SPOTLIGHT
Sponsored by: IBM

**Generative AI is changing the world, and custom models are one of the best ways to develop, deploy, and use it.**

# The Truth About Successful Generative AI

*April 2024*

**Written by:** David Schubmehl, Research Vice President, Conversational Artificial Intelligence and Intelligent Knowledge Discovery, and Kathy Lange, Research Director, AI Software

## Introduction

While 2023 was a year of experimenting with generative AI (GenAI), enterprises will focus on realizing value from key GenAI use cases in 2024. They must shift from a prototyping mindset to operationalizing a few crucial applications, balancing cost against results. Operationalization involves testing for accuracy and scale, improving performance, tracking costs, and governing the data and the models. It also means orchestration across the infrastructure for prompt engineering, integration into data sources, and interfaces that enable processing quickly, repeatably, responsibly, and sustainably under load.

The application of GenAI in enterprises is just starting to unfold, and it has the potential to revolutionize industries and transform how businesses operate. Enterprises can leverage GenAI to drive innovation, automate repetitive tasks, improve decision-making, personalize customer and employee experiences, and boost efficiencies. Companies that effectively leverage these technologies will likely gain a significant competitive advantage. IDC research estimates that GenAI's worldwide economic impact will be close to $10 trillion by 2033. Survey respondents in IDC's October 2023 *Global AI (Including GenAI) Buyer Sentiment, Adoption, and Business Value Survey* indicated that the three most important business outcomes from GenAI are increased operational efficiency, cost savings, and improved employee productivity (see Figure 1).

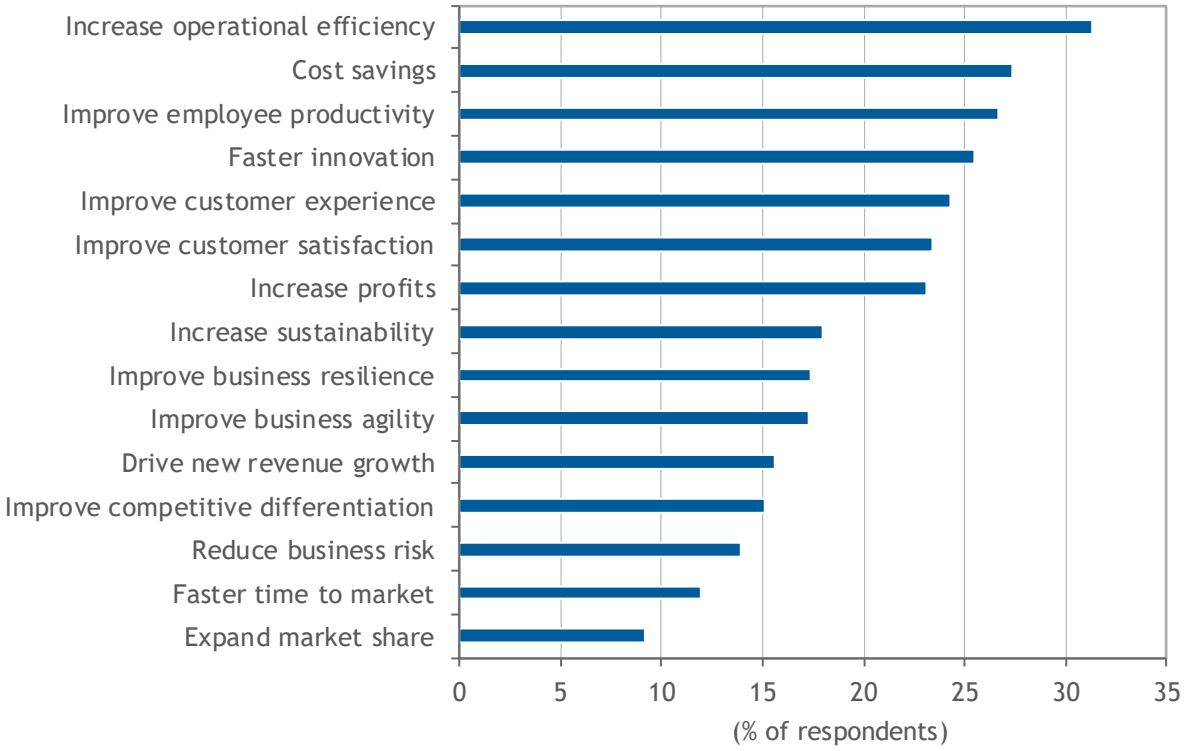
## AT A GLANCE

### KEY STATS

- 83% of IT leaders believe that the use of GenAI models leveraging their own business data will give them a significant competitive advantage (source: IDC's *GenAI ARC Survey,* 2023).
- 87% feel that their organizations are less than fully prepared to take advantage of GenAI capabilities in the next 24 months (source: IDC's *Global AI (Including GenAI) Buyer Sentiment, Adoption, and Business Value Survey*, October 2023).
- 82% of organizations plan to use multimodal foundation models.

