MIT Technology Review Insights

Produced in partnership with

Microsoft Azure

From the field: Top best practices and techniques

Customising generative AI for unique value

Preface

‘Customising generative AI for unique value’ is an MIT Technology Review Insights report sponsored by Microsoft Azure. This report seeks to understand how technology leaders are customising generative AI in their businesses and to what extent this is a priority for their enterprise-wide AI strategy. Denis McCauley was the author of the report, Laurel Ruma was the editor and Nicola Crepaldi was the producer. The research is editorially independent, and the views expressed are those of MIT Technology Review Insights.

We would like to thank the following executives for their time and insights:

Mark Austin, Vice President, Data Science, AT&T

Eric Boyd, Corporate Vice President, AI Platform, Microsoft

Tanwir Danish, Global Solutions and AI Officer, Data and Technology, Dentsu

Brian Demitros, Innovation Lead, Data and Technology, Dentsu

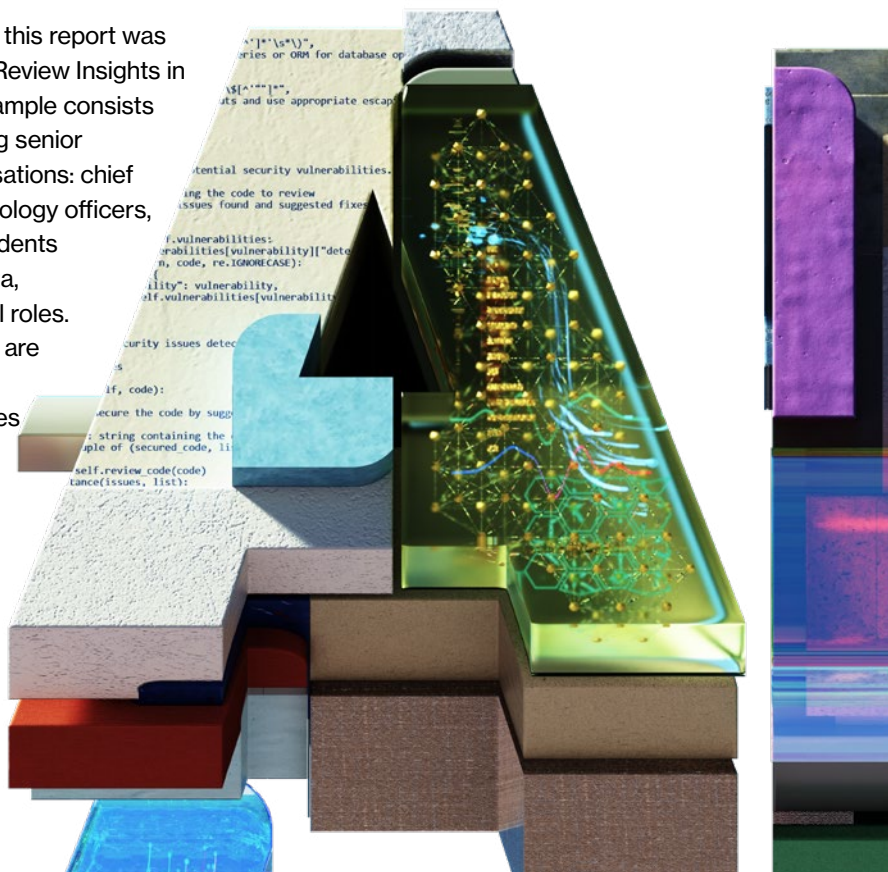
Ece Kamar, Managing Director of AI Frontiers, Microsoft Research

Gabe Pereyra, President and Co-Founder, Harvey AI

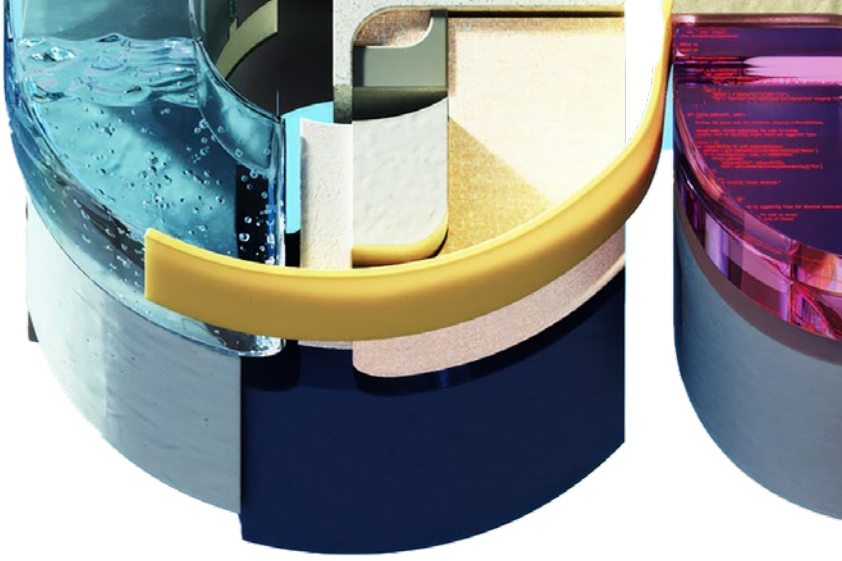
Asha Sharma, Corporate Vice President and Head of Product, AI Platform, Microsoft

Methodology

The survey forming the basis of this report was conducted by MIT Technology Review Insights in September 2024. The survey sample consists of 300 global executives holding senior technology roles in their organisations: chief information officers, chief technology officers, chief data/AI officers, vice presidents and directors of technology, data, engineering and other influential roles. The respondents’ organisations are primarily large enterprises and are headquartered in 12 countries in the Americas, Europe and Asia-Pacific. Twelve industries are represented in the survey, with the largest contingents of respondents working in financial services, technology, consumer goods and retail and manufacturing businesses.



CONTENTS



01 Executive summary

Since the emergence of enterprise-grade generative AI, organisations have tapped into the rich capabilities of foundational models, developed by the likes of OpenAI, Google DeepMind, Mistral and others. Over time, however, businesses often found these models limiting since they were trained on vast troves of public data. Enter customisation – the practice of adapting large language models (LLMs) to better suit a business's specific needs by incorporating its own data and expertise, teaching a model new skills or tasks or optimising prompts and data retrieval.

Customisation is not new, but the early tools were fairly rudimentary, and technology and development teams were often unsure how to do it. That's changing, and the customisation methods and tools available today are giving businesses greater opportunities to create unique value from their AI models.

We surveyed 300 technology leaders in mostly large organisations in different industries to learn how they are seeking to leverage these opportunities. We also spoke in-depth with a handful of such leaders. They are all customising generative AI models and applications, and they shared with us their motivations for doing so, the methods and tools they're using, the difficulties they're encountering and the actions they're taking to surmount them.

Our analysis finds that companies are moving ahead ambitiously with customisation. They are cognisant of its risks, particularly those revolving around data security, but are employing advanced methods and tools, such as retrieval-augmented generation (RAG), to realise their desired customisation gains.

The study's key findings include:

• **Customisation brings more than efficiency.** Boosting efficiency is a key motivation for customising generative AI models, according to 50% of surveyed executives, but it's not the only one. As important, say 49%, is gaining the ability to create unique solutions, 47% cite better user satisfaction and 42% seek greater innovation and creativity.

