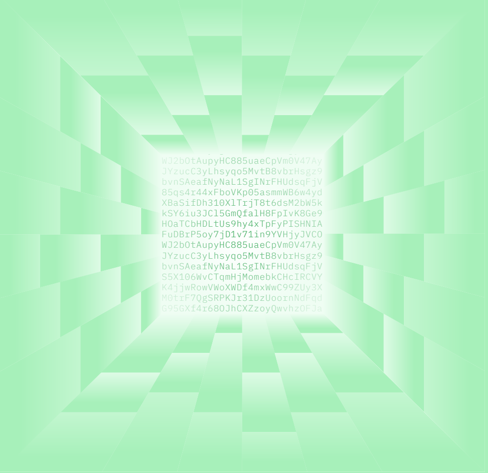


Tailor-made gen AI delivers precision power

Generative AI is unlike any technology that has come before. It's swiftly disrupting business and society, forcing leaders to rethink their assumptions, plans, and strategies in real time.

To help CEOs stay on top of the fast-shifting changes, the IBM Institute for Business Value (IBM IBV) is releasing a series of targeted, research-backed guides to generative AI on topics from data security to tech investment strategy to customer experience.

This is part 18: AI model optimization

There's a gen AI model for that

ChatGPT made everyone feel like an AI expert. But its simplicity is deceptive. It masks the complexity of the generative AI landscape that CEOs must consider when building their AI model portfolio.

Generative AI models come in many flavors. What they can do, how well they work—and how much they cost—varies widely. Who owns the model, how it was developed, and the size of its training dataset are just a few of the variables that influence when and how different models should be used.

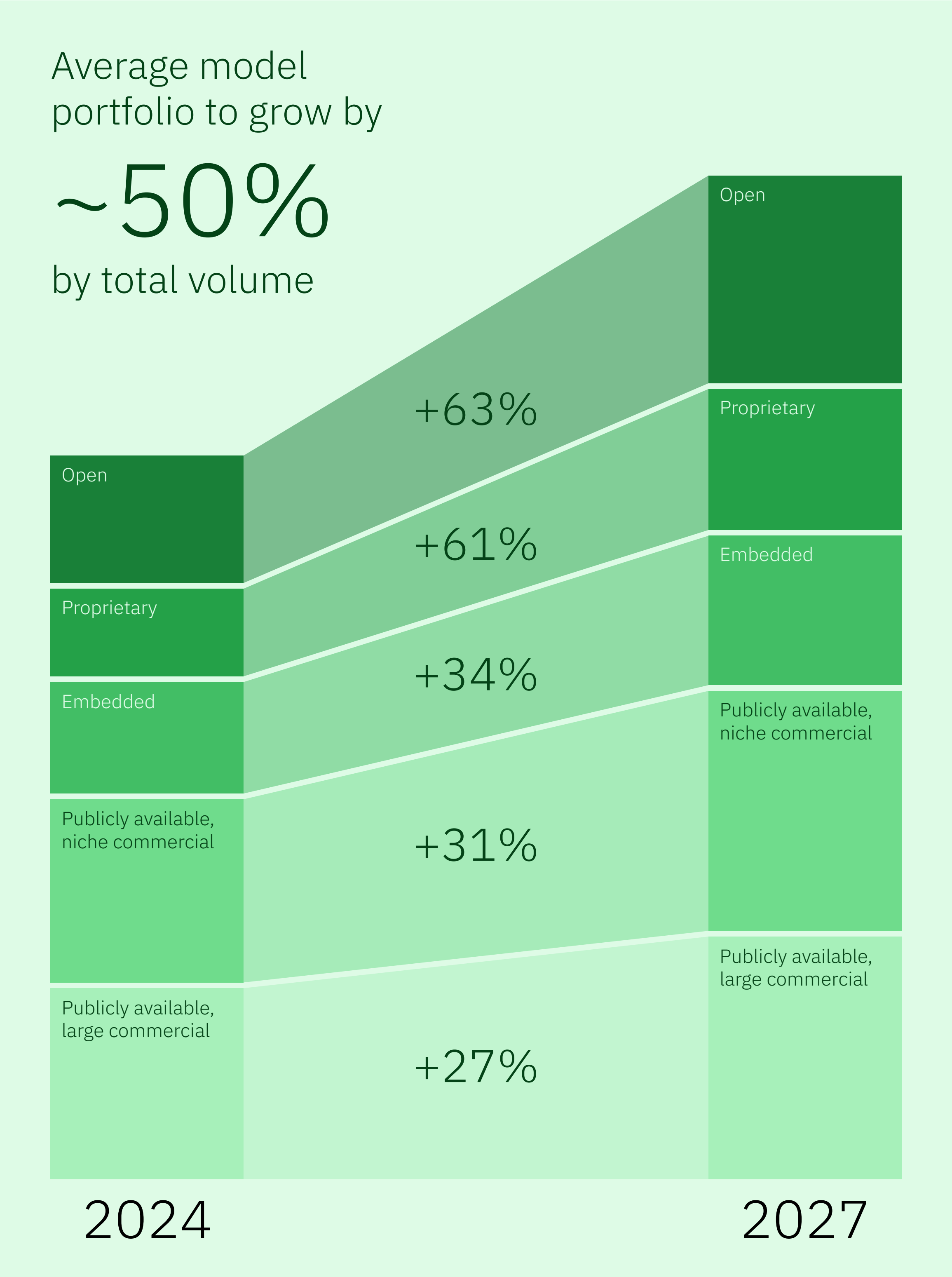
With the massive amount of data and resources it takes to train a single large language model (LLM), the question of size is monopolizing many conversations about gen AI. As a result, many CEOs wonder whether they should scale large gen AI models for their business. Or if they should develop smaller, niche models for specific purposes.

