IDC PERSPECTIVE

GenAI Engineering in the Enterprise and Why It Matters

Stewart Bond Arnal Dayaratna Kathy Lange Neil Ward-Dutton

EXECUTIVE SNAPSHOT

FIGURE 1

Executive Snapshot: GenAI Engineering in the Enterprise and Why It Matters

Fast-growing awareness of the potential of generative AI (GenAI) is driven by cloud-based services aimed at individuals — but organizations are clear that their needs will not be met by these services alone. However, implementing more sophisticated GenAI use cases involves complexity and risk. To deliver value from GenAI investments that scales across projects and over time, every organization needs a corporate discipline called GenAI Engineering.

Key Takeaways

- The GenAI Engineering discipline is charged with defining and then systematically applying clear business and technology principles and practices to maximize the value of GenAI implementations.
- GenAI Engineering's core purpose is to scale the value realized from GenAI technology and services investments by ensuring that the outputs of investments are treated as enterprise assets.
- GenAI Engineering integrates discipline across three overlapping and interlocking domains: data, AI models, and outcomes.
- GenAI Engineering shapes implementation according to three governing factors: business value, available resources, and business constraints.
- GenAI Engineering must be a collaborative discipline. The key decisions that GenAI Engineering focuses on are heavily influenced by multiple perspectives.

Recommended Actions

- Establish a GenAI Engineering discipline that brings together stakeholders from data science, data engineering, software development, architecture, and business analysis. Focus efforts on integrating perspectives across the data, model, and outcome domains.
- Align your GenAI Engineering practice with your GenAI center of excellence (COE), if you have one or are building one. But do not assume a 1:1 mapping between the activity of your COE and GenAI Engineering.
- Harvest and activate intelligence about data and models to enable data governance for GenAI. Leverage data intelligence to define acceptable data use policies in GenAI engineering practices.
- Democratize access to Gen AI tools within the organization and implement the necessary controls.
- Establish an enterprisewide governance model for model selection and for use case prioritization. Define key performance indicators (KPIs) to measure ROI on GenAI implementations.

Source: IDC, 2024

March 2024, IDC #EUR151907224

SITUATION OVERVIEW

Since the launch of ChatGPT in November 2022, fast-growing awareness of the potential of GenAI, from the boardroom to the shop floor, continues to be driven by cloud-based services aimed at individuals – including GitHub CoPilot, Jasper.ai, Dall-E, Stable Diffusion, Midjourney, and Google Bard, as well as ChatGPT.

Generic Consumer Services Created the GenAI Boom in Enterprises, Too

Just over 12 months from that launch, in January 2024, IDC conducted a large worldwide survey that demonstrated the incredible momentum behind the new technology:

- About 41% of organizations said they already had clear investment plans (across infrastructure, software, and services) in place for the rest of 2024 and 2025. Around 37% are actively exploring, testing with focused proofs of concept, the remaining 22% developing lists of potential use cases.
- Around 67% of organizations said GenAI would have an impact on their business in 2024-2025, and 32% said that GenAI has already disrupted their business to some extent.

In fact, the application of broader AI technology is already widespread among organizations worldwide, particularly within IT automation, cybersecurity, customer service automation, and business operations domains. However, before the introduction of these new cloud-based GenAI services, AI's impact was often somewhat hidden from senior leaders and society at large. The immediately obvious capabilities of GenAI services such as ChatGPT, Stable Diffusion, and GitHub CoPilot – through their ability to converse with individual users in realistic ways and generate high-fidelity, plausible content – are major contributors to the rapid growth in GenAI investment.

GenAI technology has now "broken through" to the extent that market and technology developments are regularly featured in mainstream news reports and articles. As a result, organizations the world over are scrambling to design workable responses to the trend, balancing between embracing potential business opportunities and managing risks. As a senior technology leader from a global industrial manufacturing firm recently shared, "Now, if I don't mention progress on AI in my board reports, I have a target on my back."

Enterprise Needs and Values Point Beyond the Generic

Generic services focused on the needs of individuals, however, are not currently strongly aligned with the values that organizations express. For example, 83% of organizations agree or strongly agree that "GenAI models that leverage our business data will give us a significant advantage."

Nuances shape attitudes of specific industries and geographies, but when IDC asked organizations which qualities they need from GenAI to create business value from the technology, they cite qualities such as simplicity, accuracy, privacy, security, and frugality. For example:

- Around 26% of organizations fear excessive infrastructure costs associated with model training and/or inference, and 26% of organizations also fear excessive costs associated with GenAI-enhanced application software.
- About 30% of organizations are concerned that GenAI jeopardizes control of data and intellectual property. Around 25% are concerned that GenAI use will expose them to brand or regulatory risks, while 24% of organizations are concerned about the accuracy or potential toxicity in the output of GenAI models.

Outside North America, organizations also express significant interest in technology and platform sovereignty. Wave 1 of IDC's *Future Enterprise Resiliency and Spending (FERS) Survey* in January 2024 indicates that about 27% of organizations in Europe and 24% of organizations in Asia/Pacific highlight provider sovereignty as one of the two most important factors in technology selection decisions (compared with just 14% of North American organizations).

GenAI model providers (including hyperscale and public cloud providers and their GenAI research partners) are innovating fast, but the dominant innovation focus for most is on breadth and depth of functionality. For example, vendors are introducing "multimodal" models that can manipulate multiple content types, including text, images, sound, and video.

Enterprises stand at the edge of a new era of productivity and innovation possibilities with GenAI, but enterprise leaders and practitioners must broaden their perspectives to gain value from GenAI technologies. Even setting aside concerns about accuracy, privacy, security, and cost, from a pure functionality point of view, consumer-focused services are also limited. They can deliver value in the context of some enterprise use cases, but not all.