• **RAG provides the backbone for generative AI performance.** Two-thirds of companies are using or exploring RAG as a method of customisation. Over half (54%) are also employing fine-tuning techniques, indicating that these two methods, along with prompt engineering, are used most effectively in combination.

• **Automated evaluation is gaining traction.** Over half (54%) of surveyed businesses employ manual methods to evaluate generative AI models. But 26% are either beginning to apply automated methods or are doing so now consistently.

• **Data integrity is the biggest barrier to customisation.** Around half the respondents (52%) cite the need to ensure data privacy and security as the primary difficulty they face with customisation. Most (86%) say focusing on privacy and security has become more important as they customise more actively. One-third overall (32%), and 57% of the biggest companies in the survey, deem this 'much more important'.

• **Advanced tools are empowering developers and facilitating lifecycle management.** Over half (53%) of organisations have adopted telemetry tools for tracing and debugging for their developers. Also widely used are a simplified playground of tools (by 51%) and prompt development and management (46%) to facilitate better collaboration between engineers.

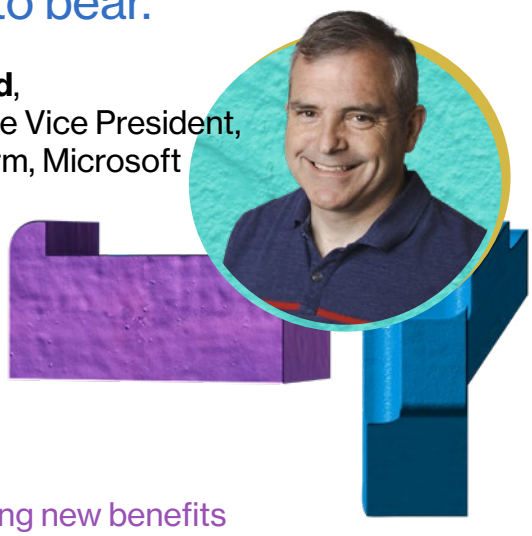
02 Embracing AI customisation

Businesses customise for a range of reasons, but tying them all together is the desire to integrate the organisation's own knowledge and expertise into the solutions they wish to build. Such integration is what enables businesses to build AI applications that are differentiated from others in the market. "If you're using an off-the-shelf model, your application's going to look very similar to everyone else's," says Eric Boyd, corporate vice president, AI platform, at Microsoft. "What is it that differentiates your application? It's usually the data and knowledge that your organisation can bring to bear."

Foundational models are not suitable for enterprise use, says Brian Demitros, innovation lead for data and technology at advertising network Dentsu. "They're trained on the internet and often contain innacurate or misleading information," he says. "You can't simply use them out of the box and expect a level of accuracy needed to support critical decision-making. Customisation is critical to get value out of them."

"What is it that differentiates your application? It's usually the data and knowledge that your organisation can bring to bear."

Eric Boyd,
Corporate Vice President,
AI Platform, Microsoft

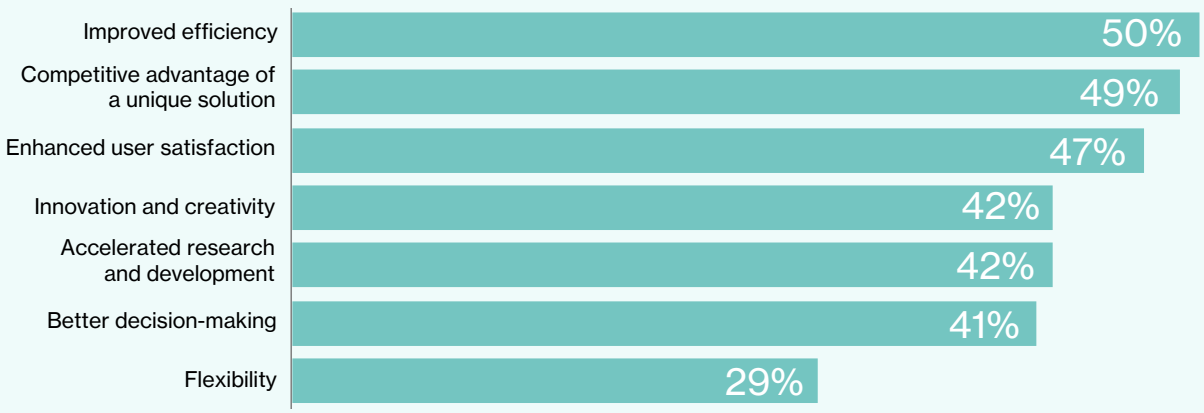


Harnessing new benefits

The executives we surveyed see three benefits above all from customisation. Half of them (50%) say it's important for delivering greater efficiency – for example, by automating tasks, streamlining workflows and optimising business processes. A similar share (49%) also say customising is important to gain competitive advantage from having a unique solution in the market. Almost as many (47%) cite enhanced user satisfaction – for example, from personalisation or greater responsiveness (see Figure 1).

Figure 1: Customising brings efficiency and more

Competitive advantage, enhanced user satisfaction, innovation and creativity and accelerated R&D are also top motivators for AI model customisation.



Source: MIT Technology Review Insights survey, 2025

Leveraging AI to uncover white spaces and drive brand growth is a key benefit for marketing and advertising agency Dentsu, according to Tanwir Danish, global solutions and AI officer, data and technology for Dentsu.

"We have proprietary methodologies for driving growth, designing and activating the total experience and optimising marketing campaign performance," he explains. "These are specialised fields, where our expertise, knowledge bases, proprietary data sources and machine learning models empower us to build enterprise-grade AI systems that deliver growth."

Gabe Pereyra, president and co-founder of Harvey AI, an AI solutions provider serving the legal industry, also cites enhanced creativity and quality as customisation benefits. "Many partners at our law firm clients use custom models to ask questions, get information, draft briefs and develop arguments that others may not have."

Using customised solutions makes these lawyers better at their jobs. Among the surveyed companies, 42% also cite innovation and creativity – such as the creation of new products, services and business models – as a key desired outcome from customisation. The same percentage cite accelerated R&D as a key benefit (see Figure 1).

"You can't simply use them out of the box and expect a level of accuracy needed to support critical decision-making. Customisation is critical to get value out of them."

Brian Demitros, Innovation Lead, Data and Technology, Dentsu

Tomorrow's large and customised industry-specific models

Large language models may not lend themselves to specialised, industry-specific purposes today, but Gabe Pereyra foresees the future development of powerful, all-purpose and self-learning AI models that can play that role. The company he co-founded, Harvey AI, is seeking to build one for the legal industry.

"In a couple of years, it will no longer make sense for a law firm to have one system that extracts data from contracts, another that analyses transcripts and another that reviews the e-Discovery corpus," according to Pereyra. "Instead, the firm's professionals should be able to access a super powerful model that's connected to all this data and can perform tasks across a wide range of thematic areas."

This will not be the same as interacting with today's foundational model, says Pereyra. "It will be customised in two ways. First, it will have access to all the specialised data," he says. "Then it will learn from all the interactions of the lawyers using the model." Therein will lie the model's real value, says Pereyra: "It will imitate all the firm's internal expertise."

