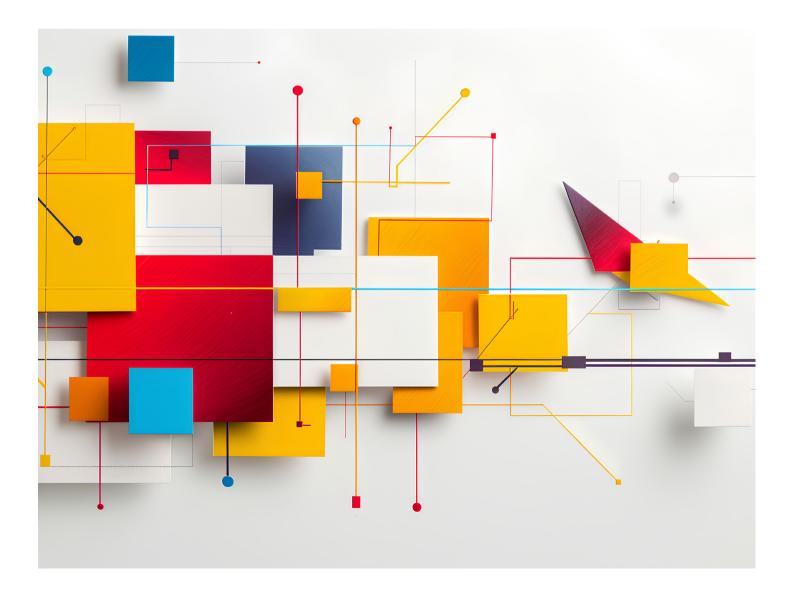
### MIT Technology Review Insights

Produced in partnership with



A new software stack for generative AI will prepare organizations for the challenges of moving from AI pilots to AI in production.

# Moving generative Al into production



enerative AI has taken off. Since the introduction of ChatGPT in November 2022, businesses have flocked to large language models (LLMs) and generative AI models looking for solutions to their most complex and labor-intensive problems. The promise that customer service could be turned over to highly trained chat platforms that could recognize a customer's problem and present user-friendly technical feedback, for example, or that companies could break down and analyze their troves of unstructured data, from videos to PDFs, has fueled massive enterprise interest in the technology.

This hype is moving into production. The share of businesses that use generative AI in at least one business function **nearly doubled this year to 65%**, according to McKinsey. **The vast majority of organizations (91%)** expect generative AI applications to increase their productivity, with IT, cybersecurity, marketing, customer service, and product development among the most impacted areas, according to Deloitte.

Yet, difficulty successfully deploying generative AI continues to hamper progress. Companies know that generative AI could transform their businesses – and

### Methodology

In August 2024, MIT Technology Review Insights conducted a poll on the challenges and choices organizations face in deploying generative AI use cases. The 250 executive respondents represent a broad range of industries and work at organizations across the globe.

### Key takeaways

Businesses are embracing the transformative potential of generative Al and are eager to move the technology into production. Implementation difficulties, however, are slowing deployment for many.

2

Executives cite AI output quality, integration complexity, high costs for model inferencing and training, and application latency as top challenges – and they are looking to compound AI systems as part of the solution.

Building a solid and adaptable AI stack – supporting a variety of base models, including integration solutions, and offering next-generation tooling – will be foundational to businesses' success with generative AI.

that failing to adopt will leave them behind – but they are faced with hurdles during implementation. This leaves **two-thirds of business leaders** ambivalent or dissatisfied with progress on their AI deployments. And while, in Q3 2023, **79% of companies said they planned** to deploy generative AI projects in the next year, **only 5% reported having use cases in production** in May 2024.

"We're just at the beginning of figuring out how to productize AI deployment and make it cost effective," says Rowan Trollope, CEO of Redis, a maker of real-time data platforms and AI accelerators. "The cost and complexity of implementing these systems is not straightforward."

Estimates of the eventual GDP **impact of generative AI** range from just under \$1 trillion to a staggering \$4.4 trillion annually, with projected productivity impacts comparable to those of the internet, robotic automation, and the steam engine. Yet, while the promise of accelerated revenue growth and cost reductions remains, the path to get to these goals is complex and often costly. Companies need to find ways to efficiently build and deploy AI projects with well-understood components at scale, says Trollope.