To highlight the gap that now exists between the focus and direction of generic consumer services, and the needs of enterprises wanting to deliver business value from GenAI, IDC predicts:

> By 2026, 60% of enterprises will squander the GenAI opportunity by failing to see GenAI artefacts as assets; failing to weave data, AI, and developer perspectives; and failing to embrace new data value chains.

To avoid this fate, enterprises must invest in a targeted capability called GenAI Engineering.

GenAI Engineering: A Discipline Every Organization Needs

Introducing GenAI Engineering

Despite the simple interfaces presented by services such as ChatGPT, Gemini, and GitHub CoPilot, the GenAI technology space is fast-moving, increasingly complicated and potentially high stakes. Every organization — even organizations expecting to implement GenAI primarily by consuming GenAI-based services from third-party providers — needs to develop a corporate discipline that is charged with bridging strategy and implementation. This discipline is charged with consistently applying key patterns and principles to GenAI implementation projects that are effectively aligned with business outcomes, resources, and constraints.

The GenAI Engineering discipline is charged with defining and then systematically applying clear business and technology principles and practices to maximize the value of GenAI implementations at scale over time.

GenAI Engineering's core purpose is to scale the value realized from GenAI technology and services investments by consistently applying transparent decision making, appropriate implementation patterns, and common tools and platforms to ensure that the outputs of investments are treated as what they are: enterprise assets.

GenAI Engineering integrates business, technology, and data perspectives when guiding GenAI implementation projects. In the same way that architecture engineers take architects' conceptual drawings and turn them into detailed plans for buildings that can be constructed and that will remain standing, GenAI Engineering sifts hype from the art of the possible, and ultimately transforms the art of the possible into delivery of things that are practical.

GenAI Engineering discipline is clearly linked to, and shaped by, GenAI governance.

GenAI Engineering Integrates Discipline Across Three Domains

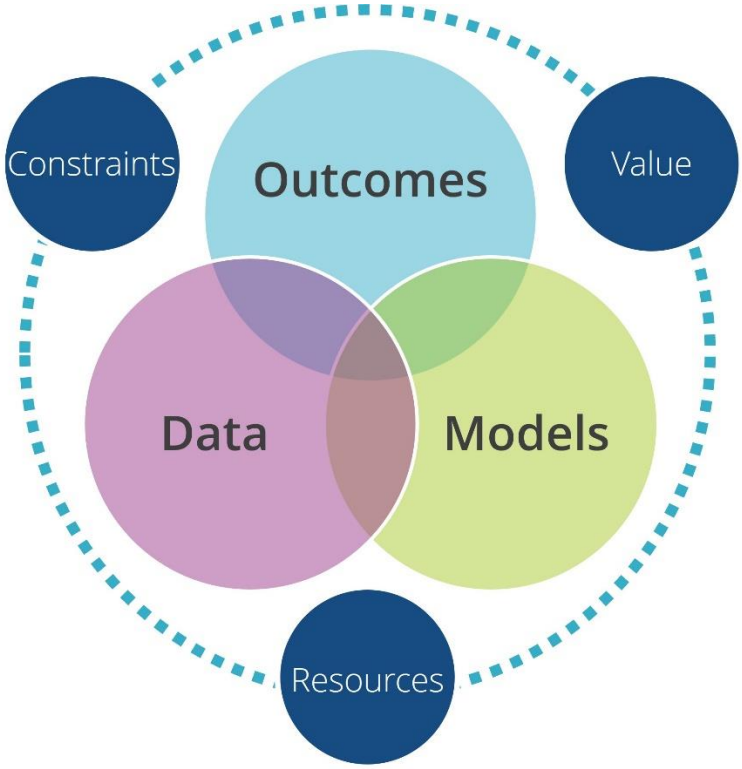
GenAI Engineering does not focus exclusively on issues of GenAI models, despite the prevalence of GenAI models in common technology industry discourse. In fact, GenAI Engineering integrates concepts and decision making between three overlapping and interdependent domains, as Figure 2 shows:

- **Data.** Consumer-focused GenAI services take simple textual input and produce content in response. But in an enterprise context, high-quality corporate data is a foundational element of how GenAI value will be delivered for almost every implementation. In the data domain, the GenAI Engineering practice aims to consistently ask and answer questions such as:
  - Where are we sourcing data used in GenAI technology implementations?
  - What data was used to train the models we are using, and is that data appropriate for the outcomes we need?
  - What data is being used in fine-tuning, retrieval-augmented generation (RAG), and vector embeddings? Is it timely, correct, and is it being used in the right context?
  - Is the corporate data that we are using securely stored and transferred, and is privacy being protected?
- **AI models.** It may seem at first glance like the world of GenAI is dominated by a small handful of suppliers offering public, shared models for rent via APIs to power use cases everywhere, but the reality is a lot more complicated. In the model domain, the GenAI Engineering practice aims to consistently ask and answer questions such as:
  - How are we sourcing models used in GenAI technology implementations?
  - What types and sources of models will work best for the outcomes we want to deliver, based on our specific needs?
  - How do we decide to fine-tune, customize, or otherwise specialize a GenAI model to deliver a given outcome?
  - How are we ensuring that the models we use (whether sourced externally, or developed internally) consistently deliver high-quality output?
- **Outcomes.** Some business functionality can be implemented using a popular, general-purpose GenAI foundation model "off the shelf," but many needs will not be satisfied this way. More foundation model providers are entering the market, offering pretrained models that are trained and designed to perform well in specific situations; and increasingly, enterprises and ISVs are realizing that there is not necessarily a one-to-one mapping between use cases and GenAI models. For many situations, foundation models will need to be augmented and specialized to perform well. In the outcomes domain, GenAI Engineering practice aims to consistently ask and answer questions such as:
  - What kind of implementation approach makes most sense for a given outcome we want?
  - What is the appropriate degree of autonomy we can give to AI components in driving the functionality of an application or system?
  - What infrastructure platforms should we use to train/retrain models?
  - Where should trained models be deployed for inferencing at runtime?

