited settings. For instance, a bank might test an AI agent in its mobile app to help customers navigate FAQs or a software vendor might beta-release an AI feature within its product. These agents need strong guardrails (to protect data and brand), but they represent the first forays into putting AI in front of end users. In recent interviews, several customers indicated these "on the horizon" plans that they started internally, then noted that, as confidence grew, "external scenarios and agents are on the horizon."[3]

One exemplary story of Stage 3 in action is that of a midmarket ISV. Initially, the company tried to code an AI solution itself with open-source models, but it fell short in accuracy and maintenance. By using Foundry's integrated approach in Stages 2 and 3, the ISV very quickly got an agent-driven solution working. In their own words, "They moved to Azure AI Foundry [now Microsoft Foundry], using GPT-4 for understanding emails, and used Foundry Agent Service to draft responses… [which they] connected to their CRM via Foundry IQ… [They then] built a functional AI email assistant in a few weeks, not months." In A/B tests, the AI agent handled 50% of incoming emails with minimal human intervention, cutting average response times from hours to minutes. This led the ISV CEO to remark that, with Microsoft Foundry, "we didn't have to become AI experts or build an MLops stack from scratch – it was all there. It just worked."[3] That quote underscores the value of Stage 3 integration: By leveraging Foundry's ready-made agent tools and Azure's platform, organizations can focus on the business logic of their AI agent rather than reinventing infrastructure.

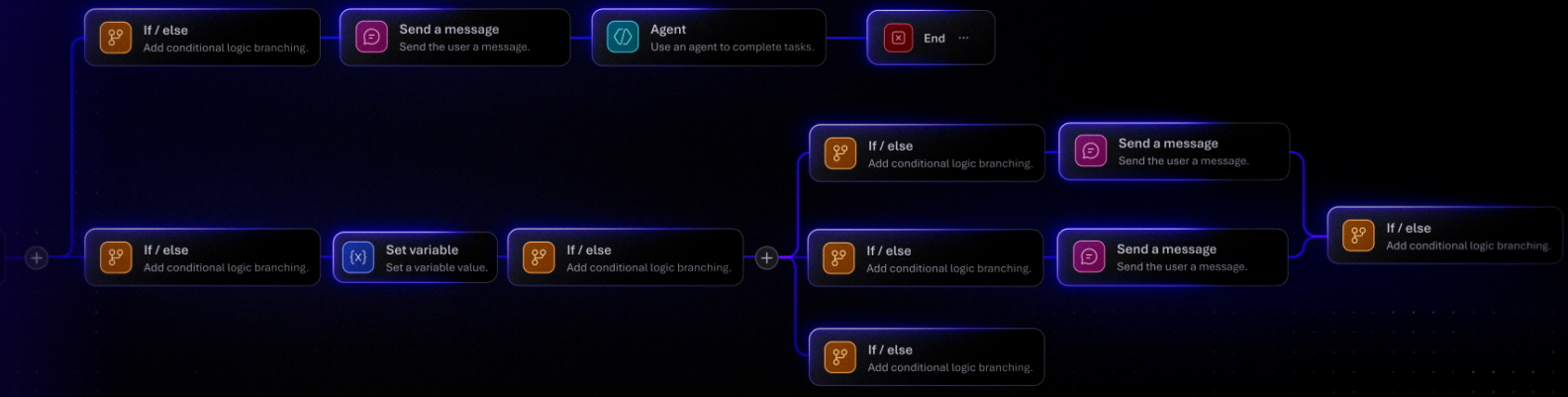
In one of our case studies, KPMG International built a custom AI agent within Microsoft Foundry to analyze regulatory and tax documents, surface relevant guidance, and help professionals deliver insights faster across global markets.[11]

Read more about KPMG International

"We use Foundry Agent Service to build and deploy agents. For one use case, we connected tools like Logic Apps and our own APIs so the agent could orchestrate an entire workflow. The first few hours were rough – onboarding had some sticking points – but templates helped. Now the agent can pull data from our ERP and draft a response automatically."

– A Stage 2 customer (developer) on building an integrated agent with Microsoft Foundry

This paraphrased insight reflects common feedback from organizations that the initial learning curve in Foundry Agent Service was quickly mitigated by templates, resulting in a productive agent that connects to enterprise systems.



STAGE 3

Challenges to overcome

Building agents that can reliably take actions introduces new complexities. Users in the research highlighted their key challenges, including understanding how to define tools, dealing with authentication and permissions for those tools, and learning how to debug multistep agent flows. One common pain point is tool-calling logic: figuring out how to prompt the agent to decide when to use a tool vs. answer directly, handling when a tool invocation fails, and so on. Some reported spending a lot of time refining system prompts and agent logic after deploying to pilot users, which "elongated our development cycle," but was necessary for a good final product. Observability becomes critical here to continuously monitor and govern the agent's behavior.

