CIO 2004 - 2024 ASSOCIATION OF CANADA
20TH ANNIVERSARY

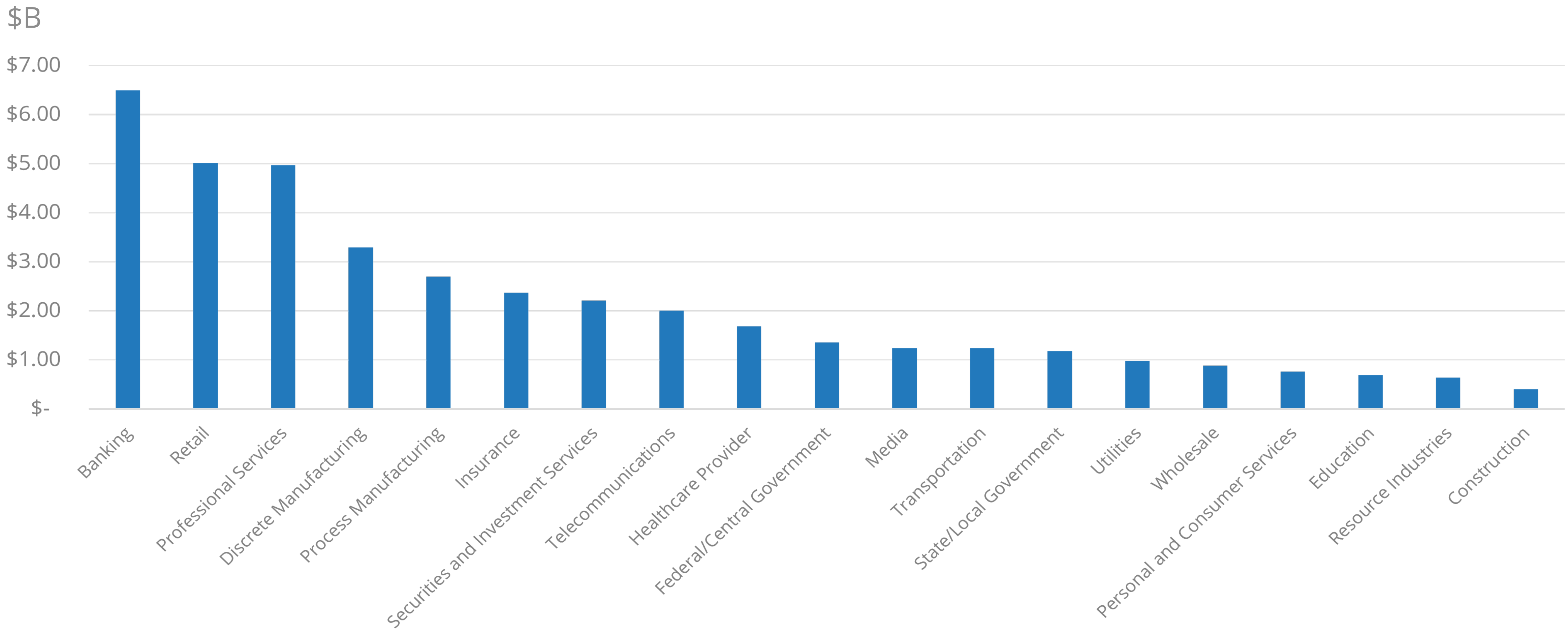
IDC

Navigating the Uncharted: Managing the risks in Generative AI implementations