The answer is, they need to do both.

And many already are. A typical organization uses 11 generative AI models today—and expects to grow its model portfolio by ~50% within three years.

Why so many? Because every use case comes with its own requirements and constraints. And different business problems demand different types of models.

For example, tasks that are highly specialized, such as image editing or data analysis, need gen AI models that are trained on smaller, niche datasets. Work that is sensitive or proprietary requires gen AI models that can be kept confidential— and close to the vest. More general tasks, such as text generation, may call for gen AI models trained on the largest datasets possible.



Average model portfolio to grow by ~50% by total volume

While CEOs should have teams that understand all the details about what sets different models apart, you do need to know that picking the right model for each task—each application of generative AI—matters. Knowing what drives cost, environmental impact, and business value will help you optimize the performance of your AI portfolio—and give your teams the tools they need to beat the competition.

IBM Institute for Business Value research has identified three things every leader needs to know:

1. There’s no such thing as an all-purpose gen AI model.
2. Gen AI costs are completely in your control.
3. Gen AI advantage is fleeting.

And three things every leader needs to do right now:

1. Give some teams the sledgehammer—and others the scalpel.
2. Find your own gen AI sweet spot.
3. Make your models work harder.

1. Agility + Generative AI

What you need to know →

There's no such thing as an all-purpose gen AI model

Generative AI helps organizations move faster with precision and agility—if they use the right model, running in the right environment, for the right purpose.

While technology leaders are best positioned to decide which gen AI model to use where, understanding the pros and cons of different model types—as well as where the competition is headed—helps CEOs make more informed investment decisions.

Model type	Example and characteristics
Open	**Granite, Mistral** – Training varies by size/specialization – Focus on transparency and accountability – Varying degrees of openness by company/model – Greater potential for innovation
Proprietary	**Custom-developed enterprise model** – Training effort borne by enterprise – Greater control over scope and data – Greater potential for differentiation
Embedded	**Models in SAP Joule, Salesforce Einstein, Adobe Firefly** – Integrated into existing enterprise software – Typically leverages existing models as software product feature – Not typically usable independently
Publicly available, niche commercial	**Google Med-PaLM** – Trained on large, specialized data sets – Focus on depth and specialization – Typically opaque – Some potential for differentiation
Publicly available, large commercial	**GPT-4** – Trained on massive data sets – Focus on breadth and depth – Typically opaque – Limited potential for differentiation

What you need to know

There's no such thing as an all-purpose gen AI model (cont.)