FIGURE 1: ***Three Most Important Business Outcomes from GenAI***

**Q** ***Which of the following are the three most important business outcomes that your organization is trying to achieve from AI initiatives?***




*n = 607*

*Source: IDC's Global AI (Including GenAI) Buyer Sentiment, Adoption, and Business Value Survey, October 2023*

## *The Impact of GenAI*

GenAI has taken the world by storm over the past year. Organizations everywhere are working to identify how they can use this transformational technology to achieve a host of desirable outcomes, including improving productivity, increasing sales, and getting products to market quicker. However, adopting GenAI takes careful consideration and planning.

### *Moving from Prototyping to Production*

Reality is coming. Creating scalable production-grade generative AI applications is complex. In the GenAI applications' prototype phase, users often experimented with the largest models available, focusing on accuracy. As organizations move toward production with GenAI, other factors to consider include overall costs, the availability of compute resources, performance (such as latency and response time), energy use, skills, and risk exposure. Allocating and managing resources are essential to maximizing utilization and minimizing potentially high workload costs and GenAI energy use. Using new tools and cultivating new skills will be necessary. Choosing the right foundation models, customizing them for a specific use case, and interconnecting the various application components into a seamless

process can be complex and expensive. Most GenAI applications involving texts and images require specialized hardware. Incorporating the enterprise data to tune the model can be time-consuming to store, gather, manage, and continually update, but that's what will drive competitive advantage.

### *Foundation Model Developers and Providers*

Foundation models are central to any GenAI application. The data to build foundation models may be text (which is used in large language models [LLMs]), image, video, audio, code, or some combination (multimodal). These are the "foundation" that allows the performance of a wide range of tasks across many different use cases. Unmodified foundation models typically perform generalized tasks, such as summarizing content, generating code, or creating product descriptions. However, for complex, high accuracy–dependent applications or applications that need to leverage enterprise data, foundation models require procedural refinement by ingesting additional high-quality data, typically from domain-specific knowledge bases, to train the model for specialized tasks (creating a new, specialized model).

Foundation model providers offer a searchable repository of foundation models that may include proprietary, open source, or third-party models, often referred to as a model library or a model hub. The same foundation model may be available from different model providers. Each provider may offer different benefits to developers. Choice and flexibility are important, as is minimizing the risks to the end user. The choice of foundation model is highly dependent on the use case and often the first task of developing a GenAI application through experimentation and comparison. There are trade-offs between foundation model performance versus computational efficiency, cost versus capabilities, training data size versus environmental impact, and accessibility versus advanced capabilities.

AI platforms support the adaptation of foundation models for more targeted applications. AI platforms are an integrated set of technologies that allow users to develop, test, deploy, and refresh the models. They include multiple components such as data preprocessing tools, machine learning libraries, model hosting and deployment services, scalability and resource management, and monitoring and management capabilities.

## *Approaches to Using GenAI*

Organizations have several approaches to using GenAI in their operations, including:

- Using publicly available commercial foundation models via their APIs or prompts and prompt engineering
- Customizing publicly available commercial foundation models via tuning and retrieval-augmented generation (RAG)
- Developing custom models using available open source models as a starting point
- Building new models from the ground up

Although the first approach served well in GenAI's experimentation phase, IDC's 2023 *GenAI ARC Survey* indicated that 83% of IT leaders believe using GenAI models that leverage their own business data will give them a significant competitive advantage.

The second approach involves a tuning method, RAG, which uses an enterprise's internal data to enable LLMs to perform more efficiently by augmenting the model prompt with additional, more specific information to generate improved

responses. It allows businesses to achieve more customized solutions while maintaining data relevance and is cheaper than developing custom models from open source or building entirely from scratch.

The last two approaches allow organizations to create tailored and function-specific models tuned to the organization's needs and use cases. Their drawbacks are the need for significant data, skills, and budgets to develop and build GenAI models.