Daniel Saroff, Group Vice President,
End User Consulting and Research
April 2024

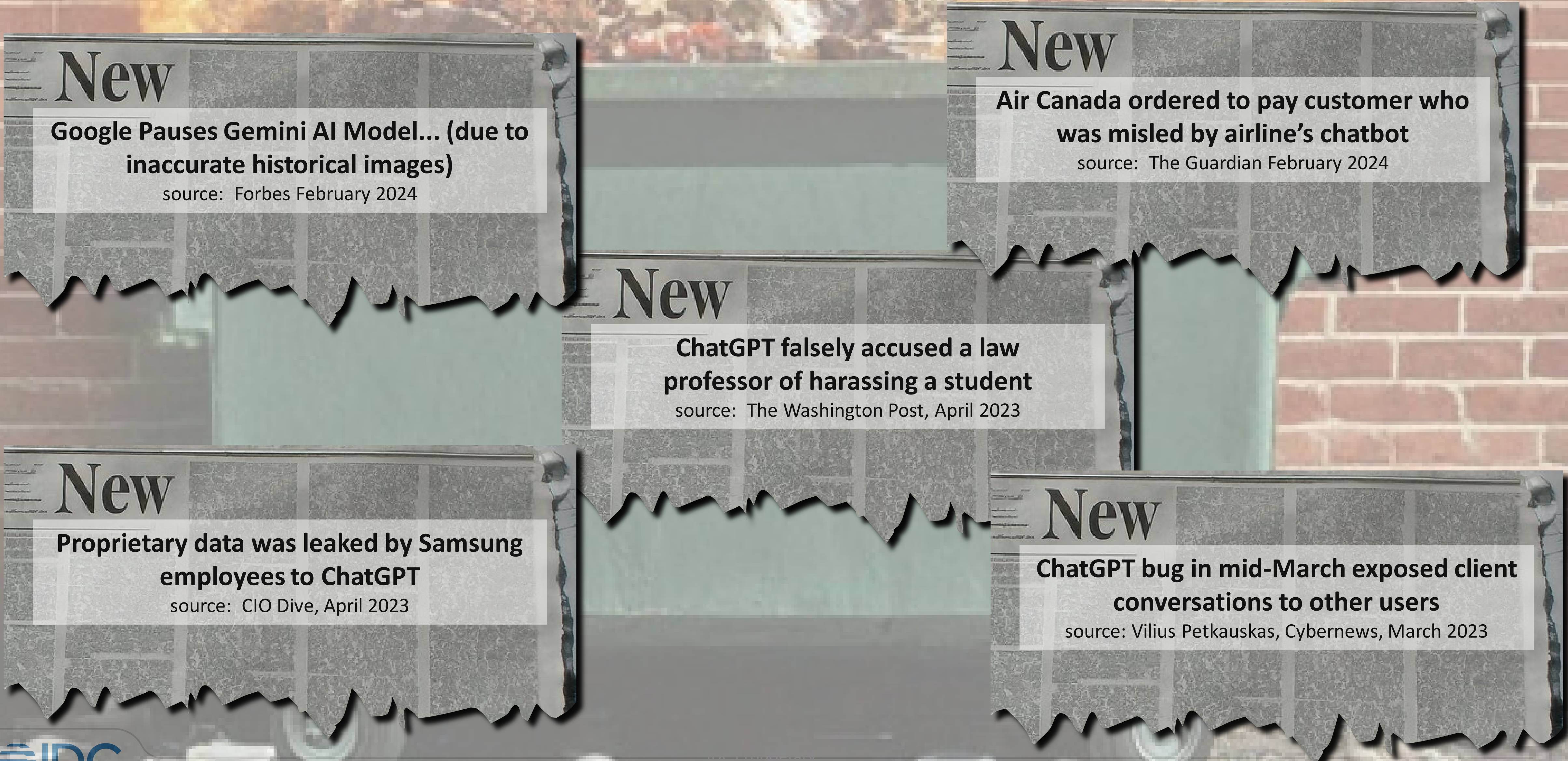
What is GenAI spending by industry expected to be in 2024?