For example, it's helpful to know that publicly available, large commercial models (such as GPT-4) only account for about a quarter of the models in use at a typical organization. Niche publicly available commercial models, such as Google Med-PaLM, account for another 23%, with open models, such as Granite and Mistral, at 16%, embedded models, such as those in SAP Joule, Salesforce Einstein, and Adobe Firefly, at 14%, and an organization's own custom-developed, proprietary models at 11%. Other models account for 12%, filling out the portfolio.

Size is one of the first factors tech leaders consider when deciding which gen AI model to use for which workflow. Large models, which are trained on hundreds of billions of parameters, offer greater breadth and depth of expertise, and can handle more complex tasks—but they come with a higher price tag and larger carbon footprint. Smaller niche models, which are typically trained on tens of billions of parameters, can perform more precisely, quickly, and efficiently when trained to do specialized work, such as translating code or content into a specific language.

Model ownership is another crucial consideration. While public, commercial gen AI models are popular—making up about half of the average organization's AI portfolio—they do have their limitations. Because they're available for any organization to purchase or license, they can't do much to drive competitive differentiation, as everyone is working from the same corpus of data. Public models can help teams work faster and more efficiently but, because they run on a public cloud, they don't offer the privacy or control businesses need to tackle mission-critical tasks.

This is where companies' proprietary gen AI models come in. Because these models are developed, owned, and controlled by the organizations that use them, leaders can dictate what data will inform their outputs. This reduces the chance a model—and the work products it informs—will be polluted by bad intel. These proprietary models also give tech leaders more flexibility to decide whether to run the model in a local environment or on the cloud, as well as how information provided by users should be stored or used to fine-tune the model's performance. This reduces the risk that private or sensitive data will be used or shared inappropriately. This is a crucial capability, as misuse, privacy, and accuracy are executives' top concerns when selecting a gen AI model.

Open gen AI models, which are built transparently with the help of open-source developer communities and can be large or small, also address these concerns. Because they are built openly, they offer visibility into the data that was used to train the model. They're also carefully scrutinized, which means risks and issues, such as whether outputs violate intellectual property or copyright laws, can be identified and addressed quickly. Companies can then modify and customize these base models to accelerate innovation, improve performance, and build trust in gen AI.

Embedded gen AI models, which stem from a variety of sources, come fully baked into a platform or piece of software, such as SAP, Adobe, or Salesforce, to meet a specific need in the context of the software's functional footprint. They add value to the product they support, but they can't be used independently.

Gen AI model adoption is set to surge over the next three years, with open models leading the way. On average, executives expect their AI model portfolios to include 63% more open models than they use today, with the need for flexibility, transparency, and customization driving this growth. They also expect to use 27% more large commercial models—which are more reliable and easier to scale—and 31% more niche commercial models, which allow for greater specialization. In the same timeframe, they anticipate using 61% more proprietary models and 34% more embedded models.

1. Agility + Generative AI

What you need to do →

Give some teams the sledgehammer—and others the scalpel

Assess your portfolio of foundation models and determine how they align with strategic workflows. Invest in large gen AI models to boost productivity and tap niche models for more targeted tasks.

See the full gen AI spectrum. Understand what distinguishes different types of generative AI models, including LLMs, a company's custom-developed proprietary models, open models, and more. Be prepared to invest in different models for different purposes.

Map the AI terrain. Ask your AI leaders to create a comprehensive catalog of all gen AI models used across the organization, including their purpose, functionality, and performance metrics. Ensure the inventory is regularly updated to reflect changes in the AI landscape.

Find perfect pairings. Be sure your teams are appropriately matching gen AI models with the right workflows based on their strengths, weaknesses, and quirks. Identify where gaps exist—but don't use a set of encyclopedias when one dictionary will do.

