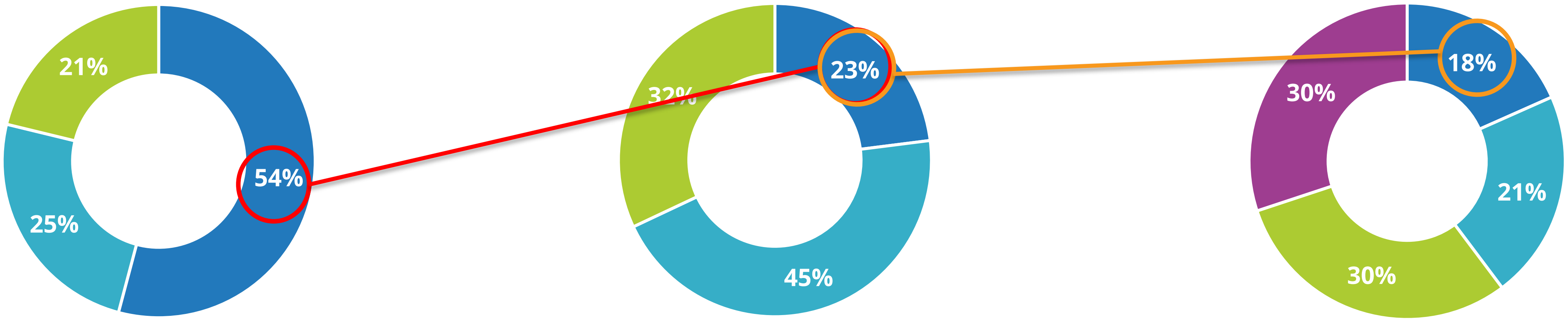
# GenAI for Competitive Advantage

## Build Versus Buy

Daniel M. Saroff
Group Vice President, Consulting and Research

# What's your organization's current approach to Generative AI?

**March 2023** | **June 2023** | **November 2023**




**March 2023**
- Not doing anything yet.
- Initial exploration of potential use cases.
- Investing in GenAI.

**June 2023**
- Not doing anything yet.
- Initial exploration of potential use cases.
- Investing in GenAI.

**November 2023**
- Not doing anything yet.
- Developing potential uses cases. Not yet investing.
- Initial models and potential use cases testing. No set spending plan.
- Investing significantly in Generative AI.

Source: Future Enterprise Resiliency & Spending Survey Wave 2, IDC, March 2023; Future Enterprise Resiliency and Spending Survey Wave June 20223
Source: Future Enterprise Resiliency & Spending Survey Wave 10, IDC, November 2023, N=889, NA: 369, WE: 220, AP: 300

# Allocation of New GenAI Investments by Enterprises Planning to Increase Overall IT spend in 2024

**You indicated that Generative AI will be an area of increased IT spending in 2024, regardless of overall IT spending plans. What is your best estimate of how this increased investment will be allocated across the following areas?**

| | Worldwide | North America |
|---|---|---|
| Dedicated Infrastructure (i.e. AI-specific server and storage hardware in on-premise/colocation datacenters as well as remote edge locations) | 24% | 27% |
| Public Cloud-infrastructure (i.e. compute and storage cloud service) | 21% | 21% |
| External AI Platforms (i.e. software/dev tools, AI models, external data sources) | 21% | 22% |
| Internal data scientists, developers, Data Ops employees, and other IT and line of business employees | 17% | 15% |
| External Services Providers (i.e. consulting, custom development, managed services etc.) | 16% | 14% |

Source: Future Enterprise Resiliency & Spending Survey Wave 10, IDC, November 2023, N=534, NA: 219, WE: 160, AP: 155

# Leading GenAI Strategic Technology Partners Worldwide

## Strategic Generative AI technology partners in the next 18 months

**41%** Cloud Providers

aws
Google Cloud Platform
Alibaba Cloud
Azure
IBM Cloud

**33%** IT Consulting Providers

accenture
Deloitte.
pwc

**30%** GenAI Tool Providers

OpenAI
ANTHROPIC
AI21labs
co:here

**Source:** GenAI ARC Survey, IDC, August 2023, WW N=1,363. Data weighted by IT spend

# Over 66% of Organizations Planning to Decrease or Freeze Overall IT Spending in 2024 **Still** Plan Increase Investments in GenAI Through Either New Funding or Cuts in Other Areas

| | Global | N. America |
|---|---|---|
| We will make significant new investments in GenAI separate from current IT budget plans | 35.7 | 34.0 |
| We will fund any new GenAI spending by more aggressively cutting spending in other areas | 31.7 | 27.6 |
| GenAI will NOT have any influence on our IT budget plans | 32.6 | 38.4 |