Worldwide Core IT Spending for Generative AI by Industry will be Over $40 Billion in 2024



IDC #US51697124 (January 2024) Source: Worldwide Core IT Spending for GenAI Forecast, 2023-2027 n = 150

Generative AI blunders (just some)



Google Pauses Gemini AI Model... (due to inaccurate historical images)
source: Forbes February 2024

Air Canada ordered to pay customer who was misled by airline's chatbot
source: The Guardian February 2024

ChatGPT falsely accused a law professor of harassing a student
source: The Washington Post, April 2023

Proprietary data was leaked by Samsung employees to ChatGPT
source: CIO Dive, April 2023

ChatGPT bug in mid-March exposed client conversations to other users
source: Vilius Petkauskas, Cybernews, March 2023

What risks are we navigating?

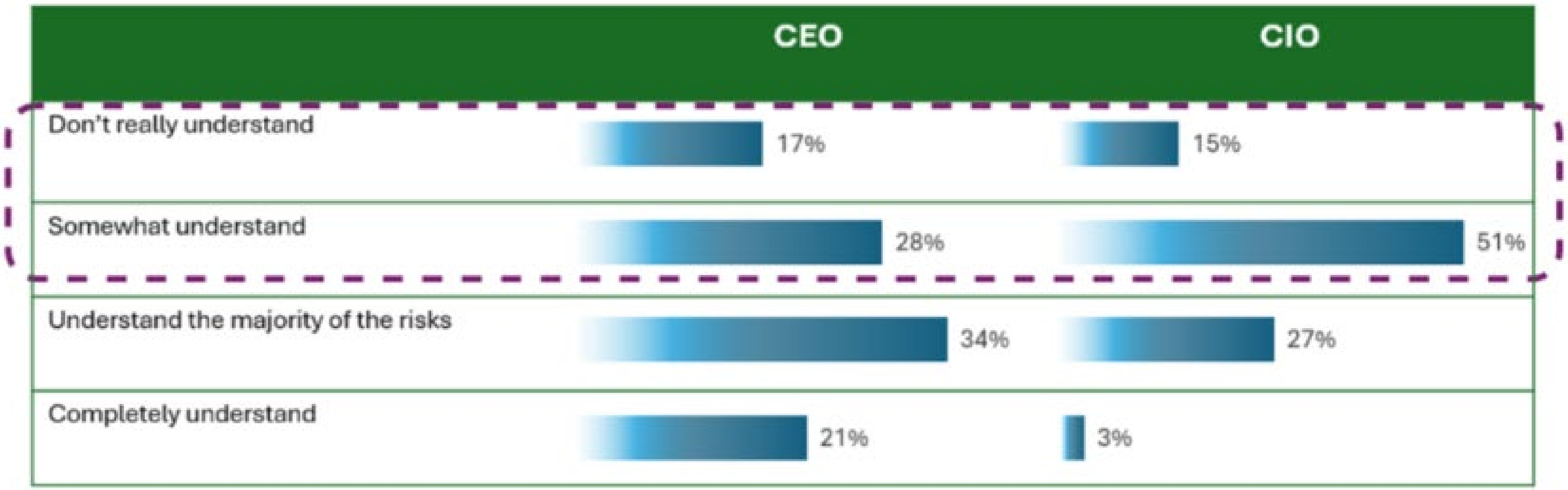
This is not a complete list…

- AI poisoning
- Bandwidth
- Bias
- Brand threat
- Copyright infringement
- Cost
- Data poisoning
- Data spill
- Environmental impacts
- Evolving regulatory landscape
- Governance
- Implementation complexity
- Integration
- Interoperability
- Lack of common standards
- Limited explainability
- Litigation
- Lying/confabulation
- Performance
- Data quality
- ROI
- Security
- Selection
- Sub-optimization
- Vendor lock-in
- …

CEO, CIO opinion on technology vendor’s understanding of the potential risk of AI.