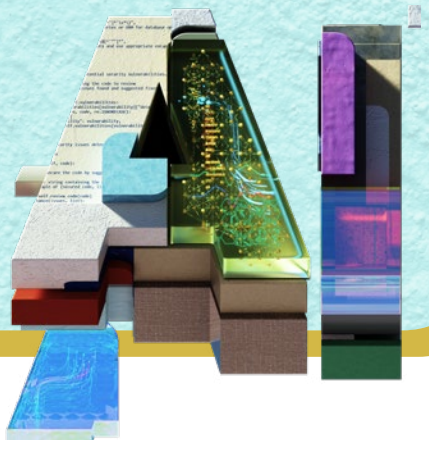
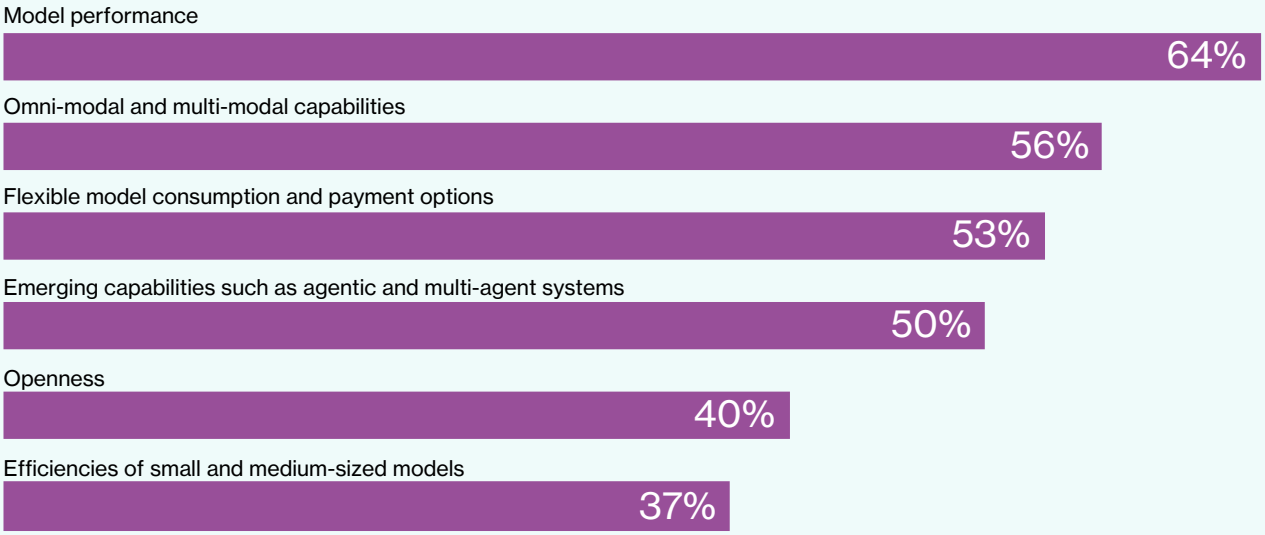


Figure 2: Key model attributes when selecting are performance, multi-modal capabilities and flexibility

Emerging capabilities like agentic and multi-agent systems are already of high importance for enterprises.



Source: MIT Technology Review Insights survey, 2025

Choosing and evaluating models

Achieving these outcomes starts with getting model selection right in the first place. What attributes do technology teams want in their generative AI models? Most of the survey respondents (64%) prioritise performance: the model must provide information that is relevant, accurate and coherent to any who query it. Omni-modal and multi-modal capabilities are also a priority for 56% of respondents. “We want end-users to be able to interact with models in different ways,” says Mark Austin, vice president of data science at AT&T. “Chat is just one interface. We’re also experimenting with voice – making sure the latency is good – as well as human-looking avatars.”

Just over half (53%) of the surveyed executives place a priority on flexible model consumption and payment options: for example, the ability to reserve capacity or access hosted fine-tuning. And 50% already want their models to have agentic capabilities (see Figure 2).

Agentic systems act as autonomous agents, performing tasks and making decisions without the need for direct human intervention. Pereyra says Harvey AI is starting to do more agentic workflows. “These are systems that can take complex legal tasks, decompose them, solve the subtasks, put them together and produce associated or higher-level product,” he says.

> “We want end-users to be able to interact with models in different ways. Chat is one interface; we’re also experimenting with voice and human-looking avatars.”
>
> **Mark Austin**, Vice President, Data Science, AT&T

"Model evaluation should be a critical application feature. Rather than evaluating once and moving on when models change, we need to move to continuous evaluation."

Asha Sharma, Corporate Vice President and Head of Product, AI Platform, Microsoft

Having agreed on desired model attributes, technology teams then need to evaluate the model options available to them. Manual methods of evaluation predominate among the surveyed companies. But a quarter of respondents (26%) say they are using automated methods to a greater or lesser extent. Around one-fifth (17%) are starting to employ automated evaluation with large data sets and almost 9% of respondents are consistently automating, showing how quickly and advanced this cohort has become. (see Figure 3).

Customising operations at AT&T

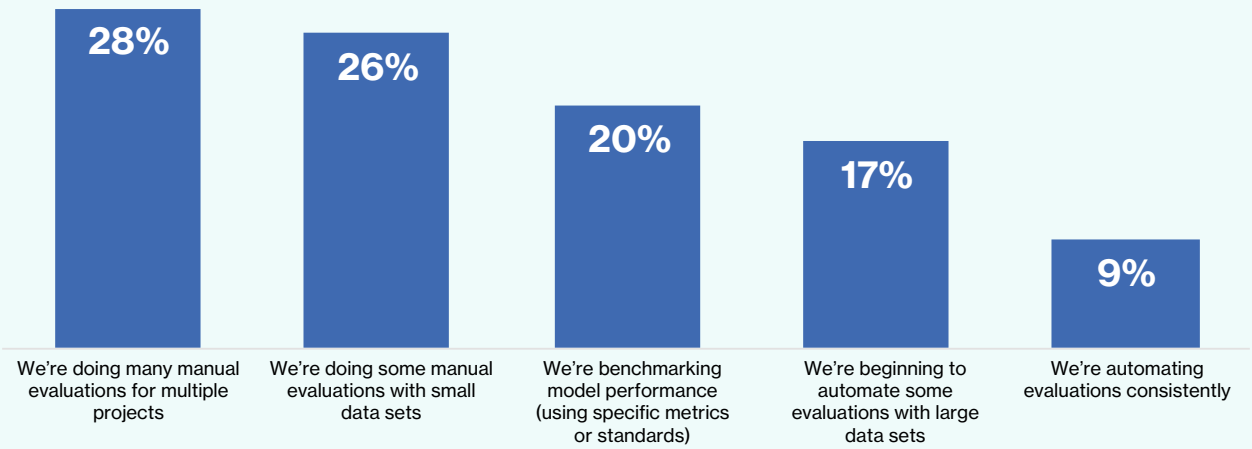
AT&T is adapting a variety of different generative AI use cases to enhance operational efficiency[1] across the business. The telecom giant is using Microsoft Azure OpenAI Service to help it migrate legacy code into modern code to accelerate developer productivity, to allow IT professionals to request additional resources like virtual machines and to enable employees to complete common human resources tasks by asking ChatGPT a question or giving it a command. The task is then passed on to the appropriate person on the employee's behalf.