### *GenAI Development*

The development and customization of foundation models can be challenging, but model providers have made the process easier to manage than building a foundation model from scratch. Organizations should be aware of certain factors:

- Choosing the right foundation model can be key to developing or customizing a successful GenAI application. Organizations can use many open source foundation models as a starting point. Model providers can offer an extensive model library representing various capabilities that can be tailored to fit the specific use case that the organization is looking for, such as recommendation systems, content summarization, or even content generation. Smaller domain-specific targeted models may be the best option because they require fewer resources and fit the exact desired use case. They can be a better, more cost-effective solution than larger, more general-purpose foundation models available.
- GenAI platforms that offer services such as tuning, grounding, debugging, and root cause analysis for the foundation models in development provide developers with the capabilities they need to create their custom foundation model significantly faster than by using separate tools.
- Deploying custom-developed foundation models is another consideration. Does the model need advanced compute capabilities, such as graphics processing units (GPUs), for inferencing? If so, how much compute is necessary? Many organizations are opting to deploy foundation models through one or more cloud providers because the compute requirements may be higher than what many on-premises organizational datacenters offer.
- Model governance is also a key aspect in developing, deploying, and using foundation models. Organizations need to be able to observe the model in operation, including monitoring, analysis, tuning, optimization, and even retraining. In addition, organizations need to identify and resolve model issues in production, such as model performance and model drift. Some model providers offer tools and capabilities for carrying out these functions.
- Although model operating compliance is still relatively nascent, many governments and standards bodies around the world are developing policies and laws to regulate the use of AI and machine learning to ensure the safety of end users. These rules and regulations will potentially look at issues such as bias, false statements, poor recommendations, data misuse, and copyright infringement to protect users from these models' adverse outputs.

The ideal environment for end-user organizations would be using a single platform for traditional machine learning models and GenAI-based foundation models and supporting both coding and noncoding interfaces. As the GenAI market matures, AI applications will become more complex, incorporating multimodels (proprietary, open source, and third party), multimodal foundation models, and multimethods (a mix of predictive, interpretive, and generative). IDC's *Global AI (Including GenAI) Buyer Sentiment, Adoption, and Business Value Survey* indicated that only 16% of organizations do not plan to use multimodal models in their operations.

In addition, the environment should have tools that help evaluate a model's performance and offer a level of transparency into how the model is working. It should also offer capabilities for tuning and training models, including handling prompt engineering. A key component of this is a vector store, which stores large amounts of data for use in training or tuning. Some environments offer prebuilt starter models for tasks such as content extraction or summarization. These can be a great starting point for organizations that want to quickly ramp up the use of GenAI applications.

### *Responsible AI*

Responsible AI has traditionally focused on explainability, fairness, robustness, transparency, and privacy. Responsible AI in the era of generative AI presents new challenges, especially in the areas of toxicity, hallucinations, and copyright. AI is evolving at breakneck speed, and the EU AI Act is the first law for AI systems in the West. The law takes a risk-based approach to regulating AI, where the obligations for a system are proportionate to the level of risk it poses. Developers of foundation models will have to apply safety checks, data governance measures, and risk mitigations before making their models public. They will also need to ensure that the training data informing their systems does not violate copyright laws.

## *Benefits of AI Platforms*

GenAI is in its infancy. The truth is that building GenAI applications is complicated, and the skills and expertise to successfully tune and integrate foundation models are in high demand and short supply. AI platforms enable foundation model customization through technology, processes, and best practices to automate and operationalize the generative AI life cycle.

AI platforms provide capabilities that include access to foundation models through centralized model hubs, allowing developers to compare multiple models and identify the most relevant foundation models to address specific business use cases. They also provide tools to track experiments, apply various tuning and grounding methods, integrate with new data, develop custom derivative models, debug and optimize performance, and deploy and monitor foundation model–based applications in production.

AI platforms are an enablement layer that supports foundation model tuning and assurance. They are evolving rapidly to support the full GenAI life cycle and differentiate through value-added capabilities. AI platforms that successfully address enterprise requirements, supporting a wide variety of users, models, and data types, will become invaluable tools within the AI toolbox and help AI-based models address broader sets of nuanced business problems.

## *Considering IBM watsonx.ai*

IBM has a rich history in AI, providing services to both partners and end customers. IBM's commitment to openness and transparency is evident through its long-standing approach. Recently, IBM and Meta jointly launched the AI Alliance, a global initiative that brings together over 50 leading industry, academic, research, and government organizations. The AI Alliance aims to advance open innovation and science in AI, prioritizing safety, diversity, and economic opportunity.

The following are some key points about IBM watsonx.ai's software and services:

- **Granite models:** Trust and transparency are at the core of IBM's Granite models, which support tasks such as summarization, insight extraction, classification, and RAG. To ensure trust and performance, IBM is developing

proprietary Granite models that specifically address harmful or objectionable content trained in accordance with IBM's AI ethics code and principles with transparency in data curation and processing. Granite models are typically a smaller, more targeted general-purpose model and are used in applications across industries and functions.