Strategic recommendations

- **Use templates and samples for agents:** Foundry Agent Service provides customizable agent code samples (for example, a pre-built helpdesk triage bot). These samples offer best practices, like how to call tools and manage the conversation flow. Development teams can start with a template to accelerate learning. As one user asked, "What's the one prescriptive guide you wish existed?" Today, the samples are intended to be that guide. This acceleration is echoed by customers: "We've seen a 30–40% improvement in time to market to build AI agents with Microsoft Azure AI Foundry [now Microsoft Foundry]. Before, we had to stitch together many models. Now we have one silo for development."[4] Another partner shared, "Our developers can go super-fast because they can get what they need in Microsoft Azure AI Foundry [now Microsoft Foundry]. They have reusable templates, centralized governance, and integrations, and Microsoft manages the infrastructure. We estimate that we reduce overall development time by 30–40%."[4]

- **Start internal, then expand externally:** Keep agents in internal pilot mode until you've resolved any issues. This was a prevalent strategy shared by Stage 3 customers in our interviews: nearly all had internal agents first to mitigate accuracy/security issues and gather feedback in a safe setting. Once the agent performs well and guardrails like content filters are put in place, test the agent with a small external user group or specific customer scenario. This phased approach builds trust with stakeholders (and regulators, if applicable) before a wide release.

- **Leverage orchestration features, but don't overcomplicate:** It's tempting to throw every tool at the agent, but a lean approach often works best initially. Identify the one or two most valuable actions an agent should perform, then implement those first. For instance, enable the CRM lookup and email send if those solve 80% of the use case, before adding more tools. Monitor agent performance and correct any identified errors. (For example, you would do this if the tool response needs formatting or the prompt needs adjusting.) Observability in Foundry Control Plane offers logs and traces to show each agent's decision for debugging.

- **Incorporate human oversight for critical tasks:** In Stage 3, AI is directly acting on business processes, so a human-in-the-loop mechanism is wise for high-stakes scenarios. For example, configure the agent to mark certain outputs as "draft" for a human to approve as the previous midmarket ISV company likely did for some email responses, or set thresholds where the agent hands off to a person. This approach captures efficiency gains while maintaining assurance.

- **Drive adoption of AI agents with workshops and hackathons:** Brainstorm agent ideas as a team or work with Microsoft to suggest co-creating a pilot agent code sample with other Microsoft Foundry customers. When Stage 3 is executed well, the organization achieves a kind of AI superpower. Not only can their system answer questions, it can also take actions automatically, bridging AI into operational workflows. This often yields significant ROI. For example, one midmarket ISV company's email agent halved response times and took on 50% of the workload, which translated to direct efficiency gains and better customer service.[1] Similarly, other Foundry customers reported that internal agents started saving employees hours per week on routine tasks.[1] These successes reinforce the business case for AI and typically lead to budget and approval for scaling up.

In our example case study, NTT DATA leveraged Microsoft Fabric data agents and Foundry Agent Service to orchestrate agents across multiple domains. These agents enable employees to interact conversationally with enterprise data, trigger insights, and automate decision workflows – all grounded in trusted corporate data. The result: over 50% faster time to market and heightened productivity across functions.[12]

Read more about NTT DATA

Now the development team stands on the threshold of Stage 4, where they will begin to scale solutions broadly and embed them into the enterprise's IT fabric with reliability, governance, and performance.

STAGE 4:

Deploying and scaling in production

Stage 4 is reached when an organization treats its AI solutions not as experiments or isolated apps, but as integral, production-grade systems. At this point, the goal is to scale up adoption across the organization and ensure that the AI is managed with the same rigor as any mission-critical application. Foundry customers at this stage typically utilize models, agents, and tools alongside Azure Machine Learning for custom model training, Azure App Services or Kubernetes for app deployment, and Foundry Control Plane for monitoring and enforcing enterprise-grade security and compliance.

Customers in Stage 4 often have multiple teams working on AI. They have multiple projects underway. For example, departments using models and knowledge for internal apps or building agents are now unifying and standardizing these efforts. AI projects blossom into full-scale applications with Azure service integrations.