It's all part of AT&T's plan to streamline rote or repetitive tasks to enable employees to focus on more complex, higher value jobs on its mission to provide better connectivity, service and value to its customers. "Using Azure OpenAI Service to help automate some of these more common tasks will be an important change to the way we operate. There will be meaningful time and cost savings," says Jeremy Legg, chief technology officer at AT&T.

AT&T deploys a number of Microsoft technologies, including Azure OpenAI Service, Azure AI Search, Azure Databricks and Azure Compute.

Figure 3: Automated evaluation gains traction

Most organisations are still relying on manual evaluation for assessing generative AI model choices, but the use of automation is picking up.



Source: MIT Technology Review Insights survey, 2025

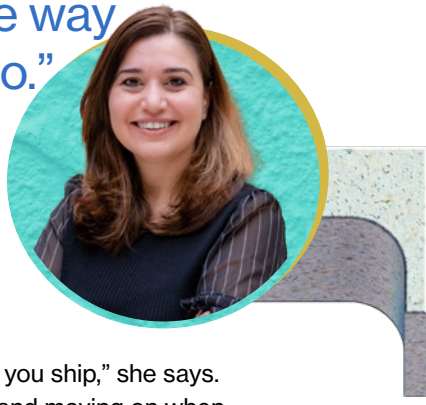
"How companies evaluate partly depends on their AI maturity," says Boyd. "If they're just getting started with generative AI, evaluation tends to be manual. As maturity grows, and they start to evaluate different data sets or different prompts, we're seeing a lot more use of automated evaluation."

According to Pereyra, automated methods currently have limitations when evaluating specific use cases. "Automated evaluation gives you a rough directional sense," he says. "[The benchmarks it provides] separate models into different generations, but they don't help you discriminate much beyond that." The optimal approach to evaluating models, says Pereyra, is to employ different evaluation methods – manual, automated and benchmarking – in tandem.

In addition to this, how teams mentally approach evaluation needs to change, says Asha Sharma, corporate vice president and head of product, AI platform, at Microsoft. "Model evaluation should be a critical application feature that you experiment with in production, just like you do with the other most important pieces of code that you ship," she says. "Rather than evaluating once and moving on when models change, we all need to move to continuous evaluation."

"At the end of the day, this is all about people. We are creating these systems to provide real value in the things we care about and doing that in the way we want them to."

Ece Kamar, Managing Director of AI Frontiers, Microsoft Research



Agents of change

If generative AI has been a game-changer for AI as a whole, agentic systems could do the same for generative AI. Taking generative AI's unique capability of creating content in the form of responses to queries further, agents perform actions based on the information the model has gathered.

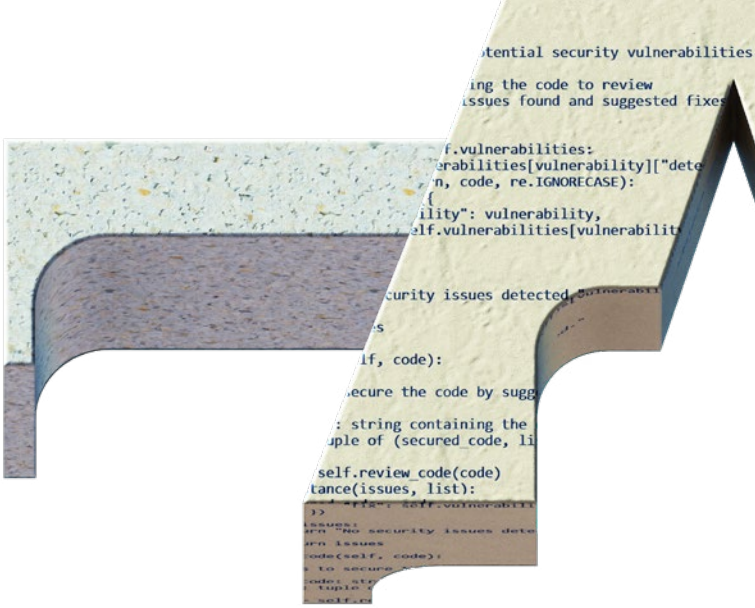
"To be really useful, AI systems need to act, perceive the result of their action and then act again," says Ece Kamar, vice president of research and managing director of Microsoft's AI Frontiers Lab. "I don't want a system that just tells me available flights, for example. I want a system that goes and books the flight for me."

Single-agent AI systems, in which an intelligent entity (a bot, for example) acts alone to perform a specific task, are already in commercial use today. Now, multi-agent systems – where multiple entities interact collaboratively (sometimes as checks and balances) to complete complex tasks – are coming to the fore.

Multi-agent systems lend themselves to autonomous problem solving in areas such as supply chain management, manufacturing operations, transport and logistics and securities trading, to name a few. The list also includes AI model customisation. Kamar foresees them being used to overcome one of the biggest difficulties companies encounter with customisation – data availability and quality.

"Multi-agent technologies will be able to create high-quality and diverse synthetic data for many domains," predicts Kamar. She provides an example: "You may lack data for a customer service scenario you have in mind. With a multi-agent setting, one agent can simulate the customer support person and another can simulate the customer, asking all kinds of questions. Another agent will monitor the conversation and make sure every piece of information provided by the customer support agent is grounded in facts, with the help of RAG."

In developing these emerging systems, it is critical to prioritise the user experience throughout the process, says Kamar. "At the end of the day, this is all about people. We are creating these systems to provide real value in the things we care about and doing that in the way we want them to."



03 The quality and performance imperative

Realising the potential of generative AI lies in its continuous improvement. As models and applications ingest more data, handle more queries and learn, their outputs become more accurate and relevant. Therefore, the importance of these models and applications is increasingly tied to business outcomes that organisations seek in terms of efficiency, competitive differentiation, user satisfaction, innovation and other areas.

Implementing a variety of methods

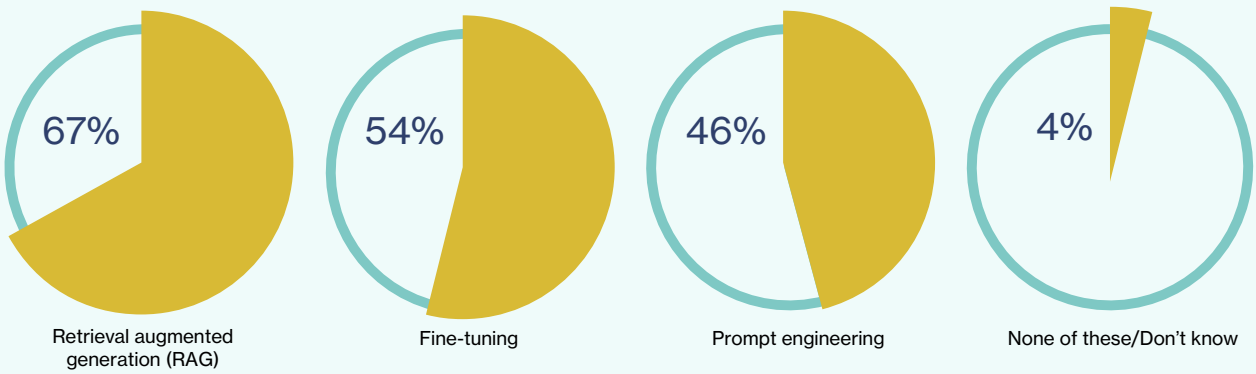
We asked technology executives their preferred methods of customising generative AI models. Their responses make clear that their organisations employ not one, but a trio of methods. Two-thirds (67%) are implementing RAG or exploring its use. Over half (54%) also employ (or are exploring) model fine-tuning for this purpose. Prompt engineering, cited by almost 46%, rounds out the array of methods employed (see Figure 4).