### "We're just at the beginning of figuring out how to productize AI deployment and make it cost effective."

Rowan Trollope, CEO, Redis

### Challenges and deployment complexity

Moving generative AI into production isn't easy. When asked about production challenges, nearly three-quarters of business leaders (72%) said they worry about the quality of outputs from their AI systems. In the same MIT Technology Review Insights poll, about 6 in 10 said they also worry about integration, high costs for model inferencing, and high costs for model training.

Because most companies are not training their own generative AI models, proper integration is necessary to deliver the right data and context to the model, says Harrison Chase, co-founder and CEO at LangChain, which focuses on integration and orchestration solutions. "At a high level, one of the big problems is getting the right context to the model," he says. "You may want to take a result from a previous LLM, and basically you need to get all this data and pipe it through." AI developers need "an orchestration layer that helps with that orchestration of context," he adds.

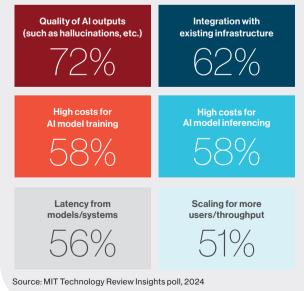
In addition, companies need to be able to determine how much context to provide to a model. Generative Al systems typically perform better with more context, but more context equals more cost, so finding the right tradeoff is important, says Trollope.

A good estimate for the amount of context provided is the number of tokens used by a model. "For many, it's all about the cost per token, so if you can get that cost per token down, you can make inferencing more efficient," says Trollope.

Costs were a pressing concern in the survey, perhaps in part because determining the benefits returned by generative AI remains difficult. Calculating return on investment for AI systems and products is a complex process with significant uncertainties, and deployment and operations costs can be hard to quantify. **Deloitte**  **research found** that businesses are "struggling to define and measure the exact impacts of their Generative AI initiatives," with 48% using key performance indicators and 38% creating business-specific frameworks for evaluating generative AI investments.

Business will continue to hold back until they can reliably determine return on investment, says Chase. "Companies want to estimate the ROI ahead of time, and it's still pretty hard to do," he says. "Quantifying the ROI ahead of time or even after the fact is still really challenging, and I think that's a big blocker to even getting to the building stage."

## Challenges faced by businesses putting generative Al apps into production Do you face the following challenges when building generative Al apps in production?

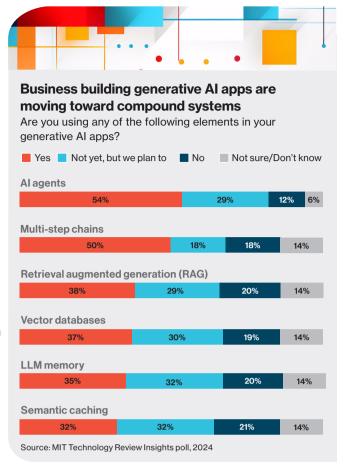


#### The compound AI advantage

In many cases, compound AI systems, which bring together several AI models, technologies, or capabilities, are emerging as effective solutions to deployment challenges. Compound AI systems, sometimes called AI agents or agentic systems, might link different AI models that specialize in different tasks, combine several AI-related technologies, or integrate individual AI modules whose different capabilities combine to tackle more complex tasks. These types of systems have attracted broad interest among survey respondents, with a majority (54%) reporting that they are already using AI agents and another 29% planning to in future.

Compound AI systems that chain together multiple task-specific models can be used to reduce costs and improve performance, says Chase. "Once the application starts to take off, then companies ask, can I use a cheaper model here?" he says. "Maybe I use an expensive model in one place, but cheaper models in another place. Or maybe I take that expensive model and I break it up into two or three cheaper model calls, because that's what's needed for cost and latency." Semantic routing is another common solution here. "It routes the user to the right tool, which might be another model, or it might not even be Al. It might route the user request to a scheduling tool or even a human to address their need," says Trollope. Fully half of survey respondents say they are currently using multi-step chains in their generative AI applications, and another 18% say they plan to in the future.