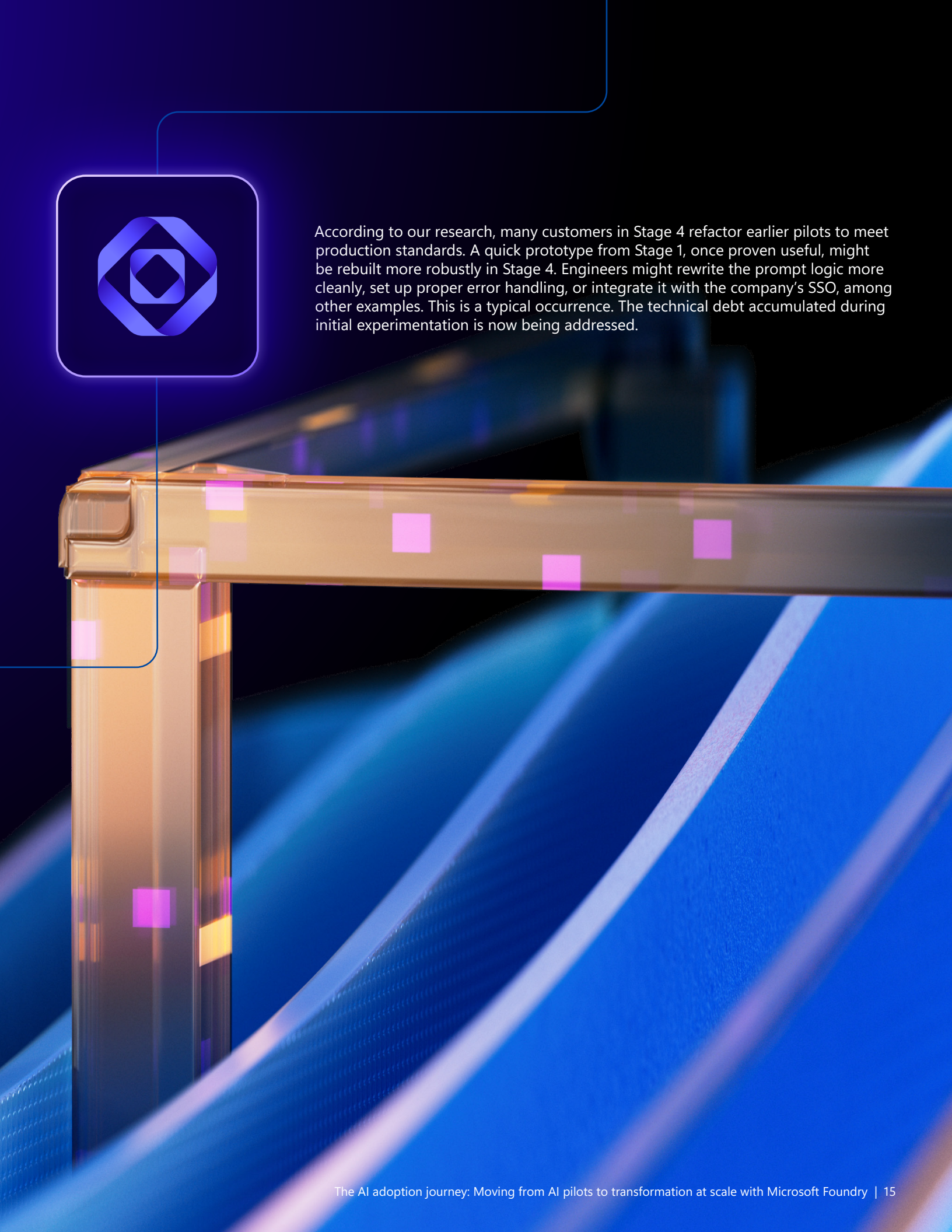
Common use cases to explore

- **Robust DevOps and MLOps for AI:** Our internal research showed 63% of high-maturity organizations implement observability metrics for AI.[13] That means they have dashboards to track model accuracy, latency, usage trends, and more. Foundry Control Plane provides this out-of-the-box, and Stage 4 users heavily rely on it for continuous monitoring. In addition, AI projects are moved into the standard IT pipeline. Teams use Azure DevOps/GitHub for version control and CI/CD of their AI code (prompts, agent configs, etc.). Many high-maturity orgs also implement evaluation frameworks and regular performance testing for their models.[3]

- **Scaling infrastructure and integrations:** As usage grows, so does the need for scalable infrastructure. Organizations in Stage 4 often deploy AI services across Azure Kubernetes Service (AKS) or use Azure Functions or Azure Container Apps for serverless scaling, depending on the scenario. In our customer research, 95% of AI deployments use Azure Kubernetes Service, which highlights how scaling companies choose robust infrastructure.[14] Additionally, they start to use more Azure services in concert. For example, Azure API Management can be used to expose AI endpoints securely.