Each method performs a distinct role. RAG scours and retrieves data from external and internal sources to ensure that model outputs are relevant and based on the most up-to-date information available. Fine-tuning is a necessary complement, ensuring that the model is retrieving the internal data it needs to perform the specific tasks set for it. Prompt engineering performs a third role, that of guiding the design of instructions, or prompts, that users give to a model to obtain the desired information.

"In effect, RAG gives AI the company's memory," says Sharma. "While general purpose models are powerful, they're missing context like product, policies and the ways the organisation does business." She observes companies benefiting from RAG today in areas such as: customer support, providing real-time access to up-to-date product documentation, for example; employee productivity, by turning thousands

Figure 4: RAG is core to customisation, in combination with other methods

Fine-tuning and prompt engineering complement RAG as preferred methods of model customisation.



Source: MIT Technology Review Insights survey, 2025

"With the help of AI agents and a customised RAG framework that taps into enterprise-grade proprietary AI models and data repositories, we can now support some of the most impactful decisions our brands make to drive growth."

Ece Kamar, Managing Director of AI Frontiers, Microsoft Research

of scattered documents into an instantly accessible corpus of expertise and skills; and compliance and risk management, as RAG lets companies control exactly what information an AI model can access and reference.

When it comes to customising generative AI applications, RAG is the simplest way to get going, according to Boyd. "It doesn't take long to wire a search engine up to a model on top of an organisation's internal data," he says. "Once that happens, a generative AI application can be created very quickly. That's why there's so much enthusiasm for it."

From there, RAG use usually requires other methods to get optimal results. With his firm's work in the legal industry, Pereyra finds RAG intuitive to use for basic searches, but more difficult as the queries become complex. He finds that fine-tuning works well with models designed to perform specific tasks.

Demitros from Dentsu agrees, "RAG is best used in combination with other methods. We're also doing fine-tuning, prompt engineering and exploring innovative approaches. And for a lot of what we're doing, there's a heavy software layer on top, which includes security and privacy as well as decision-making support."

Customising for accuracy with RAG

When devising an advertising strategy for a client, a vital piece of analysis is finding out what contributions different media channels make to the client's sales. Dentsu found that using general purpose LLMs made retrieving this information simpler and faster than before. However, says Brian Demitros, its accuracy left a lot to be desired. "We were getting 40 to 50% accuracy in the answers," he says. "That's obviously not acceptable, so we had to do a lot of custom development."

Working on a campaign for a retail client, Dentsu created its own guardrails, embeddings and vector stores, to harness its institutional expertise in data analysis for the retail and marketing domains, according to Tanwir Danish. "Our models now accurately answer to retailer-specific questions on marketing performance and budget allocation." He continues, "With the help of AI agents and a customised RAG framework that taps into enterprise-grade proprietary AI models and data repositories, we can now support some of the most impactful decisions our brands make to drive growth."

To further enhance decision-making, Dentsu integrated an agentic decision layer, enabling AI-driven recommendations for optimising marketing budget allocation. "We can do this because we have a library of AI and machine learning models that identify key drivers of marketing performance. Our optimisation models simulate business outcomes, such as sales, under different scenarios," Danish adds. This AI-powered approach is now central to how Dentsu leverages generative AI to shape campaign strategies for clients. "We've achieved around 95% accuracy in retrieving the most relevant data and insights – an enormous improvement," says Demitros. "Using generative AI without a customised layer is simply not viable when supporting business decisions for some of the world's largest brands."

Customising efficiency for the legal world at Harvey AI

Harvey AI – a legal artificial intelligence platform designed specifically for lawyers and law firms – is customising generative AI for use cases[2] including summarising and comparing documents, referencing case law and facilitate research and analysis.

Deployed across hundreds of law firms and legal teams, Harvey's platform helps lawyers and professional services providers deliver complex legal results more efficiently. "The reason it's been so hard to build technology for industries like legal is the workflows are so varied and complex and no two days are the same," explains Gabe Pereyra, Harvey's president and co-founder.

"We're now looking at leveraging NVIDIA-accelerated computing on Azure to train our own open-source models," he says. The company can also mix managed compute service and dedicated capacity for both its language models and embeddings in every region where it operates, helping it more efficiently scale throughout for research and product development. "With our need to collocate compute and models, Azure makes that easy," explains Pereyra.

Harvey AI uses a mix of products from Microsoft, including Azure OpenAI Service, Azure Database for PostgreSQL, Azure Blob Storage and Azure High Performance Computing (HPC).

The devil's in the data

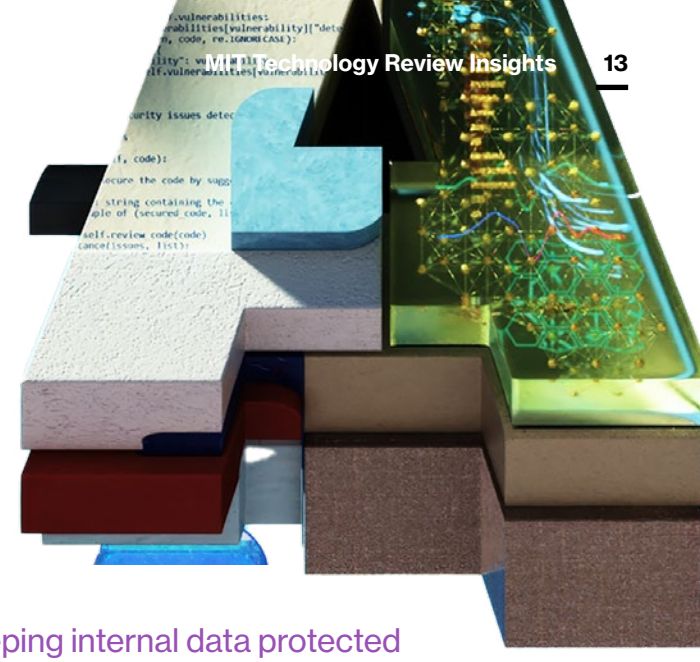
The chief barriers to customisation cited in the survey revolve, not surprisingly, around data integrity – a term that takes in its accuracy, its relevance and its safeguarding. Just over half the respondents (52%) say their main difficulties lie in ensuring data privacy and security. Almost as many (49%) cite data quality and preparation. And 45% report that they don't have the ability to measure the impact of customisation on a model's output and performance (see Figure 5).

Conveying the value of improvements to AI is a challenge that's tied to AI's growing pervasiveness, says Sharma. "When AI is embedded throughout your processes, isolating the impact is like trying to measure the ROI for electricity." But organisations are starting to flip the measurement question, she says. "Instead of asking what AI accomplished, they're asking what AI-enabled humans can accomplish. For example, a health care provider we work with doesn't just track the number of radiology images their model analyses; it also tracks the additional patient consultations that radiologists have been able to take on."

Figure 5: Data integrity is the biggest challenge to model customisation

The ability to measure ROI and finding the right developer talent and skills are also some of the biggest barriers.