2. Cost + Generative AI

What you need to know →

Gen AI costs are completely in your control

CEOs know they need generative AI—but at what cost? As gen AI stretches its tendrils into every area of the organization, business leaders say they first consider how they will achieve cost-efficiency at scale as they are selecting the right models to use in each circumstance.

When they consider barriers, 63% cite model cost and 58% cite model complexity as their top concerns.

Why is cost such an important consideration? Because it can vary widely depending on the model being used. For example, larger models come with more data storage and compute costs—which can result in higher cloud-related bills. Large models also require more frequent updates, fine-tuning, and maintenance, which come with talent costs. On the other hand, niche models have lower compute, data storage, and energy costs—and reduce the environmental impact of an organization’s AI portfolio. They can also be deployed faster and require less upkeep, which keeps people costs low.

Picking the right model size for the right task plays a big part in helping organizations manage gen AI costs. For example, complex tasks that leverage multiple skillsets and demand high accuracy, such as long-form writing, high-stakes decision-making, and testing research hypotheses may call for a larger, more expensive model. Niche models, which are more cost-efficient, are better for more targeted tasks—especially those where speed and efficiency are essential, such as real-time chat assistance, spam detection, data augmentation, and prototyping. Teams can also use advanced techniques, such as chain of reasoning, to break complex work into bite-sized tasks that niche models can handle, reducing their reliance on more cost-intensive LLMs.

As technology matures, niche models will become more proficient at handling a broader set of tasks, giving organizations a chance to get granular with cost management. By using models that are “fit-for-purpose,” meaning they’ve been designed, trained, and validated to meet specific requirements and objectives, teams can use only the resources they need for each task. And if companies use large models to train more focused niche models, they can make model development more cost-efficient.

In the near future, leaders may be able to improve cost management by using an enterprise gen AI control center to streamline decisions about which model should be used for which task. Adding a user-friendly experience layer that connects models, assistants, and prompts across the portfolio could let leaders bake in cost controls—as well as security, privacy, and compliance guardrails—to help ensure models are being used appropriately and efficiently by every employee, every time.

Executives say **cost** is the #1 barrier to adopting gen AI models.

2. Cost + Generative AI

What you need to do →

Find your own gen AI sweet spot

Discover the value of versatility. Right-size the gen AI models you use for each task to control costs and boost overall AI ROI.

Cultivate a model-agnostic mindset. Remain agile to adopt the models that have been optimized for price and performance, striking the right balance between accuracy, resource usage, and speed.

Engineer for efficiency. Tailor model scope to the deployment environment, favoring smaller, faster niche models for mobile and real-time applications, and larger models for high-accuracy, complex tasks.

Cut the fat. Establish clear performance metrics and benchmarks for each gen AI deployment. Use data-driven insights to see where gen AI is delivering the intended value—and where costs need to be reined in.

3. Competitiveness + Generative AI

What you need to know →

Gen AI advantage is fleeting

The competitive edge that generative AI delivers today will be table stakes tomorrow. As teams gain more gen AI experience—and the models themselves get smarter—CEOs must prioritize continuous improvement.

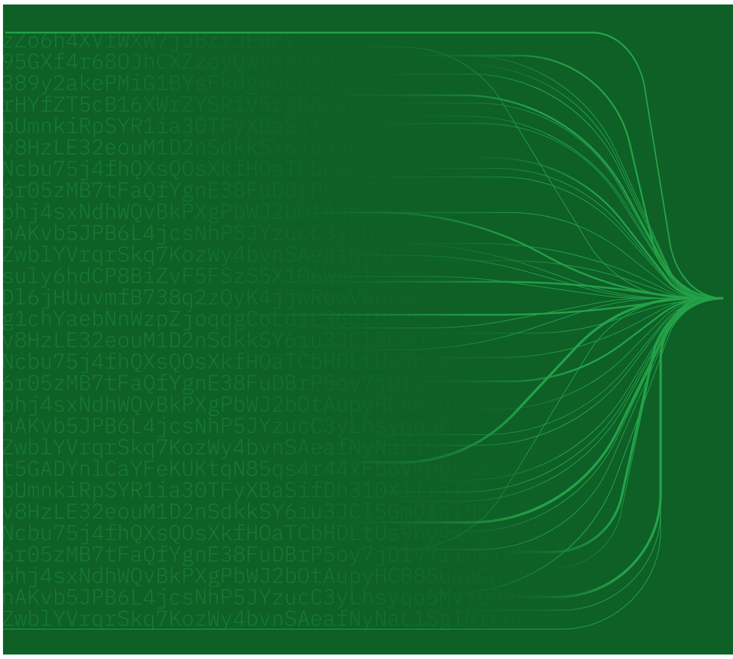
Organizations that commit to continual optimization can expect a notable performance boost. According to our research, organizations that use either fine-tuning or prompt engineering techniques report model outputs that are roughly 25% more accurate than others. Greater accuracy drives better forecasting, resource allocation, and personalization—all of which can boost the bottom line.