- **Enterprise governance, security, and compliance:** Stage 4 organizations treat AI with the same governance as other enterprise apps. They apply comprehensive security, using private networking (such as deploying Microsoft Foundry components in a VNet so they're not accessible publicly), while also enforcing identity and access controls for who can publish changes to an agent or model (using Microsoft Entra ID roles) and logging all activities for audits. Many also integrate Content Safety in Foundry Control Plane to filter model inputs/outputs for policy compliance, which is especially important if the AI is customer-facing. Compliance teams are involved for AI use to meet industry regulations (finance, healthcare, etc.), and they appreciate features like data encryption, isolated compute, and traceability of AI decisions. Among genAI leaders, 90% have structured Responsible AI frameworks, 81% have oversight committees, and 80% have monitoring systems in production, indicating that ethics and governance are not just compliance requirements but drivers of trust and scalable AI adoption.[2]

- **Wider deployment and adoption:** At this stage, the AI solution is rolled out to a broad user base (could be thousands of employees or customers at scale). For instance, a pharmaceuticals enterprise ended Stage 4 with five departments using Foundry Agent Service and Azure App Services, after initially starting with just the R&D division.[3] Similarly, we see companies launching AI capabilities company-wide or adding an AI feature to a major product line for all customers. The success metrics here move from pilot metrics (like time saved for a small team) to enterprise KPIs.[3] Ultimately, Stage 4 is about embedding AI such that it drives tangible business outcomes at scale, like increasing revenue, reducing costs, accelerating innovation, and more.



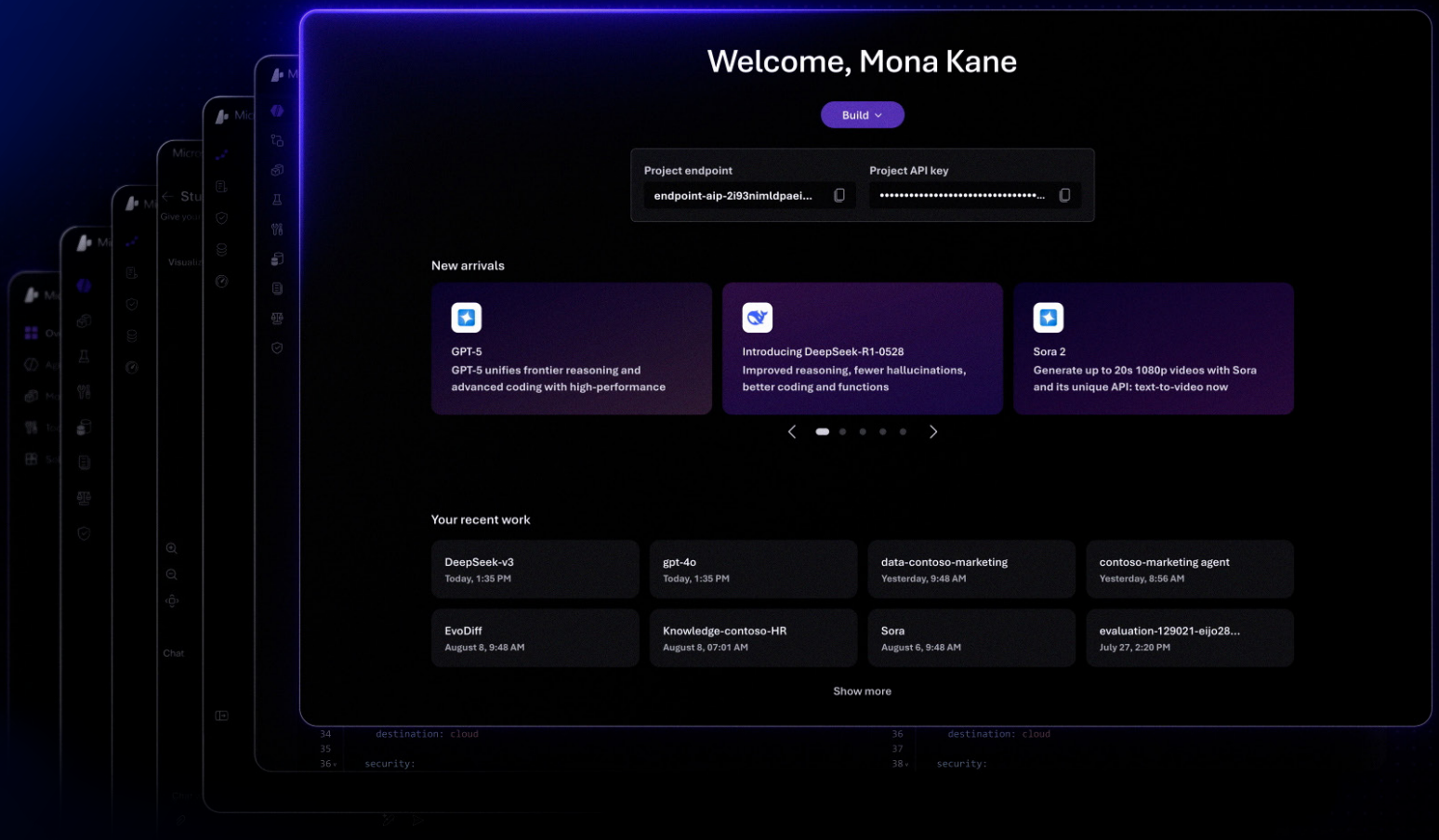
According to our research, many customers in Stage 4 refactor earlier pilots to meet production standards. A quick prototype from Stage 1, once proven useful, might be rebuilt more robustly in Stage 4. Engineers might rewrite the prompt logic more cleanly, set up proper error handling, or integrate it with the company's SSO, among other examples. This is a typical occurrence. The technical debt accumulated during initial experimentation is now being addressed.