We explore the issues shaping the evolution and the criticality of these three domains later in this report.

FIGURE 2

GenAI Engineering Core Domains and Governing Factors



Source: IDC, 2024

GenAI Engineering Shapes Implementation According to Three Governing Factors

Just like any engineering discipline, a critical feature of GenAI Engineering practice is that it is rooted in the real-world practicality of delivering solutions. In line with this, GenAI Engineering guidance and recommendations are shaped by three interrelated governing factors for each project or program it touches:

- **Value.** Is the outcome in question focused on improving employee productivity, work quality, efficiency, business scalability, revenue, or profit, or some combination of these? How much value can be delivered through a given project, and how does that affect the amount and type of investment that the organization can make?
- **Resources**. What resources are available for delivering a given outcome in terms of data, skills, tools, models, and platforms, and infrastructure? How should the availability of these resources shape sourcing and implementation decisions?
- **Constraints.** Which industry/national regulations may need to be considered? Are there any internal policies which might affect implementation (for example, ethics policies, policies about use of sovereign technologies or partners)? What are the most important risks that need to be considered and minimized?

Over the next three sections, we will explore in more detail the reasons why GenAI Engineering is critical for every enterprise.

GenAI Engineering in the Models Domain: Dealing With a Cambrian Explosion of GenAI Models

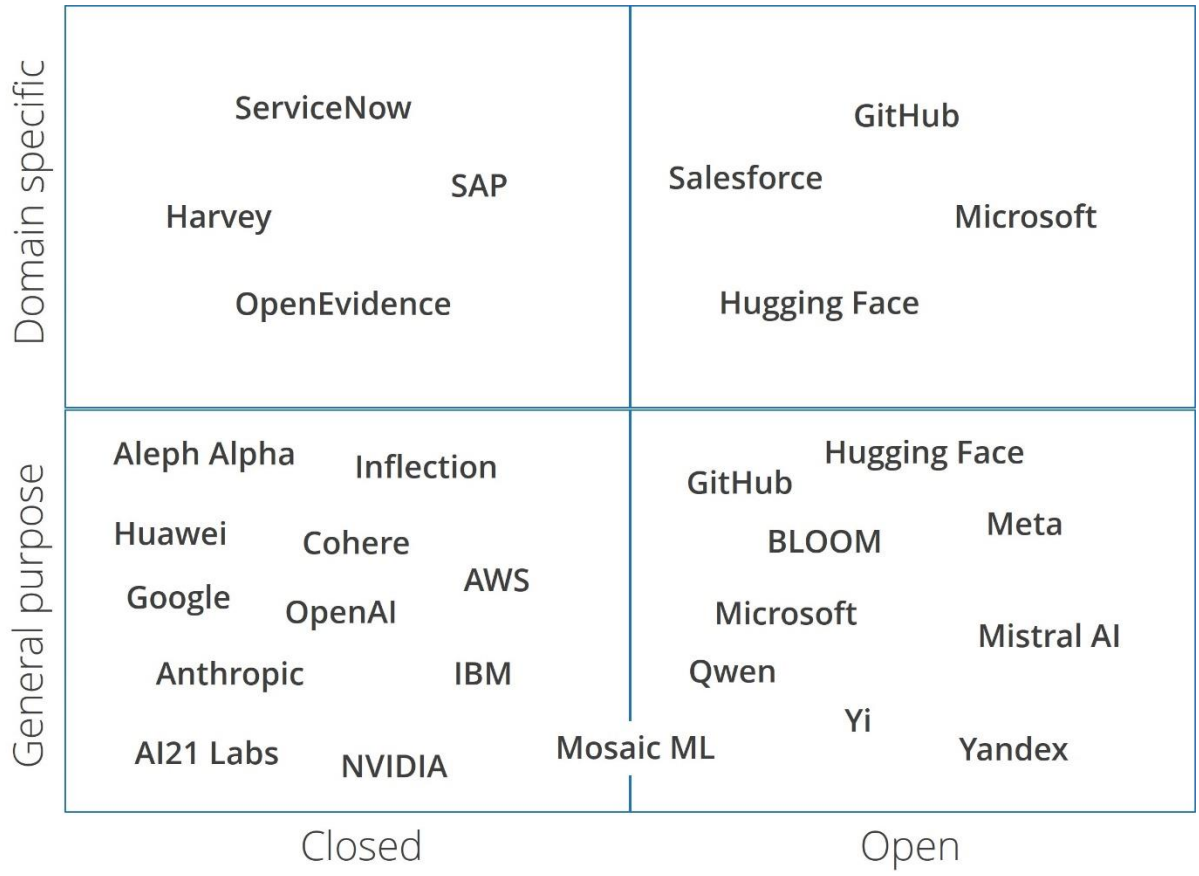
In line with the high levels of public awareness of products such as ChatGPT, Dall-E, and GitHub CoPilot, when thinking about providers of GenAI models, senior enterprise leaders are most likely to refer to OpenAI (and Microsoft, OpenAI's strategic and exclusive platform partner).