Source: Future Enterprise Resiliency & Spending Survey Wave 10, IDC, November 2023, N=129, 400

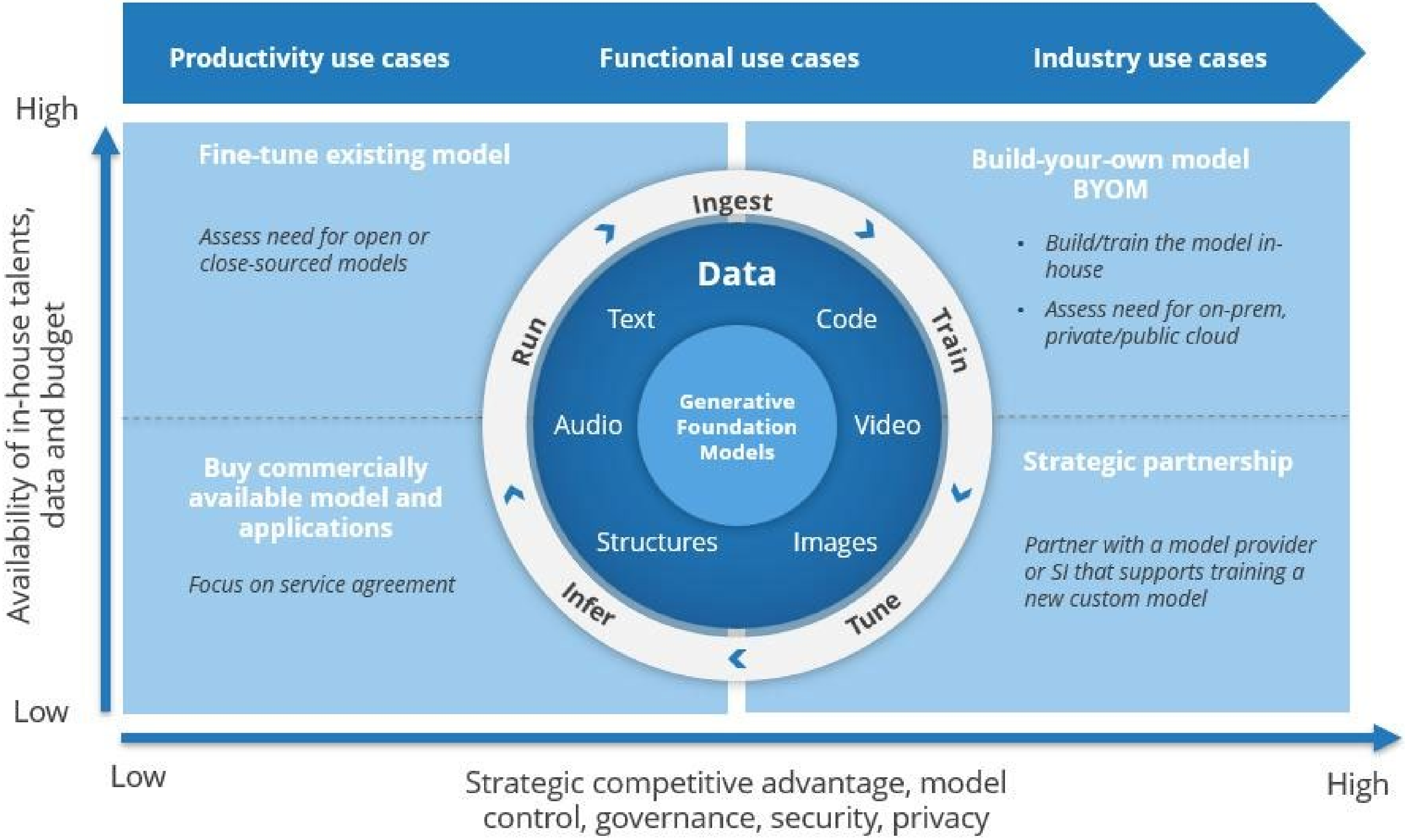
# Build Vs. Buy

# Build and Buy and Configure and Create

# Build Vs. Buy

## A false dichotomy – really configure and create.

**Organizational GenAI goals will be influenced by their individual business and technology circumstances. This will result in a mix of build and buy approaches across different use cases.**




Source: IDC, Generative AI: Approaches for Competitive Advantage, #AP49701623

# Typical Use Cases

## Three broad categories of GenAI use cases deliver increasing business value to the extent they leverage enterprises' native data.