STAGE 4

Challenges to overcome

The challenges here are the classic challenges of any IT scaling, with a twist of AI complexity:

- **Cost management:** With scale can come significant cost. AI inference (especially with large model calls) can be expensive at volume. Stage 4 teams monitor with Foundry Control Plane to optimize performance and costs. Some use auto-scaling to handle peak vs. off-peak and implement caching of frequent answers. Consider using cheaper model instances for less critical tasks. We recommend monitoring what drives costs (for example, one agent calling the model too often) and optimizing that in code. Customers highlight the cost advantages: "Microsoft Azure AI Foundry [now Microsoft Foundry] is very cost-efficient versus going to a GPU provider to run models from open-source libraries. It's 30–40% cheaper with Microsoft than if we tried to do it on our own."[10] Another partner observed, "Microsoft Azure AI Foundry [now Microsoft Foundry] makes our whole workflow and AI product development easier. We are still experimenting, but it's probably 25–30% time and cost savings."[10]
- **Keeping up with rapid changes:** AI tech evolves quickly. Stage 4 organizations may struggle to keep their solutions up-to-date – new model versions (GPT-5, GPT-4.1, etc.) or new model, tool, or agent features are deployed frequently. It can feel like aiming at a moving target. Microsoft's own updates (Ignite conferences, tech community posts) are key to watch at Stage 4.
- **User adoption and change management:** Rolling out AI to large user populations requires training and change management. Some employees might be skeptical or slow to trust the AI solutions. Stage 4 initiatives should include user education – show users how the AI works, clarify it's there to assist (not replace jobs, which is often a concern), and share success stories. Collect user feedback continually and improve the UX of your designed AI solutions. Essentially, ensure the AI is actually being used and delivering the expected value at scale, not just enabled.



Strategic recommendations

- **Implement full-stack monitoring and observability:** By now, it is crucial to have telemetry at every layer. Use Observability in Foundry Control Plane for model/agent telemetry (latencies, success rates, content filters triggered, etc.). Use Application Insights or Azure Monitor for the surrounding app (HTTP request rates, exceptions). Monitor infrastructure (CPU/GPU usage, memory) if on AKS or VMs. And set up alerts on key metrics (like a sudden drop in answer accuracy or a spike in error rate) so the team is notified to investigate. High-performing orgs treat this like a DevOps practice, and some even have an operational dashboard as part of their Network Operations Center, given how mission-critical the AI system becomes.
- **Optimize and iterate on models:** Teams should regularly evaluate if a model update or fine-tuning could improve performance. Perhaps a smaller, optimized model can handle 70% of queries, and only fall back to GPT-4 for the rest to save cost. Or perhaps new model versions (like GPT-5) dramatically improve results, as is the case for one automotive company that noted, "Issues we struggled with in GPT-4 were miraculously solved by GPT-5," justifying the upgrade.[3] Thus, staying on top of model developments and retraining when beneficial are key. Microsoft Foundry makes swapping or fine-tuning models automatic or with minimal coding. For example, development teams can A/B test with a model router to keep AI sharp and in support of business objectives.
- **Expand use cases and drive platform stickiness:** Stage 4 is not an endpoint – it's an ongoing expansion. Perhaps your success with an internal agent can be extended to customer-facing scenarios (with appropriate safeguards). A Microsoft survey indicated many Foundry customers remain unaware of other services beyond the ones they started with. So, if you began with Azure OpenAI and Foundry IQ (powered by Azure AI Search), consider trying Azure Speech for a voice interface or Azure Document Intelligence to process forms. Continuously evangelize internally how the Foundry ecosystem (and its integrations like Power Platform or Dynamics 365 connectors, if relevant) can be leveraged. This helps achieve the vision of an AI-frontier organization.
- **Maintain governance and oversight:** As AI becomes widespread, an AI governance committee or review process is critical. This group should review things like: Are AI outputs adhering to our responsible AI guidelines? Do we have processes if the AI makes a harmful mistake? Do we comply with emerging regulations (e.g., data residency, AI transparency requirements)? Foundry provides tooling like content moderation, tracing, and logs for audits. Consider implementing "watermarking" or disclaimers in user-facing responses to indicate AI-generated content, which is increasingly viewed as a best practice. Essentially, keep your AI deployment not just scalable and efficient, but also responsible and trustworthy at scale.