However, the Microsoft-OpenAI duo is far from the only source of GenAI model innovation. Fueled by venture capital and corporate investment, competitors have flooded into the space, including:

- GenAI research-focused vendors such as Anthropic, AI21, and Cohere
- Hyperscale public cloud providers (and Microsoft competitors) AWS and Google
- Enterprise technology platform vendors such as IBM, Oracle, ServiceNow, and Adobe
- Sovereignty-focused providers such as Mistral, Aleph Alpha, Qwen, Yi, and Yandex
- Industry-specialized providers such as Harvey (insurance) and OpenEvidence (medicine)
- A vibrant and fast-growing open source model community, with thousands of GenAI-related projects hosted by Hugging Face and GitHub

FIGURE 3

Representative Examples of GenAI Model Providers



Source: IDC, 2024

Open source communities are a particularly energetic vector of innovation: open source projects are quickly evolving model capabilities in terms of model size and efficiency, training, and inferencing cost, explainability, and more.

The Role of AI Platforms

GenAI is in its infancy, and the truth is that building GenAI applications is complicated and that the skills and expertise to successfully tune and integrate foundation models are in high demand and short supply.

GenAI Engineering practitioners can accelerate implementation efforts and lower some requirements for deep technical skills using AI platforms, which enable foundation model customization through technology, processes, and best practices to automate and operationalize the generative AI life cycle.

AI platforms provide capabilities such as access to foundation models through centralized model hubs, enabling developers to compare multiple models and identify the most relevant foundation models to address specific business use cases. Furthermore, AI platforms 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.

As GenAI foundation models and AI platform tools evolve, buyers should expect technology suppliers to address critical concerns about security, modularity, transparency, ease of use, collaboration, enterprise visibility and governance, and reduction of cost and complexity. They will become more performant, and more automated, and address trustworthiness, which will enable rapid implementation of GenAI solutions. These tools will become standard components in an organization's AI/ML toolbox and enable models to address broader sets of nuanced business problems.

See *IDC PlanScape: Unleashing Generative AI Value from the Next Wave of Foundation Models and AI Platforms* (forthcoming) for a deeper dive into how AI platforms help with the management of GenAI models.

GenAI Engineering in the Outcomes Domain: Dealing with a Range of Implementation Choices

To maximize their GenAI-related opportunities, enterprises must explore the opportunities and trade-offs associated with multiple model sourcing and project implementation approaches. These include buying or renting prebuilt services that embed pretrained GenAI capabilities; buying or renting prebuilt GenAI models directly; fine-tuning and integrating third-party GenAI models and services with corporate data and processes; and in some situations, custom-building (or heavily customizing) GenAI models.

Organizations must assess three broad types of GenAI use cases, and each use case type aligns with a particular approach to implementation:

- **Use cases focused on task productivity.** These are basic use cases such as summarizing a report, generating a job description, or generating code. This functionality is being infused into existing applications (e.g., Microsoft 365 Copilot and Duet AI for Google) in which a standalone application addresses a specific request or output that enhances the employee's work task or knowledge.
- **Use cases focused on business function or process improvement**. These use cases will tend to integrate a model (or multiple models) with corporate data for a specific function (e.g., marketing, sales, service, procurement) through grounding and/or model fine-tuning.

Well-established enterprise applications and platforms from vendors such as Salesforce, SAP, Oracle, and Workday are beginning to incorporate GenAI into their offerings to provide these capabilities.

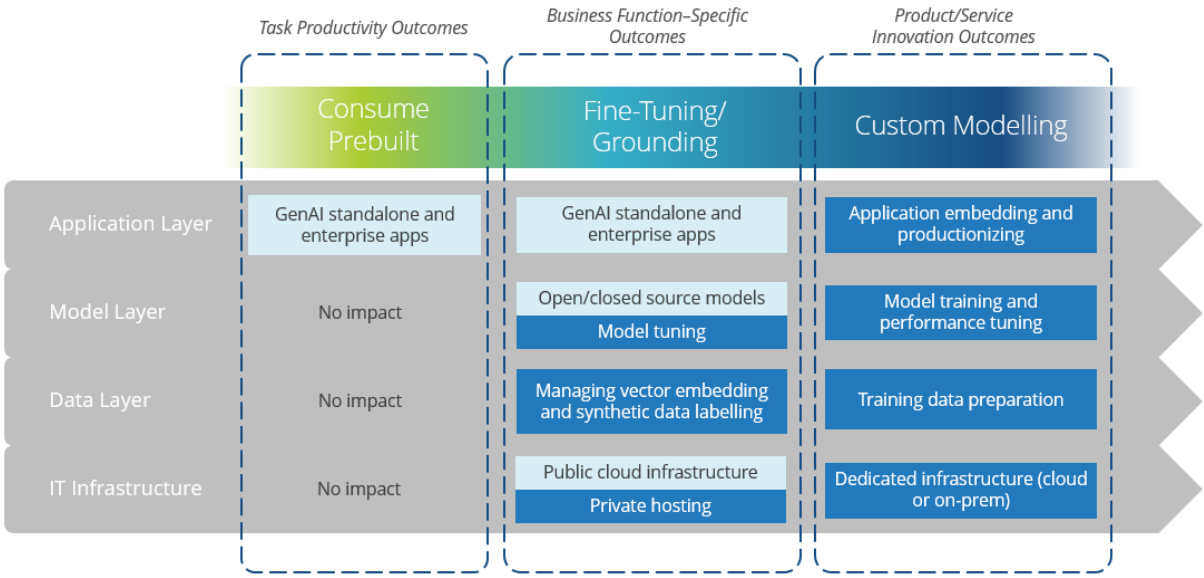
- **Use cases focused on industry-specific product/service innovation**. These advanced use cases will generally require more custom work (and in most cases require building your own GenAI model) to address industry-specific processes or activities. These will impact every industry and every phase from product development (e.g., NVIDIA for drug discovery), operations (e.g., visual inspection in manufacturing), and customer experience (e.g., product recommendations in retail).