**Use case categories range from productivity to functional to industry. They provide increasing differentiation while necessitating increased levels of control over model architecture, security, data privacy, and governance**

| Use Cases Categories | Business Impact | Drivers | Possible Implementation Approach | Use Case Example |
|---|---|---|---|---|
| Productivity use case | • Increases task productivity<br>• Drives operational efficiencies | • Limited talent in-house<br>• Limited budget<br>• Low risk appetite<br>• Early adoption<br>• Limited institutional data | • Commercial applications with embedded GenAI<br>• Native GenAI standalone applications (e.g., Microsoft Copilot, Jasper AI, and so forth | • Summarizing a report<br>• Generating a job description<br>• Generating code in Java/C#/Python |
| Functional use case | • Increased functional effectiveness<br>• Contextualized experiences | • Well-harmonized institutional data<br>• Available talent in-house<br>• Budget available<br>• Medium risk appetite | • Fine-tuning open source models<br>• Fine-tuning models available from model hubs and AI platforms<br>• Retrieval-augmented generation (RAG) | • Provision of hyper-personalized sales and marketing<br>• Hyper-personalized wealth and investments knowledge management<br>• Generative product design and prototyping |
| Industry use case | • Enable new digital business models, products, and services<br>• Industry-specific competitive moats | • Stringent regulatory and privacy requirements.<br>• Talent in-house or partner<br>• Quality and quantity institutional data | • Fine-tuning third-party or industry models<br>• Custom-built models (BYOM) | • Generative drug discovery in life science<br>• Generative material design for manufacturing |

Increasing Maturity

Source: IDC, Generative AI: Approaches for Competitive Advantage, #AP49701623

# Build Approaches Elaborated

## Build approaches deliver focus and differentiation, but IDC also see foundation models evolving to support more specialized industry and business process–specific use cases.

**The build approach can self-implemented or through a partnership with a vendor or service provider if inhouse talent is unavailable.**

| Approach | Description | Associated Life-Cycle Steps and Key Considerations | Sample Key Technology/Tools (Vendor) |
|---|---|---|---|
| Prompt engineering | Use prompting techniques to influence generated outputs of a pretrained model. | **Run:**<br>• Prompt engineering skills<br>• Open source or commercial models<br>• On-premises model or API access<br>• Modalities | • Open source models: GPT-3 (OpenAI), Hugging Face Transformers, BERT, ELECTRA (Google)<br>• Commercial models: GPT-3.5 Turbo (OpenAI), IBM Watson (IBM), DeepL |
| RAG | Combines a domain-specific collection of documents with pretrained models to contextualize output, without touching the LLM | **Ingest:**<br>• Large volume and quality of domain-specific documents<br>**Infer:**<br>• Vector database selection<br>• Embedding model selection | • Vector databases: Vespa, NGT, Pinecone, Milvus, Elasticsearch, Faiss (Meta), Weaviate<br>• Embedding models: ELMo, SBERT, FastText, Word2Vec, GloVe |
| Fine-tuning | Using domain data and human supervision to tune pretrained models to improve domain-specific performance | **Ingest:**<br>• Training data quality and governance<br>**Tune:**<br>• Availability of matching foundation models<br>• ML expertise for parameter tuning and MLOps capability<br>• Labor and skills for ongoing RLHF<br>• Compute capacity and cost<br>• Regular re-tuning for data refresh | • ML framework: TensorFlow, PyTorch<br>• LLM orchestration framework — LangChain, LlamaIndex<br>• Specialized generative model libraries: Hugging Face Transformers, StyleGAN (Nvidia)<br>• Model optimization platforms: Optuna, HyperOpt, Ray Tune<br>• Cloud-based ML platforms: Vertex AI (Google), AWS SageMaker, Runway ML, Einstein Platform (Salesforce), WatsonX (IBM) |
| Custom modeling (Specialized / Industry) | Foundation models built on large corpus of proprietary data | **Ingest**:<br>• Large volume of quality training data and governance<br>**Train and tune:**<br>• ML expertise for parameter tuning and AI/MLOps capability<br>• Experienced data science and ML team<br>• Dedicated compute and storage capacity | • Compute and storage capacity: Nvidia, AWS, AMD, Intel, Google Cloud, Microsoft Azure<br>• Cloud-based AI platforms: DGX Cloud (Nvidia), Azure AI (Microsoft), WatsonX (IBM), Vertex AI (Google)<br>• Data platforms: Oracle, Snowflake, Cloudera, Databricks |