In Stage 4, the transformation intended at the outset is largely realized: The organization has gone from tentative early trials to running AI as a core part of its business. They have models, agents, and tools working in concert.

Reviewing the four stages of AI adoption

Let's review the four stages to AI adoption:

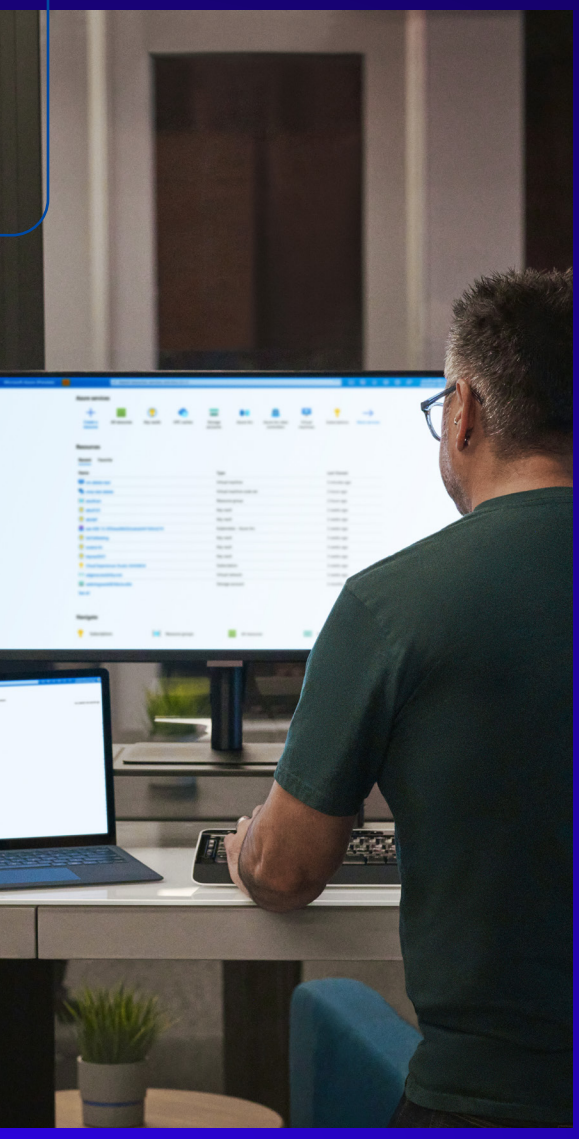
Stage 1 – Experimenting and early pilots: Start small with a targeted pilot (e.g., internal chatbot), using Foundry Models and Foundry Tools. Establish value and get initial buy-in.

Stage 2 – Grounding with enterprise data and customizing with Azure Machine Learning: This boosts relevance and user trust, turning proofs of concepts into truly useful solutions.