Source: MIT Technology Review Insights survey, 2025



04 Risk factors

Generative AI adoption has brought a greater focus from organisations on the privacy, safety and security of customised models and their data. Executive concerns about threats to these come through clearly throughout the global survey, but the experts we interviewed believe the concerns will begin to recede.

The vast majority of surveyed executives (86%) say the advent of generative AI, and the capability to customise it, has heightened application safety. Around one-third (32%) say it has become "much more important." The bigger the organisation, the greater the degree of concern with the potential for breaches. Among respondents from the biggest companies (with at least USD 50 billion in annual revenue), 57% deem safety to be much more important now, compared with just 14% of those working in the smallest firms in the survey (see Figure 6).

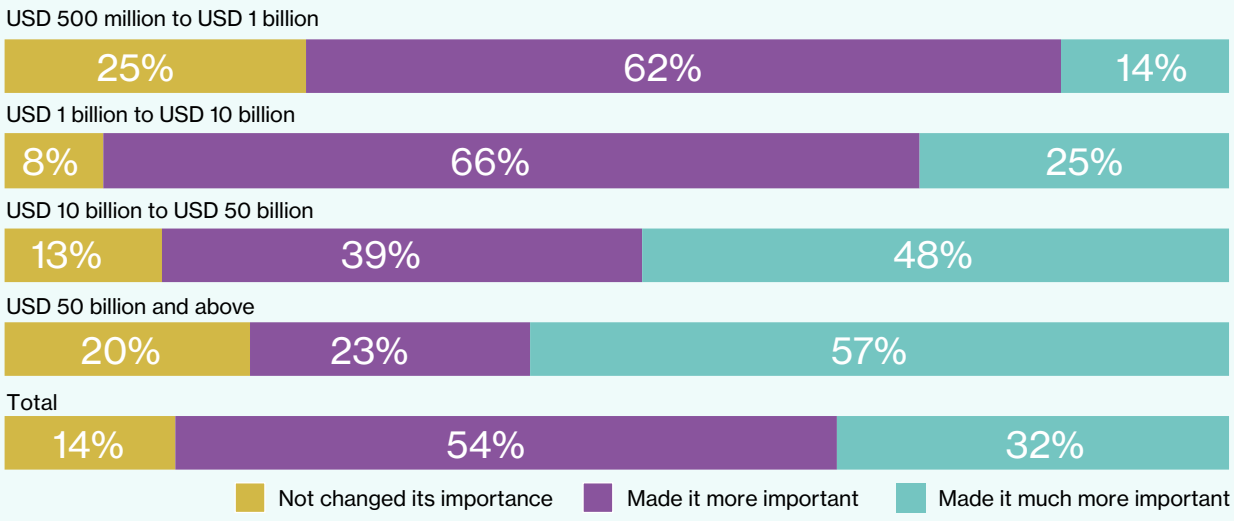
Keeping internal data protected

Probably the biggest privacy concern is with internal data finding its way into public foundation models. Customisation increases the risk of this happening. "As you train information into the models, the models themselves are not capable of following any sort of role-based authentication," says Boyd. "If you put sensitive information into the model, access needs to be restricted to people whom you trust with that information."

This is one reason why Dentsu integrates multiple layers of security and privacy safeguards into its models. "For us, first-party data belongs solely to our clients," says Demitros. "We set a very high bar for its use and do not leverage it to enhance our proprietary models. Under no circumstances can it be used to benefit another client."

Figure 6: The bigger the business, the greater the concern for security

Generative AI has made the privacy, safety and security of custom apps a greater priority for organisations of all sizes, but especially the largest.



Source: MIT Technology Review Insights survey, 2025

"There is some tension today with the trade-off between the value of customisation versus its perceived risk. As AI maturity grows and more organisations customise, the perceived risk will decline."

Gabe Pereyra, President and Co-Founder, Harvey AI

"There are so many moving parts to these models, and many of the security concerns are valid if you're deploying them in the wrong way," says Pereyra. He believes, however, that most issues with data leakage or contamination will eventually go away. "There is some tension today with the trade-off between the value of customisation versus its perceived risk," he says. "As AI maturity grows and more organisations customise, the perceived risk will decline."

Ensuring model and application integrity

Generative AI's risk exposure extends beyond data privacy threats. The surveyed companies are being proactive against a range of threat vectors, among which the most prominent are hallucinations (cited by 60%), the use of compromised or malicious models (58%) and prompt injection attacks (55%), which attackers use to manipulate a model's outputs (see Figure 7).

Avoiding hallucinations

Hallucination is actually less a security issue than one of algorithm, search tool or data quality, resulting in a model generating incorrect or misleading information. There is no silver bullet to eliminate hallucinations, according to Austin of AT&T. "You need a variety of approaches to check, catch and minimise them," he says. "Good prompting helps. And using RAG to produce citations for every model answer is proving effective for us."

Responsible AI and media insights at Dentsu

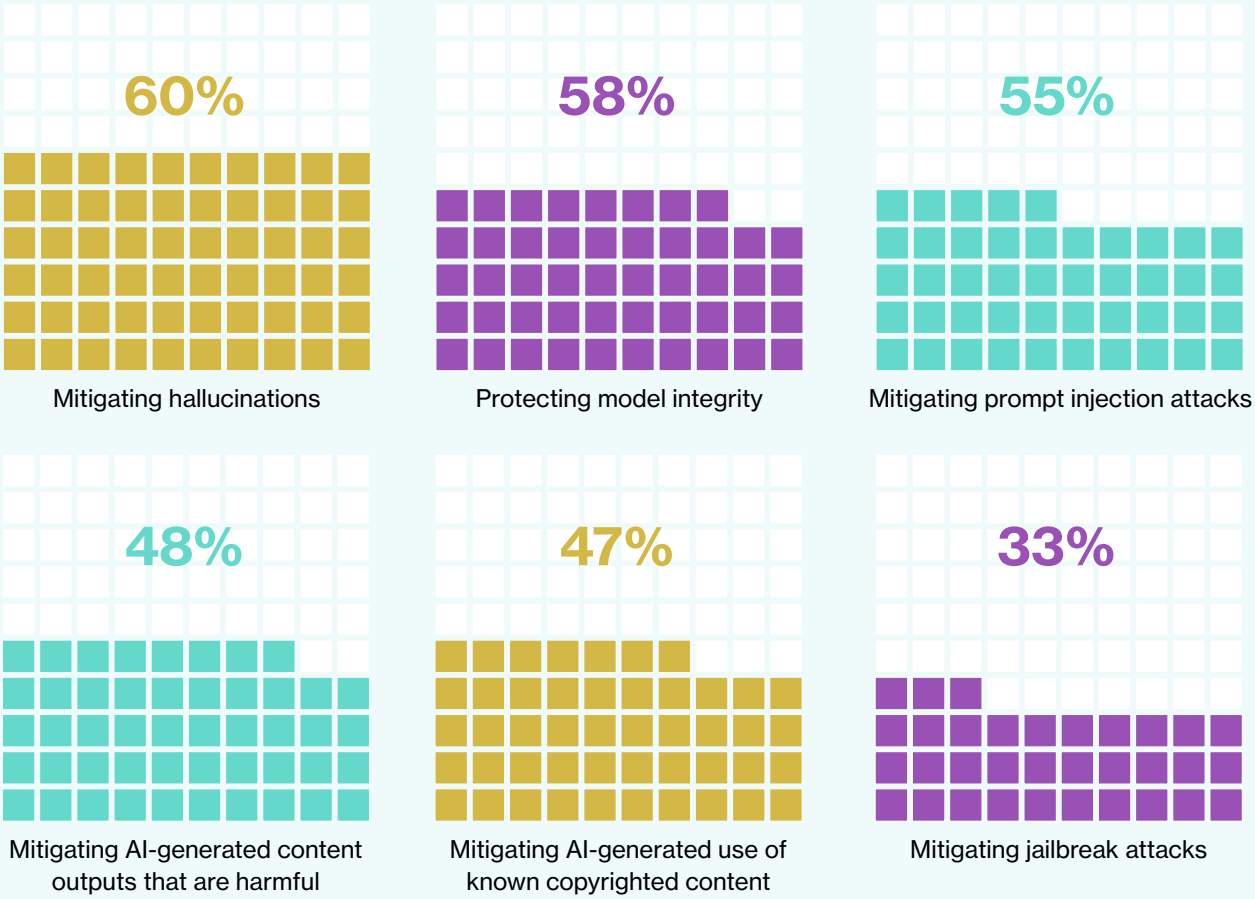
Media and advertising company Dentsu is speeding up the process[3] of accessing increasingly complex media and consumer analytics with a chat-based predictive analytics copilot. The new agent can interact with natural language and draws on Dentsu's extensive media metrics, including forecasting, budgeting, modelled client data and best practices. What once took a team of data scientists weeks to sift through multiple systems, can now be done 90% faster, enabling Dentsu to quickly respond to trending topics and emerging technologies for more innovative campaigns. "Delayed campaigns slow customer service and may result in missed opportunities," says Becca Kline, senior director of analytics at Dentsu.