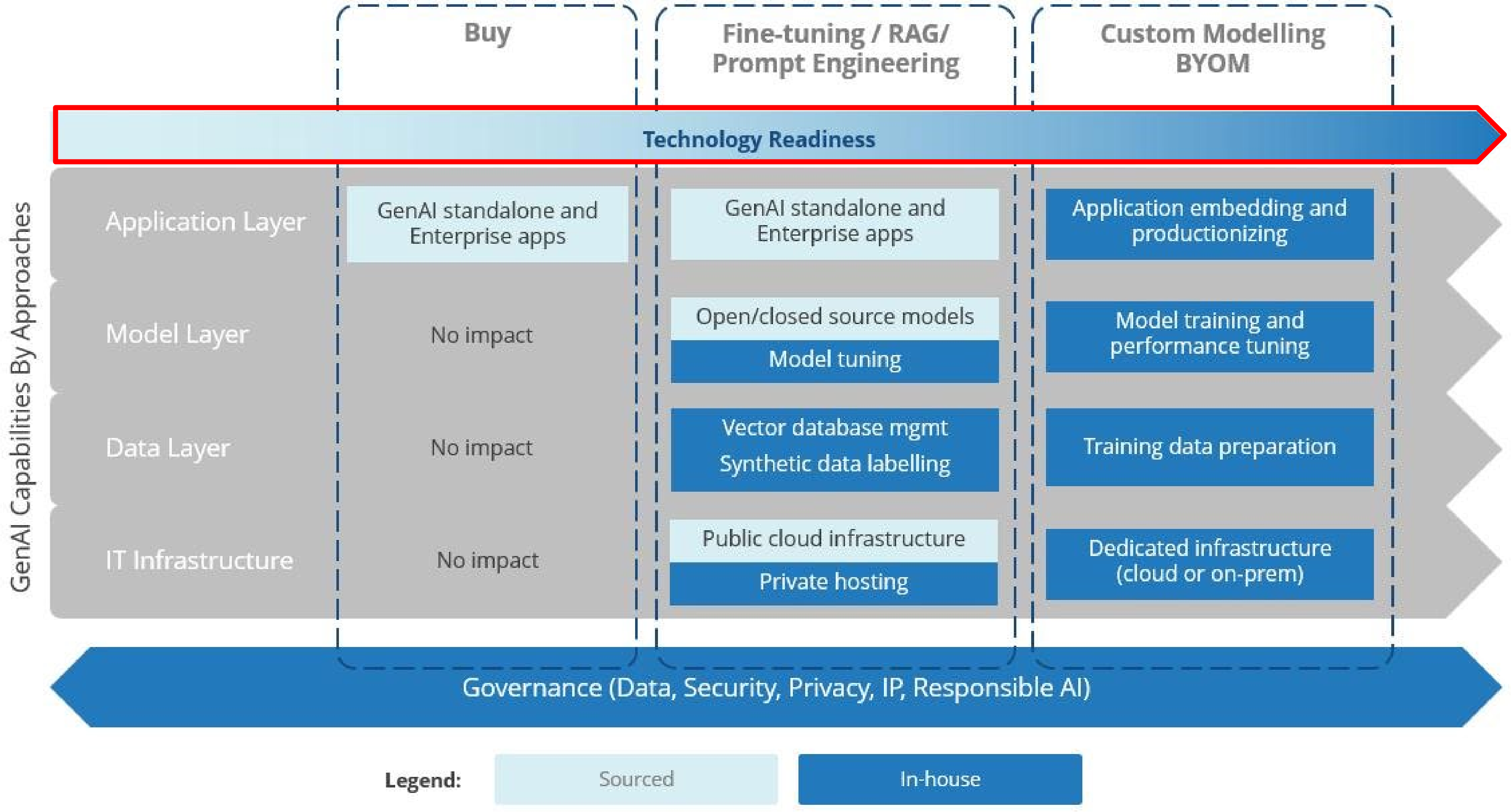
Note: Technology examples are not necessarily exclusive to the stated GenAI approach and are not an endorsement. They are used as examples.

Source: IDC, Generative AI: Approaches for Competitive Advantage, #AP49701623

# Technology Readiness for GenAI Approaches

Building the GenAI capability can range from fairly lightweight to an extremely involved process impacting multiple layers of the technology stack.

**The technological readiness of the organization helps determine build versus buy decisions. For example, the greater the readiness implies better controls and governance — a key enabler for GenAI.**




Source: IDC, Generative AI: Approaches for Competitive Advantage, #AP49701623




# Takeaways

# Some Takeaways

- Understand your near-term and long-term foundational GenAI readiness
- Get your data house in order
- Start with understood, good, recent, and relevant data
- Build for the future with a platform and integration in mind
- Understand your genAI business priorities – productive, functional, industry
- It's not build versus buy; it's both. Choose what makes sense for each component of your strategy

Daniel Saroff
Group Vice President Consulting and Research
dsaroff@idc.com
+1.508.935.4462

IDC.com linkedin.com/company/idc blogs.idc.com