Organizations will ultimately need to explore all these use case types and implementation approaches. Organizations report that on average, they expect that in 2024, 34% of investments will be spent on task productivity use cases, 38% of investment will be spent on business function or process improvement use cases, and 28% of investment will be spent on product/service innovation in 2024 (based on wave 1 of IDC's *FERS Survey* in January 2024).

Figure 4 provides a high-level overview of the use case implementation choices that enterprises face as they consider different types of GenAI use cases.

FIGURE 4

Implementation Choices Are Complex and Must Be Managed



Source: IDC, 2024

For use cases that are best implemented through model fine-tuning/grounding or custom modelling, GenAI engineers will also need to consider infrastructure implications that can affect timing, costs, and data governance. See *Build-Versus-Buy Decision Making: Optimizing AI-Ready Infrastructure ROI* (forthcoming) for a deeper dive into AI ready infrastructure developments to watch.

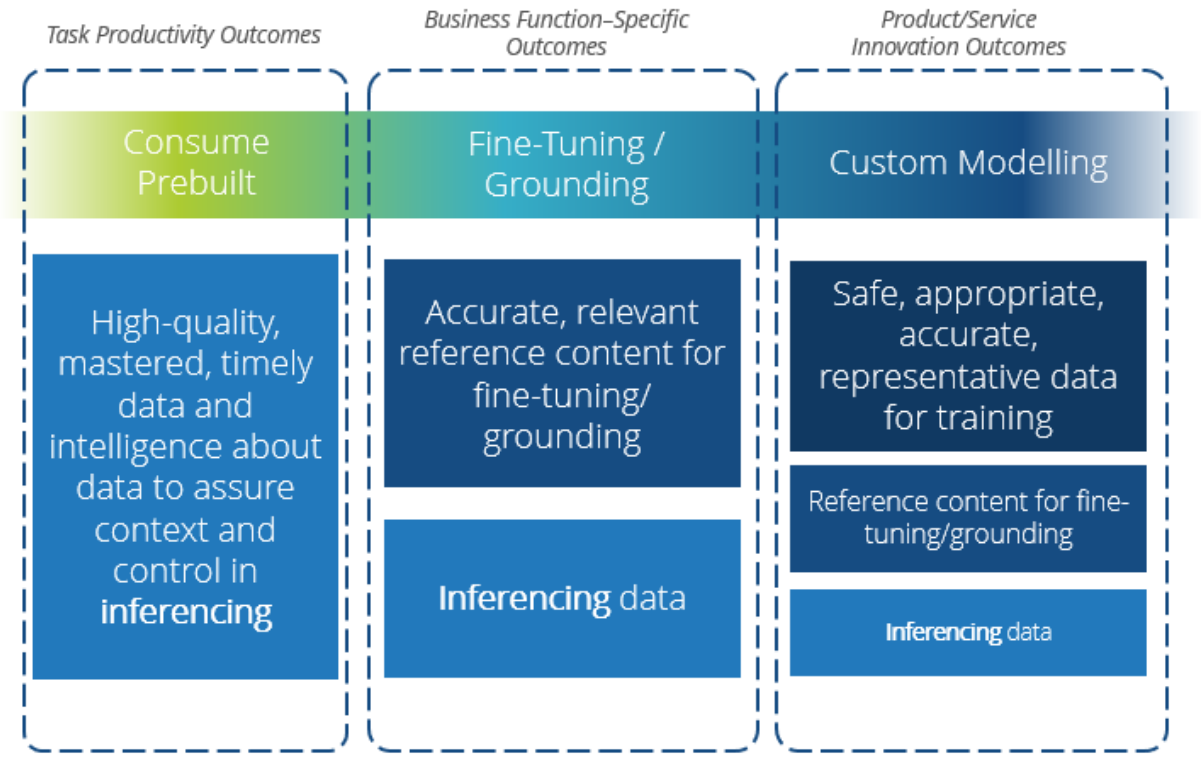
GenAI Engineering in the Data Domain: Reflecting Data's Value

As Figure 5 shows, no matter what type of GenAI use case is being considered, and no matter the resulting use case implementation choice, the availability of high-quality corporate data is a major contributor to effective implementation.

- In task productivity use cases – even where a prebuilt GenAI model is being consumed "off the shelf" through a public API or a consumer-friendly hosted service – at runtime, corporate data, content and/or knowledge will very often provide critical context that guides the service to provide targeted value (via a technique sometimes known as "prompt tuning"). For example, in a hypothetical service that generates marketing copy for an event, the user will need to provide clear information to the service about the type of event, the event title, the audience, the location, the event format, and so on. A hypothetical service that provides medical guidance to healthcare professionals based on a domain-specific GenAI model must be provided with reliable information about a patient's symptoms and clinical history to provide relevant output.
- In use cases focused on improving business function or process performance, GenAI models at the heart of use case implementations will need to be "grounded" to incorporate data, content, or knowledge; and/or be fine-tuned through additional training; this improves their performance in a specialized domain. For example, in a hypothetical service that provides easy-to-use search and retrieval functionality across a large customer service knowledge base, the GenAI model at the core of the service must be integrated with that knowledge base to ensure it only answers questions using information from that knowledge base.
- Use cases focused on industry-specific product or service innovation will commonly rely on custom-built or extensively customized GenAI models that are highly specialized for the domain in question. In these use cases, GenAI models will need to be trained from scratch, or incrementally trained for specialization, using high-quality training data.