The business is using Microsoft Azure AI Foundry for the copilot to create a system that interacts smoothly with its suite of business apps. The application architecture consists of loosely coupled microservices and multiple generative AI agents that act autonomously, but can collaborate using API-based integration and the GraphQL protocol. "The idea for this copilot was to sit within Dentsu's suite of applications as another kind of microservice that maintains the same look and feel and inherits a shared set of governance modules," says Callum Anderson, global director for DevOps and SRE at Dentsu. "We have to be responsible in how we use AI for all our clients," he explains. "Everything we did, we considered through the lens of how we governed and ensured the AI is responsible."

Dentsu uses many Microsoft technologies, including Azure OpenAI Service, Azure API Management, Azure Kubernetes Service and Azure Data Factory.

Figure 7: Top generative AI threat vectors

Reducing hallucinations, protecting model integrity and mitigating prompt injection attacks are the threat vectors respondents are most focused on.

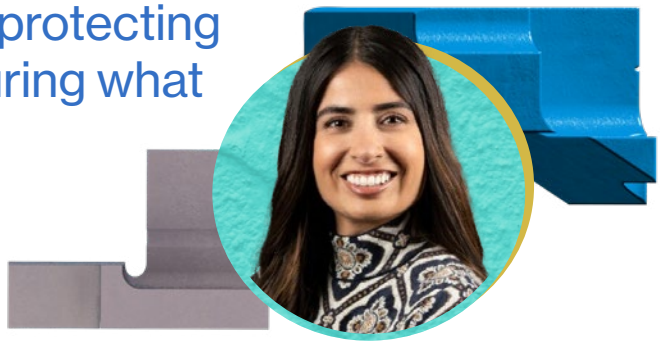


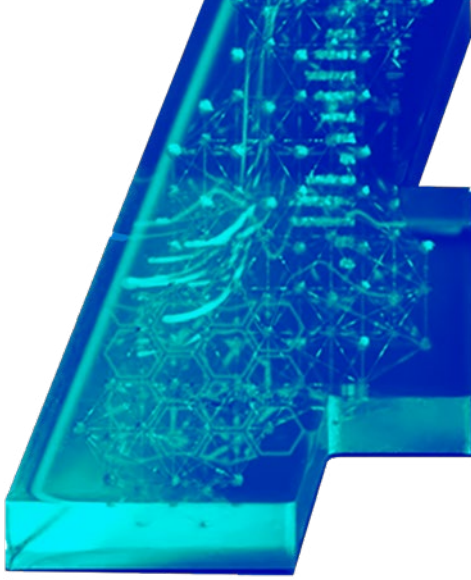
Source: MIT Technology Review Insights survey, 2025

Sharma sees a shift occurring in how some companies are thinking about model protection. "The purely security mindset is giving way to what I call an AI trust architecture," she says. "We're entering an era where model security and privacy become less about restrictions and more about enabling innovation. It will not just be about protecting what we have, but also about securing what we can create."

"We're entering an era where model security and privacy become less about restrictions and more about enabling innovation. It will not just be about protecting what we have, but also about securing what we can create."

Asha Sharma,
Corporate Vice President and Head of Product, AI Platform, Microsoft



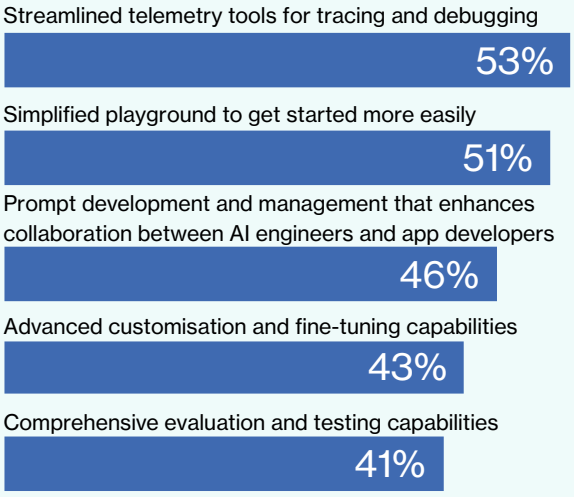


05 Continuous operations

Managing entire portfolios, with several dozen models and hundreds of AI-enabled applications poses a set of complex challenges. According to Demitros, generative AI is a double-edged sword for lifecycle management. One edge is the challenge of keeping pace with advances in generative AI capabilities. "When the base functionality changes, the next iteration blows away all the benchmarks that you've customised around, so now your entire product, or a portion of it, needs to change," he says. "You have to be willing and able to scrap product much more rapidly."

Figure 8: Tools of the trade for next gen software development

Streamlined telemetry, a simplified playground and prompt development and management are respondent organisations' top development tools for generative AI software.



Source: MIT Technology Review Insights survey, 2025

The other edge of the sword is a massive opportunity. Demitros notes that software developers only get to spend part of their day coding. Much of their time is spent on administrative work, such as documentation and meetings, he says. "Generative AI is becoming very helpful in offloading a lot of those administrative tasks, freeing up more dev cycles."

Empowering dev teams

We asked our global survey respondents how they are empowering their development teams building with AI applications. We learned that companies are implementing a range of different tools and techniques. Just over half (53%) are making streamlined telemetry tools for tracing and debugging capabilities available to teams. A similar share (51%) of surveyed companies are providing a simplified playground of development tools so teams can get started developing custom AI applications more easily. And 46% are using prompt development and management features that accelerate the creation, evaluation and deployment of model prompts. This additionally helps to enhance collaboration between AI engineers and app developers (see Figure 8).