And yet, only 42% of executives say they always use prompt engineering, the process of designing inputs that will deliver the desired outputs, to enhance model accuracy.

But model optimization is only one piece of the puzzle. As portfolios evolve, so must model governance. This involves regularly updating the way the organization manages and controls model inventory, as well as who has permission to develop, train, and fine-tune models. Organizations also need clear processes for tracking model performance metrics, dealing with drift—where the accuracy of a model degrades over time—and correcting for bias in model outputs. This is all on top of the work teams must do to stay in line with rapidly changing regulations.

Organizations must also continually improve their AI infrastructure—their hybrid cloud strategies—to adopt more powerful AI models as they’re developed. As both data volumes and model complexity increase, tech infrastructure must be able to handle the heavier load. Then there’s the matter of scaling. As more teams use gen AI in all its forms, organizations need to evolve their infrastructure or cloud environment to meet increased demand.

What does this look like? Today, at least half of organizations are focused on optimizing network infrastructure, accelerating data processing, or distributed computing. Overall, 63% of executives say their organizations are using at least one infrastructure optimization technique.



- Fine tuning and prompt engineering improve model accuracy by 25%.

3. Competitiveness + Generative AI

What you need to do →

Make your models work harder

Don’t be satisfied with early successes. Continually push teams to aggressively improve model performance and outpace the competition by using the latest AI techniques and infrastructure.

Raise the gen AI bar. Add enterprise data into pre-existing gen AI models—in a private cloud or on-prem environment—to create value that is unique to your organization. Use fine-tuning, prompt engineering, and other optimization techniques to stay three steps ahead of the competition.

Future-proof your AI infrastructure. Invest in cloud-based services or specialized hardware, as well as open frameworks, so you can capitalize on continual AI-driven disruption.

Don’t get sidelined. Advance gen AI faster than your peers by establishing a clear governance framework. Question assumptions about your regulatory preparedness and become your own toughest critic.

IBM Institute for Business Value

The CEO's guide to generative AI

AI model optimization

The statistics informing the insights in this report are sourced from a proprietary survey conducted by the IBM Institute for Business Value in collaboration with Oxford Economics. The survey queried 200 US-based executives on their perspectives regarding AI model optimization in June 2024.

IBM Institute for Business Value

For two decades, the IBM Institute for Business Value has served as the thought leadership think tank for IBM. What inspires us is producing research-backed, technology-informed strategic insights that help leaders make smarter business decisions.

From our unique position at the intersection of business, technology, and society, we survey, interview, and engage with thousands of executives, consumers, and experts each year, synthesizing their perspectives into credible, inspiring, and actionable insights.

To stay connected and informed, sign up to receive IBV's email newsletter at ibm.com/ibv. You can also find us on LinkedIn at https://ibm.co/ibv-linkedin.

ibm.co/ceo-generative-ai-model-optimization

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America | July 2024

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

This report is intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. IBM shall not be responsible for any loss whatsoever sustained by any organization or person who relies on this publication.

The data used in this report may be derived from third-party sources and IBM does not independently verify, validate or audit such data. The results from the use of such data are provided on an "as is" basis and IBM makes no representations or warranties, express or implied.

WPXGRV7D-USEN-01