FIGURE 5

Data Value and Model Value Are Always Intertwined



Source: IDC, 2024

The Role of Data Governance

Data used in GenAI projects must be controlled and governed to address organizations' concerns about the use of internal data with external models, leaks of sensitive personal information and/or corporate intellectual property, brand or regulatory risks, and concerns about accuracy, specificity, and toxicity of GenAI output. Nearly half (40%) of the 881 organizations surveyed worldwide by IDC in January 2024 believe that implementing data sharing and operations practices that ensure data integrity is the most important process and policy for ensuring success with GenAI. Unfortunately, only a third or less of that same population believe that they can ensure high-quality data, strictly control sensitive data, understand what data has intellectual property ownership issues, and track and control their internal data use with external GenAI models.

Enabling data governance for generative AI is not a technology-only solution; it requires a strategic vision and capabilities across multiple dimensions of data and model intelligence, technology, processes, and people. A strategic vision needs to have clear objectives, a set of ethical principles, data ownership and use policies, a business case with associated capital and operating budgets, and established performance metrics that promote transparency and accountability.

Enabling data governance for generative AI requires organizations to gain intelligence about the data it stores, manages, and uses within its data estate as well as combine it with model intelligence to ensure data is being used appropriately with appropriate models. Data intelligence leverages business, technical, relational, and operational metadata to provide transparency of data profiles, classification, quality, location, lineage, and context. This provides people, processes, and technology with trustworthy and reliable data. Model intelligence is the metadata associated with

the model, such as model definition, intended uses and limitations, classification, model function, modality, libraries, sensitivity, security guardrails, versions, training data used, licensing, and so on.

See *IDC PlanScape: Enabling Data Governance for Generative AI* (forthcoming) for a deeper dive into the importance of data governance for GenAI implementation.

Data Value Chains Are Critical for GenAI

To illustrate the mechanics of processing corporate data to maximize its value for GenAI use case implementation, we introduce the concept of the GenAI data value chain.

The GenAI data value chain encapsulates the entire life cycle of data management processes that transform raw data into insightful, valuable outcomes in GenAI contexts. This value chain is essential for:

- **Foundation model development.** Involving data collection (gathering diverse, unstructured data), preparation (chunking and tokenization), training, evaluation, continuous learning, and adherence to ethics and compliance
- **Foundation model transformation:** Focused on model adaptation (using specialized data sets) through techniques such as fine-tuning and RAG
- **Bespoke application development.** Designing and integrating applications that leverage foundation models for specific use cases

For foundation model development, key data value chain elements are:

- **Chunking.** Facilitates efficient handling and processing of large data sets and enables parallel processing and memory overhead reduction
- **Tokenization.** Prepares data sets for foundation models by breaking down text into manageable tokens that enhance the ability of models to process inputs and generate outputs

For foundation model transformation, key value chain elements are:

- **Vector databases.** This element plays a critical role when transforming foundation models for specific applications or tasks. These databases are used for efficiently storing and retrieving the embeddings that represent the knowledge the model has learned. This is particularly important in tasks that require quick access to vast amounts of high-dimensional data, such as semantic search or personalized content recommendation.
- **Embeddings.** In the transformation phase, the model's ability to create accurate and contextually relevant embeddings is key. These embeddings, which are high-dimensional representations of words, phrases, or other data types, are highly curated to enable the model to deliver specialized outputs for specific tasks and use cases.

For the development of bespoke applications leveraging foundation models, key value chain elements are:

- **Vector databases and embeddings.** In bespoke applications, vector databases are used to manage the embeddings generated by the foundation model. This is critical in applications involving real-time data processing such as chatbots, where quick retrieval of relevant information based on user queries is essential.
- **Chunking and tokenization.** These processes are vital in pre-processing data in bespoke applications. For instance, in a custom language processing tool, tokenization ensures that user input is correctly interpreted by the model. Similarly, chunking can be used to process large volumes of data or long documents in a manageable way.

The data value chain for GenAI enables the realization of the following benefits with respect to both foundation models and generative AI-based applications:

- **Quality improvement.** Enhancing accuracy and relevance of content, reducing errors, and accurately replicating real-world scenarios
- **Performance enhancement:** Increasing efficiency in task processing, accelerating learning, and improving real-world application generalization
- **Latency reduction.** Using high-quality data for more efficient algorithms, speeding up training and inference, and reducing computational overhead.
- **Security improvements.** Boosting the model's ability to detect biases, enhancing resistance to threats, and minimizing exploitation risks
- **Cost of ownership reduction.** Reducing the need for retraining, minimizing deployment risks, and extending model relevance and longevity.

See *IDC PlanScape: Understanding the Generative AI Data Value Chain: Key Concepts and Considerations* (forthcoming) for a deeper dive into the importance of data value chains for GenAI implementation.