Telemetry tools for tracing and debugging point the way toward complete observability of code generation when working with AI models. This is the next generation of AI application development, according to Sharma. "AI debugging shows how reasoning flows through your system from initial prompt through model decision to final output," she says. "It also enables performance optimisation, showing teams which prompts are most effective and how different approaches impact accuracy and costs." And tracing enhances transparency. "When something unexpected happens, these tools lets you trace back through the AI's decision-making process," says Sharma. "That builds trust."

Identifying use cases

For all the internal capabilities that businesses are creating to manage their generative AI operations, many will require help in several areas for some time to come. When asked where they currently need support, by far the most common response is the identification of use cases, cited by 76% of survey respondents. Scaling (mentioned by 47%), establishing prototypes (44%), performance and quality monitoring (44%) and preparing solutions for deployment (42%) are other major areas where external support and advice are needed.

Figure 9: Help needed to make sense of use cases

The vast majority of respondents say their organisation needs help early in the AI development process.

Identifying the business use case and success criteria for a custom AI project	76%
Scaling a successful generative AI solution to more users (and with larger data sets)	47%
Monitoring performance and quality continuously (adapting to ongoing change)	44%
Establishing prototypes by adjusting prompts and swapping models	44%
Preparing a generative AI solution for deployment	42%
Evaluating performance and quality of your AI solution	38%
Discovering and selecting the right model for your use case	37%
Connecting your AI model selection with your organisational data	34%
Managing feedback on a deployed generative AI solution	24%
Reverting learnings from one generative AI project into a new idea	3%

Source: MIT Technology Review Insights survey, 2025

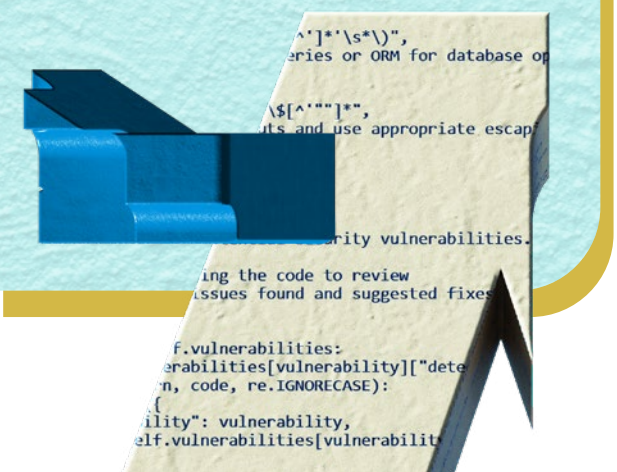
Automation for lifecycle management

According to Mark Austin, AT&T has 55 generative AI use cases in production today. One of the biggest, he says, is an agentic framework to automate generative AI across the full lifecycle of software development.

The initiative evolved from Ask AT&T, a generative AI application the company launched in 2023 to help employees interact with data. The first thing Austin and his team saw was about 40% of the questions being put to it were about coding. The benefit of this was obvious, he says: “We naturally don’t want people pasting their code out on the internet and asking things like ‘How do I fix this?’ We saw that they were asking such questions internally, and that was hugely important.”

The internal application has since evolved to become more than a coding tool, encompassing the full software development lifecycle. Austin explains: “It starts with someone writing in plain English that they need to develop or modify code to do something specific. The tool takes that and generates a plan and frameworks for how the application will be built, whether from scratch or atop an existing one. Then it writes the code, writes the test scripts to test the code, checks for software vulnerabilities, all the way through to deployment.”

Austin and his team have found employees are accepting around one-third of the code that they’re getting from Ask AT&T. “It helps us go faster,” he says. “It also helps us to develop more securely by catching software vulnerabilities or security issues on the front end. And it’s a big time saver.”



06 Conclusion

Estimates vary but, over time, generative AI is likely to add hundreds of billions of dollars, or more, to the world's GDP. If it does, customisation is likely to unleash a significant portion of that value. As we have illustrated throughout this report, it is when organisations can tailor generative AI models and applications to their specific needs, and make the most of their own data and expertise in doing so, that the technology's full potential can be realised. As powerful as they are, today's one-size-fits-all foundation models cannot achieve this.

Customisation is not without its challenges, even for the biggest of organisations with substantial resources at their disposal. It requires high-quality and well-governed domain-specific data. It requires deep collaboration among teams of specialists across AI, applications, data and infrastructure. And it demands confidence that the right safeguards are in place to protect modified models and applications against data leakage and malign actors.

Our research highlights several aspects of generative AI customisation that businesses – particularly those with lower levels of AI maturity – should consider carefully as they customise more of their models and applications. Prominent among them are the following:

Rigorous evaluation is worth the time spent. Manual study of available models and applications prior to selection, along with benchmarking, is the right approach to ensure a smooth path for later customisation. There is a strong case for automated evaluation as the organisation's generative AI use cases grow in number and the data sets they need grow in size.

Customisation methods work best in combination. RAG is an effective and widely used method for improving generative AI models, but its utility depends on the use case. More often than not, it is most effective when employed with other methods, particularly fine-tuning and prompt engineering, that perform complementary roles.

Customisation with data security in mind. Leakage of sensitive data into public models is a real concern. But strong model and data governance is an effective safeguard. And teams can act to augment this, such as by adding more security and privacy capabilities to its models and applications.

Embrace the holistic abilities of customisation. It's not just about tailoring individual models and applications to the needs of specific use cases. Generative AI can be customised to design and development processes across entire portfolios. Businesses should explore its advantages for improving applications portfolio management as a whole.

About MIT Technology Review Insights

MIT Technology Review Insights is the custom publishing division of *MIT Technology Review*, the world's longest-running technology magazine, backed by the world's foremost technology institution – producing live events and research on the leading technology and business challenges of the day. Insights conducts qualitative and quantitative research and analysis in the US and abroad and publishes a wide variety of content, including articles, reports, infographics, videos and podcasts.

About Microsoft Azure

Microsoft Azure is a leading cloud platform for building, deploying and managing custom AI applications at scale. Launched in 2010 as a pivotal shift from on-premises data centres to the cloud, Azure continues to grow with extensive capabilities that go far beyond infrastructure. With comprehensive services and tools for developers, AI, data and apps, Azure delivers a cohesive approach to cloud computing that's unmatched. Its open, flexible platform is designed to empower companies of all sizes, across industries and at any stage of AI transformation.

Endnotes

1. 'AT&T improves operations and employee experiences with Azure and AI technologies', Microsoft, May 18, 2023, https://www.microsoft.com/en/customers/story/1637511309136244127-att-telecommunications-azure-openai-service.
2. 'Harvey makes lawyers more efficient with Azure AI infrastructure', Microsoft, December 3, 2024, https://www.microsoft.com/en/customers/story/19750-harvey-azure-open-ai-service.
3. 'Dentsu reduces time to media insights by 90% using Azure AI', Microsoft, November 19, 2024, https://www.microsoft.com/en/customers/story/19582-dentsu-azure-kubernetes-service.

Illustrations
Illustrations assembled by Peter Crowther Associates Ltd.

MIT Technology Review Insights

www.technologyreview.com

insights@technologyreview.com