- **Model variety:** IBM offers a wide range of models and model types, including encoder/decoder models for various tasks. Users have ample choice for code, language, and other domains.
- **Client protection:** Differentiated client protection is a hallmark of IBM's approach to developing foundation models. The company stands behind its Granite models and indemnifies clients against third-party intellectual property (IP) claims. Clients can use IBM's Granite models without the need for indemnification, and IBM's IP indemnification liability does not have a cap.
- **Multimodel and multilingual support:** IBM's platform supports multiple models, including IBM developed, open source, and third party. It also offers multilingual (LLM) support beyond English, covering languages such as Japanese, Spanish, Portuguese, French, and German.
- **Multicloud deployment:** IBM's deployment flexibility is based on OpenShift, allowing users to bring models to the data or vice versa. This approach avoids lock-in to a specific cloud provider.
- **Empowering expertise:** IBM's GenAI expertise is accessible through the company's client engineering and consulting teams. With over 1,000 specialized consultants in the IBM Consulting Center of Excellence for generative AI, the company provides proven solutions to more than 100 companies.

In summary, IBM's commitment to openness, trust, and performance positions the company as a leader in the AI landscape, with the AI Alliance serving as a collaborative force for responsible innovation and scientific rigor.

Foundation models provide an unparalleled opportunity for organizations to capitalize on a large, diverse corpora of untapped external data sources that they would likely not be able to collect, compile, or train.

### *Challenges*

Cost and skills continue to be the top inhibitors for organizations trying to scale AI. The generative AI technology landscape is changing rapidly, with new products and services delivered daily from both large tech providers and start-ups. Data and application complexity also plague many organizations. Some 87% feel that their organizations are less than fully prepared to take advantage of GenAI capabilities in the next 24 months.

IBM must continue to add functionality to watsonx.ai to further automate and orchestrate steps within the GenAI life cycle to reduce complexity and increase model velocity.

GenAI is largely driven by the developer community. Increased marketing to individual developers and small businesses should be considered. IBM must provide outreach to organizations that traditionally haven't used IBM products if it wants to get maximum exposure for watsonx.ai.

## Conclusion

Foundation models provide an unparalleled opportunity for organizations to capitalize on a large, diverse corpora of untapped external data sources that they would likely not be able to collect, compile, or train. Combining this data with rich internal data can fuel new custom models that yield unique insights and higher model accuracy for domain-specific tasks.

Excitement about GenAI's potential is high, but the associated challenges, including costs, lack of skills and governance, and concerns about data or IP loss, inhibit organizations from realizing its value. As GenAI foundation models and AI platforms evolve, they must address critical concerns about security, modularity, transparency, ease of use, collaboration, enterprise visibility and governance, and reducing costs and complexity.

Enterprises should choose a technology partner that can help address these challenges while prioritizing their GenAI use cases, evaluating GenAI approaches, investing in internal GenAI skills, and developing policies for responsible GenAI use.

# About the Analysts

***David Schubmehl,*** *Research Vice President, Conversational Artificial Intelligence and Intelligent Knowledge Discovery*

Dave Schubmehl is research vice president for IDC's Conversational Artificial Intelligence (AI) and Intelligent Knowledge Discovery research. His research covers information access and artificial intelligence technologies around conversational AI technologies, including speech AI and text AI, machine translation, embedded knowledge graph creation, intelligent knowledge discovery, information retrieval, unstructured information representation, knowledge representation, deep learning, machine learning, unified access to structured and unstructured information, chatbots and digital assistants, and rich media search in SaaS, cloud, and installed software environments.

***Kathy Lange,*** *Research Director, AI Software*

Kathy Lange is research director for IDC's AI and Automation practice, focused on machine learning life-cycle tools and technologies. Ms. Lange's core research coverage includes machine learning life-cycle technologies and platforms, trends, end-user requirements, business models, use cases, associated regulations, and market sizing for this critical, fast-growing segment. In addition, her research covers AI life-cycle automation, model pipelines, trustworthy AI, and data annotation and labeling services.

## MESSAGE FROM THE SPONSOR

Not all models are created equal, however, all models need governance to deliver responsible AI. Learn how IBM watsonx.governance enables responsible AI.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

**IDC Research, Inc.**
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com

**This publication was produced by IDC Custom Solutions.** The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2024 IDC. Reproduction without written permission is completely forbidden.