Retrieval-augmented generation (RAG) is a compound Al technique that adds a retrieval component to the generative one. The retrieval capability allows the system to search through files, documents, and data, finding highly-relevant and grounded information on which to base its outputs. This allows organizations to adopt general-purpose Al models and then tailor and refine

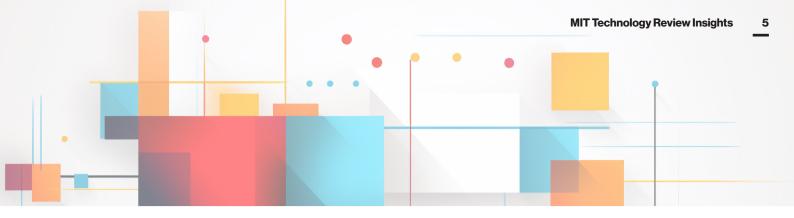


them for their own use. Interest in RAG is high among survey respondents, with 38% saying they are using the technique now, and an additional 29% saying they plan to soon.

Semantic caches and vector databases are two additional components that can add efficiency to AI systems. Semantic caches group together similar and duplicate model queries based on their meaning and context (not just the literal words used), and then store and reuse model responses as appropriate. Vector databases are the specialized databases used to store the high-dimensional vectors, or "embeddings," that represent these queries and responses and that enable

"Companies want to estimate the ROI ahead of time, and it's still pretty hard to do. Quantifying the ROI ahead of time or even after the fact is still really challenging, and I think that's a big blocker to even getting to the building stage."

Harrison Chase, Co-founder and CEO, LangChain



matching by similarity. Survey respondents are adopting both of these tools, with 37% saying they are currently using vector databases and 32% using semantic caching. Interest in these technologies is also growing quickly, with about another third of respondents (30% for vector databases and 32% for semantic caching) saying they plan to adopt them in future.

Redis estimates that between 30% and 80% of queries to a large language model (LLM) are duplicates, so the efficiency gains from caching can be substantial. **One research group found** frequent repetition even at the single-user level, calculating that an average 31% of participants' ChatGPT queries were repetitive of their own previous queries.

Semantic caches, then, can significantly speed up such requests and significantly reduce model inferencing costs, says Trollope. "We have customers who have 90% of their LLM calls answered using our semantic caching, which is built on a vector database," he says. "Those users don't have to wait for the LLM, and the customer controls their costs."

#### Building the stack

As organizations turn their AI ambitions into reality, and shift their generative AI applications into production, they will need to focus on building the technology stack to support them. A first step will be determining which base model – or models – they want to build on.

The majority of organizations surveyed (67%) have begun by building generative AI applications with third-party closed-source models, such as OpenAI's. But many have aspirations to incorporate other types of models as well.

Open-source models are trending, likely driven by companies' desires to control security and costs. Nearly three-quarters of respondents say they are either using open-source cloud-based models now (42%) or plan to do so in the future (32%). Another 41% report interest in open-source models on-prem: 17% are using these currently, and 24% plan to in the future.

Developing integrations between technologies is also important to the adaptable AI stack. While prompts do not necessarily readily transfer between models, common application programming interfaces (APIs) can help companies swap in more efficient or better performing components.

Companies including Redis and LangChain support this flexibility by providing standard interfaces or integrations between top AI tools. "The APIs may not have all the same parameters, and so this is where using something

			•••		
Businesses are exploring a range of model options Are you using any of the following AI model(s) in your generative AI apps?					
Yes Not yet, but we plan to No Not sure/Don't know					
3rd party, closed-source models in the cloud (such as OpenAl's)					
67%		18%		12% 3%	
Open-sourced models in the cloud					
42%		32%		22%	4%
Proprietary models in the cloud					
40		25%	28	%	7%
Proprietary models on-prem					
20%	<b>22</b> %	38%		20%	
Open-sourced models on-prem					
17%	24%	37%		22%	
Source: MIT Technology Review Insights poll, 2024					

like LangChain can help," says Chase, "because we have a standard interface for all the different models, and so you can swap between them."

Survey respondents show a great deal of awareness and sophistication about what the next steps in generative Al will be, and a large share say that their companies have mature Al capabilities in place or on the immediate roadmap. Yet the gap persists in what companies have been able to put into production so far.

Trollope says there is a lag in the adoption of tooling. Organizations need to deploy the next-generation AI tools and solutions that will enable them to build, manage, and integrate the generative AI applications of their imaginations. "The industry needs to catch up to the core research breakthrough, with the implementation tooling that's necessary to make it easy," he adds. "Once that tooling is in place, I think we'll see AI development accelerate."

As businesses build their AI stacks, in collaboration with partners like Redis and LangChain, they are also readying themselves for the AI innovations of the future. "Enterprises need models that run quickly with a high speed data platform that can ensure when you pull up data, it comes up fast," says Trollope. "You can't have a lot of latency in these systems."