GenAI Engineering: A Collaborative Discipline

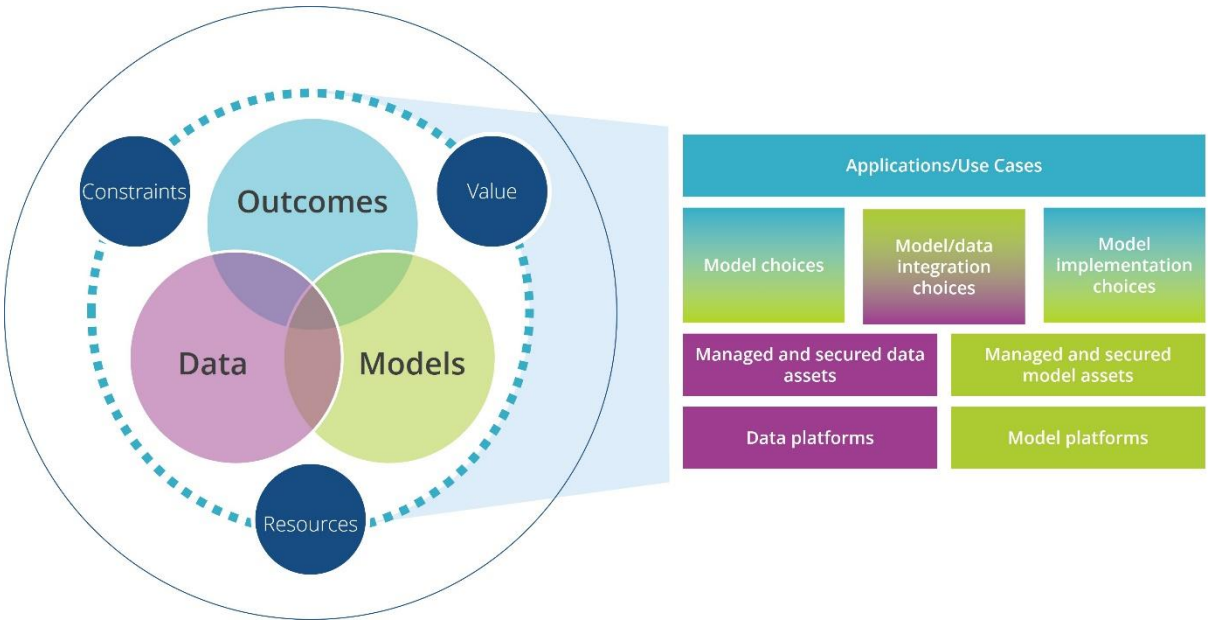
Enterprise excitement about the potential of GenAI is driven by novel technology — but people will be the key to ultimate GenAI business success. Organizations report that investment in skills is high on the agenda: 30% of planned GenAI investment will be in people and skills, according to IDC's *GenAI Awareness, Readiness, and Commitment (ARC) Survey* in August 2023. Moreover, 36% of organizations are creating mandatory GenAI awareness and acceptable use training, according to wave 1 of IDC's *FERS Survey* in January 2024.

Focusing on developing skills is critical, because organizations report that readiness to take advantage of GenAI opportunities within key communities is only moderate at best. Wave 1 of IDC's *FERS Survey* indicates that only 24% of IT, data and FinOps staff, 21% of DevOps and data science staff, and 26% of security, risk and compliance staff, and 25% of IT support staff are judged to be completely ready to support GenAI plans and investments.

GenAI Engineering is a discipline that drives collaboration between all these stakeholder groups to create success. The three domains that are core to GenAI Engineering — data, models, and business outcomes — are traditionally "owned" by distinct teams that do not always work together. But collaboration is essential, because the key decisions that GenAI Engineering focuses on must be heavily influenced by multiple perspectives, as Figure 6 shows.

FIGURE 6

GenAI Engineering Is a Collaborative Discipline



Source: IDC, 2024

For example:

- Decisions about model choices must be informed by the needs of individual projects and outcomes; and equally, outcome practicality will be informed by decisions about model choice and sourcing.
- Decisions about how to use corporate data to fine-tune, customize, and integrate models must be informed by the availability and quality of available data. Equally important, data acquisition and management strategies should be informed by demand based on new GenAI use cases and model choices.
- Decisions about how and where to implement models (for training, fine-tuning, integration, and inferencing at runtime) must be informed by the needs of individual projects and outcomes. Equally, outcome practicality will be informed by policies that dictate infrastructure and platform investment priorities.

Key Roles Involved in GenAI Engineering

A wide range of personas must be prepared to work together to implement GenAI Engineering practice and advance its use within projects and programs. Key personas span both technical and non-technical roles, and include:

- **Chief information security officers (CISOs).** CISOs in generative AI safeguard data integrity and enforce security policies. They mitigate risks to ensure compliance with data protection laws and ethical standards. Their focus on preventing data breaches and establishing a secure AI infrastructure is critical to maintaining the integrity and trustworthiness of AI systems, applications, and infrastructures.
- **Chief data officers (CDOs).** These C-levels oversee data acquisition, governance, and quality and ensure alignment with organizational goals and compliance with ethical and regulatory standards. They are also responsible for enhancing the effectiveness of AI

models by providing diverse, accurate data. Furthermore, CDOs foster a data-centric culture within organizations that is crucial for the successful implementation and adoption of generative AI technologies.

- **Data engineers.** Data engineers design and maintain the data infrastructure vital for generative AI. They also manage data pipelines and ensure the quality and accessibility of data for AI model training. Their role involves optimizing data processes for efficient storage and retrieval, crucial for handling the vast data volumes in AI systems. Additionally, they take responsibility for scaling infrastructure to meet evolving AI demands.
- **Data scientists.** These are key in developing and refining generative AI models. They select algorithms, train models, and extract insights from data. They are responsible for model training as well as the transformation of models using techniques such as fine-tuning and retrieval augmented generation.
- **Prompt engineers.** These bring data science, application development, and data engineering skills to bear to assemble prompts at time of inference; they also interpret results returned by the foundation model. This role is still emerging, and there is a possibility that the activities performed will be part of the three roles it borrows from, rather than it be an individual and unique role in the organization.
- **Non-technical domain experts.** These are in many cases key providers of domain expertise required to implement prompt engineering and tuning activities that can help to specialize outputs from foundation models for particular GenAI implementations.
- **Database administrators.** These manage the storage infrastructures that are central to generative AI. They ensure secure, efficient data storage and retrieval, optimizing database performance for AI processing. Their role includes implementing data recovery strategies, maintaining data integrity, and scaling databases to support the growing needs of AI applications.
- **Developers.** These build and integrate AI-driven applications. They translate complex AI models into practical software solutions and customize them for specific purposes. Like data scientists, developers may take responsibility for the transformation of foundation models by means of methods such as fine-tuning and retrieval augmented generation.