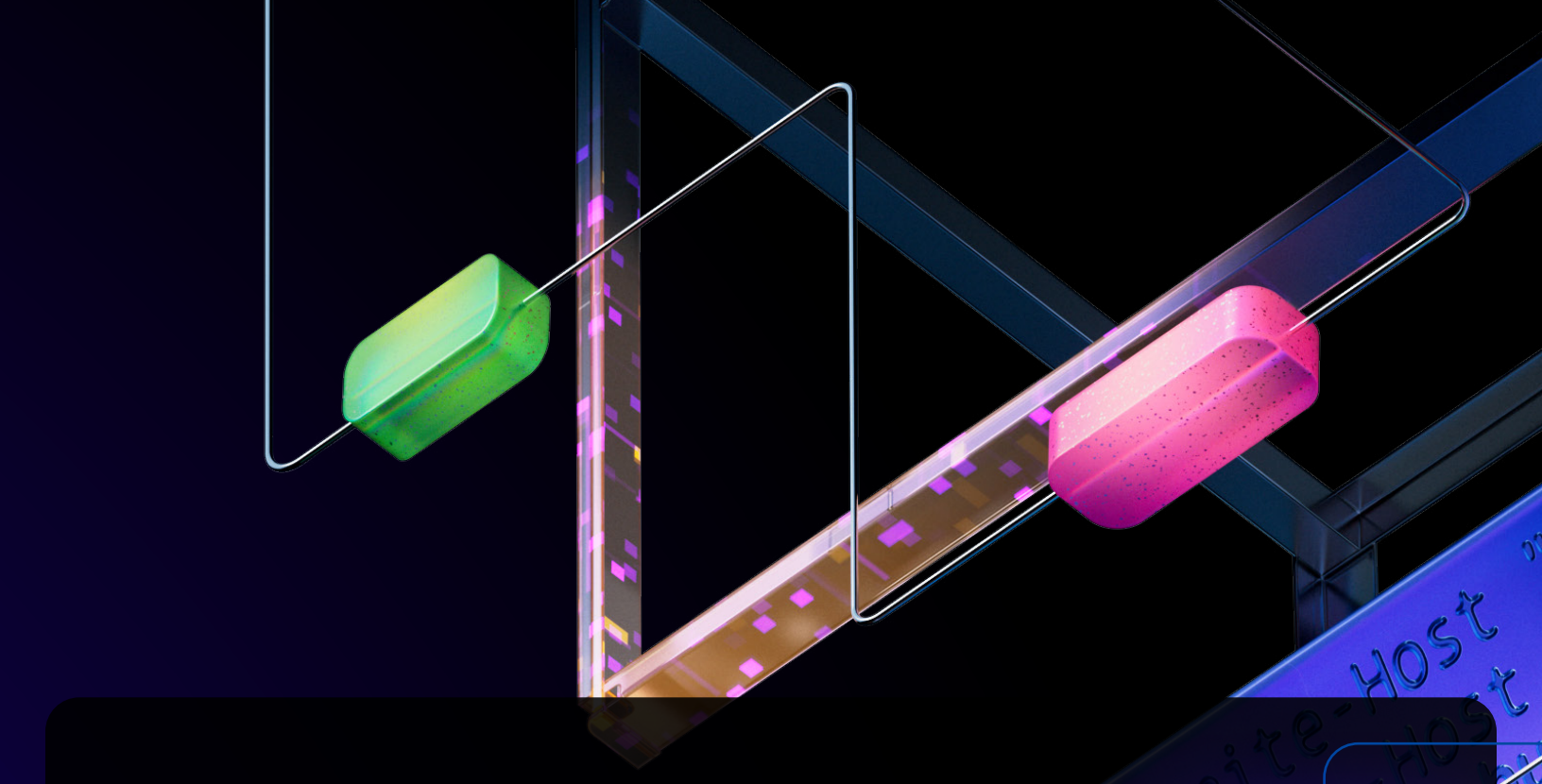
Stage 3 – Building intelligent agents and workflows: Integrate tools/APIs so the AI can perform tasks (not just chat), automating workflows.

Stage 4 – Deploying and scaling in production: Attach cloud services and implement monitoring and governance so AI can become an enterprise-grade solution.

Each stage builds on the previous, and thanks to the unified Foundry portal, SDK, and API, this progression can be achieved incrementally. Foundry integrates all the necessary pieces – models, agents, tools, and control plane – so that advancing through the stages is more a matter of turning on the next module than starting from scratch at each step.