"The industry needs to catch up to the core research breakthrough, with the implementation tooling that's necessary to make it easy. Once that tooling is in place, I think we'll see AI development accelerate."

Rowan Trollope, CEO, Redis

### Managing latency in generative Al applications

Even as the capabilities of generative AI grow, its users expect more from its responsiveness. Real-time voice interactivity, for example, is a powerful and intuitive technology—and as it becomes more widely available, users will demand that it function at the speed of human conversation. Reasoning models like OpenAI's o1 can provide much better answers if they are given time to think. All of this will have to be done quickly to meet user expectations.

"You cannot have any lag in these experiences," says Rowan Trollope, CEO at Redis. "Latency is the new downtime." Survey respondents agree—56% report that latency from models or systems is a challenge they've encountered when bringing generative AI apps to production.

Harrison Chase, CEO at LangChain, sees organizations recognizing the need for speed as they develop their generative AI applications. "A sneaky big challenge," he says, "is figuring out the right UX for these applications. One of the consistent themes we hear is that people often spend as much time on the UX as they do on the prompt engineering, because a lot of the UX has to do with latency."

As organizations move to more sophisticated compound Al solutions, latency will become an increasing concern. "Compound Al equals compound latency and compound cost," says Trollope. "In a multi-model system, each model adds its own latency as you add them up in the stack." As companies explore these more complex Al systems, their effects on speed will remain a prominent challenge. "Moving generative AI into production" is an executive briefing paper by MIT Technology Review Insights. We would like to thank all participants as well as the sponsor, Redis. MIT Technology Review Insights has collected and reported on all findings contained in this paper independently, regardless of participation or sponsorship. Teresa Elsey was the editor of this report and Nicola Crepaldi was the publisher.

### 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. And through its growing MIT Technology Review Global Insights Panel, Insights has unparalleled access to senior-level executives, innovators, and entrepreneurs worldwide for surveys and in-depth interviews.

### From the sponsor

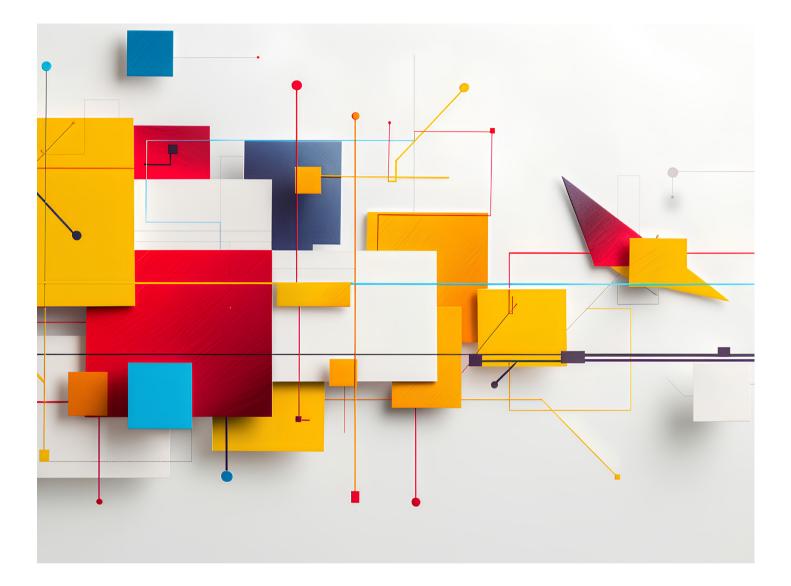
Redis is the world's fastest data platform. From its open source origins in 2011 to becoming the #1 cited brand for caching solutions, Redis has helped more than 10,000 customers build, scale, and deploy the apps our world runs on. With cloud and on-prem databases for caching, GenAI, and more, Redis helps digital businesses set a new standard for app speed. Located in San Francisco, Austin, London, and Tel Aviv, Redis is internationally recognized as the leader in building fast apps fast. Learn more at **redis.io**.



#### Illustrations

Cover art and spot illustrations created with Adobe Stock.

While every effort has been taken to verify the accuracy of this information, MIT Technology Review Insights cannot accept any responsibility or liability for reliance by any person on this report or any of the information, opinions, or conclusions set out in this report.



### MIT Technology Review Insights

www.technologyreview.com insights@technologyreview.com