GenAI Engineering and the Role of a COE

With 32% of organizations worldwide prioritizing the creation of a GenAI center of excellence, what is the relationship between a GenAI COE and GenAI Engineering?

The answer is simple: a GenAI COE is a natural "home" for a GenAI Engineering capability. However, there are two important wrinkles to bear in mind:

- Common practice in COE setups is for COE resources to take responsibility for more than core engineering or best practices sharing activities. For example, COEs in adjacent spaces (enterprise automation, analytics, and so on) commonly play an "internal marketing" role, sharing success stories and helping teams make business cases for investment. If you plan to set up a GenAI COE, you should not limit its activities to GenAI Engineering.
- GenAI Engineering is a discipline, built around a set of practices — it is not a team. Good practice is for disciplines such as GenAI Engineering to be operationalized in ways that reflect broader business-technology operating models. This means that for many enterprises, a "hub and spoke" hybrid community model will be optimal (where a core team focuses on setting policies and establishing good practices, but multiple teams, federated across business units, brands or departments, implement those policies and practices)

ADVICE FOR THE TECHNOLOGY BUYER

GenAI Engineering is a critical discipline for every enterprise that wants to deliver sustainable, scalable value from GenAI technology investments. It brings a systematic approach to the implementation of principles and practices that drive consistent decision making throughout the life cycle of every GenAI project and program — right from the early stages of use case selection and prioritization, through solution design, model selection, architecture, implementation and ongoing monitoring and management.

The GenAI Engineering discipline should direct a phased approach to delivering GenAI value, starting with a solid data foundation. As use cases are explored and projects are executed, the discipline should focus on creating a standardized but modular software stack that can be used to accelerate and govern implementation across initiatives with diverse requirements.

People and Practice

GenAI Engineering must be collaborative. Establish a GenAI Engineering discipline that brings together stakeholders from data science, data engineering, software development, architecture, and business analysis. Focus efforts on integrating perspectives across data, model, and outcome domains. Center your work on establishing policies and practices that help teams consistently answer key implementation questions, wherever in the business they may be.

Align your GenAI Engineering practice with your GenAI COE, if you have one or are building one. But do not assume a 1:1 mapping between the activity of your COE and GenAI Engineering.

The Importance of Intelligence

Harvest and activate intelligence about data and models to enable data governance for GenAI. Risks of GenAI models producing hallucinations and incorrect output as well as risks of leaking personal or corporate sensitive information run high if your organization does not know where the highest-quality, most recent, and business-relevant data and content exists within the enterprise data estate, including personal or corporate sensitivity classifications. The risks of using a model that is inappropriate within the context in which it is being used also run high if there is not enough information about where the model came from, what it does, how sensitive it is (based on classification), and who or what is responsible and accountable for it.

Leverage data intelligence to define acceptable data use policies in GenAI engineering practices. Organizations without these policies will not be able to use GenAI responsibly.

GenAI Engineering and Governance Are Two Sides of the Same Coin

Democratize access to Gen AI tools within the organization, while implementing the necessary controls. Offer flexibility for the rapid development of custom solutions. Address data latency issues to optimize UX and increase productivity.

Establish an enterprisewide governance model for model selection and for project prioritization. Define key performance indicators (KPIs) to measure ROI on implementations.

Enabling data governance for GenAI requires people that are working within the constraints and guardrails of process and policy and have clear accountabilities and responsibilities. Leverage responsible, accountable, consulted, and informed (RACI) matrices to clearly define roles and responsibilities among the CDO, CISO, CPO, data architects, data analysts, data engineers, prompt engineers, application developers.

LEARN MORE

Related Research

- *IDC Market Glance: Generative AI Technologies for Model Build and Orchestration, Application Development, and Deployment and Data Management, 4Q23* (IDC #US51422323, December 2023)
- *How Generative AI is Transforming Developers, Development, and the Developer Experience* (IDC #US51317123, October 2023)
- *Artificial Intelligence in Data Intelligence and Integration Software for the Future of Enterprise Intelligence* (IDC #US50923223, September 2023)
- *Generative AI: The Path to Impact* (IDC #EUR151153223, August 2023)

Synopsis

This IDC Perspective discusses the importance of GenAI Engineering in enterprises, highlighting its potential to drive business growth and innovation. It provides an in-depth analysis of the current state of GenAI adoption, the challenges organizations face, and the need for a systematic approach to maximize the value of GenAI implementations. This content also emphasizes the role of data governance, collaboration, and a phased approach in successful GenAI implementations. It concludes with advice for technology buyers on establishing a GenAI Engineering discipline and leveraging data intelligence.

"Every organization needs to develop a corporate discipline that is charged with bridging strategy and implementation; this discipline must also be charged with consistently applying key patterns and principles to GenAI implementation projects to ensure they deliver sustainable value," said Neil Ward-Dutton, VP AI, Automation and Analytics Europe, IDC. "That discipline is GenAI Engineering."

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

IDC U.K.

IDC UK
5th Floor, Ealing Cross,
85 Uxbridge Road
London
W5 5TH, United Kingdom
44.208.987.7100
Twitter: @IDC
blogs.idc.com
www.idc.com