- **Champion a cross-functional AI team:** The most successful cases of customers had a blend of stakeholders – IT, developers, business SMEs, and executive sponsors – collaborating from pilot to scale. This provides technical robustness, business relevance, and leadership support at each stage.
- **Measure and communicate value at every stage:** Establish KPIs early (even if rough), such as time saved, accuracy improved, or user satisfaction, and track them. Share quick wins from Stages 1 and 2, and later, substantial ROI from Stages 3 and 4. For example, 20 hours/week saved or 50% automation of emails are powerful stats to secure ongoing investment.
- **Use a crawl-walk-run approach for new features:** When introducing something new – be it a tool or exposing AI to customers – start with a limited scope ("crawl"), learn and iterate ("walk"), then scale up ("run"). This de-risks innovation. The modular nature of Foundry enables teams to pilot a feature with one team, then gradually roll it out org-wide once validated.
- **Continuously upskill your teams:** AI tech is fast-moving. Invest in training developers on new Foundry capabilities, encourage architects to attend Foundry workshops, and create internal knowledge shares. Some companies create "AI Champions" in each department to facilitate adoption. Microsoft Learn modules for Microsoft Foundry and the Azure Tech Community can be useful resources here.
- **Stay engaged with Microsoft's roadmap:** Microsoft Foundry features are constantly evolving. Being an early adopter can provide a competitive advantage. Joining preview programs or customer advisory boards can give your organization a head start on the latest models, agents, and tools. Microsoft Azure is dedicated to supporting organizations at every stage of their cloud and AI journey. Whether teams are exploring new possibilities, planning strategies, implementing solutions, scaling across the enterprise, or realizing measurable outcomes, Azure provides tailored programs, expert guidance, and best practices. Azure meets customers wherever they are, combining resources, funding opportunities, and technical support to accelerate adoption and maximize value. This approach empowers organizations to innovate confidently and achieve frontier outcomes with the right strategy and support.



We're here to help at every stage of the journey

Exploring
You're learning cloud and AI, experimenting in select areas of your organization

Planning
You're defining cloud and AI strategy, running proofs of concept and planning deployments

Implementing
You're moving from proofs of concept and pilots into production

Scaling
You've deployed cloud and AI solutions for multiple functions and are scaling across your organization

Realizing
You're producing measurable cloud and AI value through strong processes and governance

Microsoft Unified
24/7 access to experts and hands-on support

Microsoft Learn
A free, centralized learning platform with interactive tutorials, certification tracks, and hands-on labs

Microsoft Marketplace
One storefront for pre-built cloud solutions and services

Azure Essentials
Curated technical guidance and best practices for cloud and AI adoption

Azure Accelerate
Access to funding, expert guidance, and 700+ specialized partners

Conclusion

In conclusion, the journey from initial AI trials to platform-scale deployment can be achieved – and the common customer progression observed (models → knowledge → agents → integrated platform) provides a proven roadmap. By following the stages outlined above and heeding the lessons from others who have gone before, Azure customers and technical decision-makers can accelerate their AI adoption with confidence. Microsoft Foundry is there to support each leg of the journey, providing the "assembly line" of AI capabilities needed to go from a single brilliant idea to AI at enterprise scale.

We stand at an inflection point. Businesses have seen flashes of AI's potential – a chatbot here, a proof-of-concept there – but struggle to translate that into real, deployed solutions at scale. Microsoft Foundry's unified platform is built to solve exactly this, helping organizations go from pilot to production faster using whatever mix of models and tools fit their needs.

By encouraging model experimentation, grounding with knowledge, integrating agentic capabilities, and embracing the broader Azure ecosystem to support AI app and agent deployment, enterprises can turn pilots into transformative solutions. The journey is iterative and sometimes non-linear, but each step builds momentum. With the right strategy and the capabilities of Microsoft Foundry, AI adoption becomes a continuous cycle of value creation – guiding organizations to not only adopt AI, but to truly embed it into the fabric of businesses.

Build with Microsoft Foundry

- Get started with Microsoft 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
- Watch our Foundry demo on YouTube
- Visit the Path to Production for AI Agents

[1] Generative AI at Scale: Delivering Benefits That Include Increased Revenue, Cost Savings, and Augmented Innovation, CIO Whitepaper
[2] AI Readiness Advisor, Whitepaper
[3] Internal Microsoft Report
[4] Microsoft Foundry Customers IDI Report
[5] Forrester TEI Report of Microsoft Foundry
[6] Assembly Software saves law firms up to 25 hours per case with Azure, Microsoft Customer Stories
[7] Insight Partner 2024 Survey of enterprise tech leaders
[8] Cognizant makes performance management more effective and meaningful with Microsoft Azure Machine Learning, Microsoft Customer Stories
[9] YoungWilliams cuts call center response time 99% with Azure AI Foundry Agent Service, Microsoft Customer Stories
[10] Azure AI Foundry: The AI App and Agent Factory, Microsoft Azure Blog
[11] KPMG is redefining the audit with agentic AI using Azure, Microsoft Customer Stories
[12] NTT Data Transforms Its Enterprise with Agentic AI in Microsoft Fabric, Microsoft Foundry, Microsoft Customer Stories
[13] Gartner Survey Finds 45% of Organizations With High AI Maturity Keep AI Projects Operational for at Least Three Years
[14] Azure Telemetry Data, Microsoft

Microsoft Azure