45% of CEOs and 66% of CIOs feel technology vendors do not completely understand the risk potential of AI

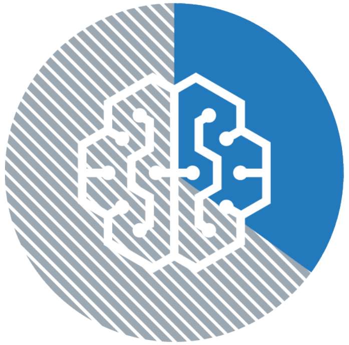
Opinion on technology vendors’ understanding of the potential downside risks of AI



	CEO	CIO
Don’t really understand	17%	15%
Somewhat understand	28%	51%
Understand the majority of the risks	34%	27%
Completely understand	21%	3%

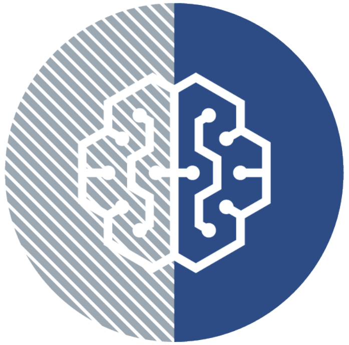
Source: IDC Worldwide CEO Survey 2024; IDC CIO Quick Poll (January 2024)

“Implementation” of GenAI is not always standalone. It is being infused into enterprise applications



35%

The average number of applications using some form of AI/ML or DL TODAY

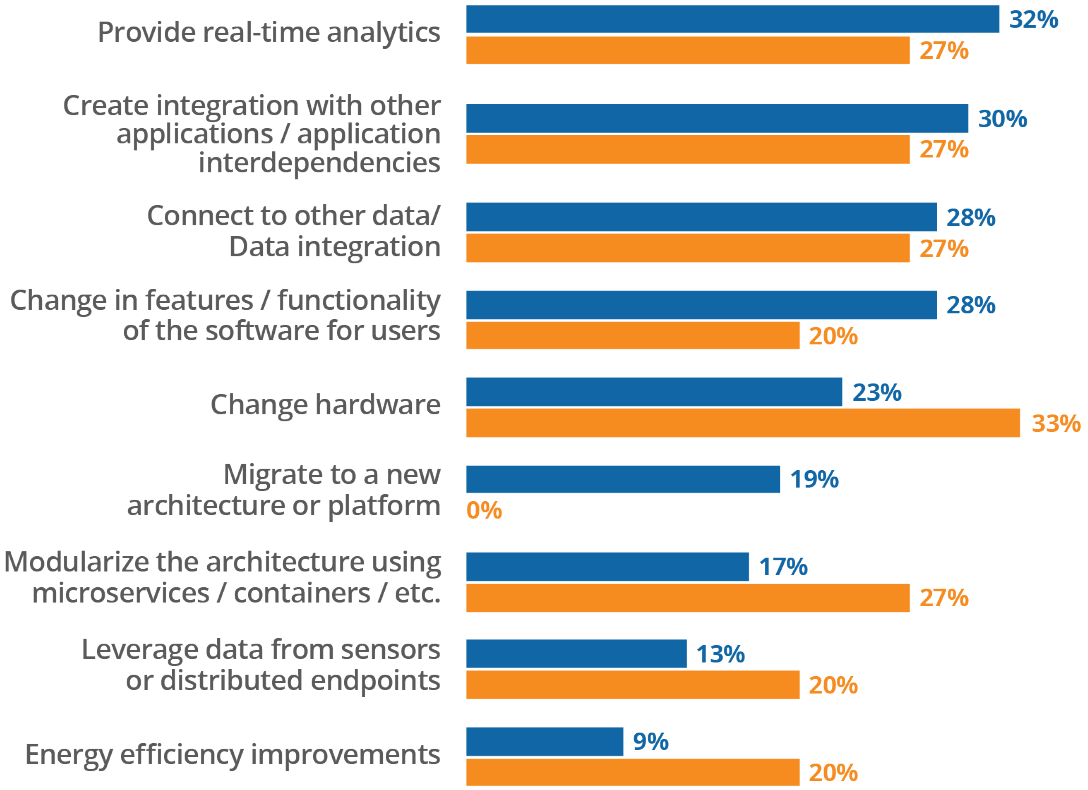


50%

IN TWO YEARS

AI applications will be integrated with other applications across the cloud portfolio

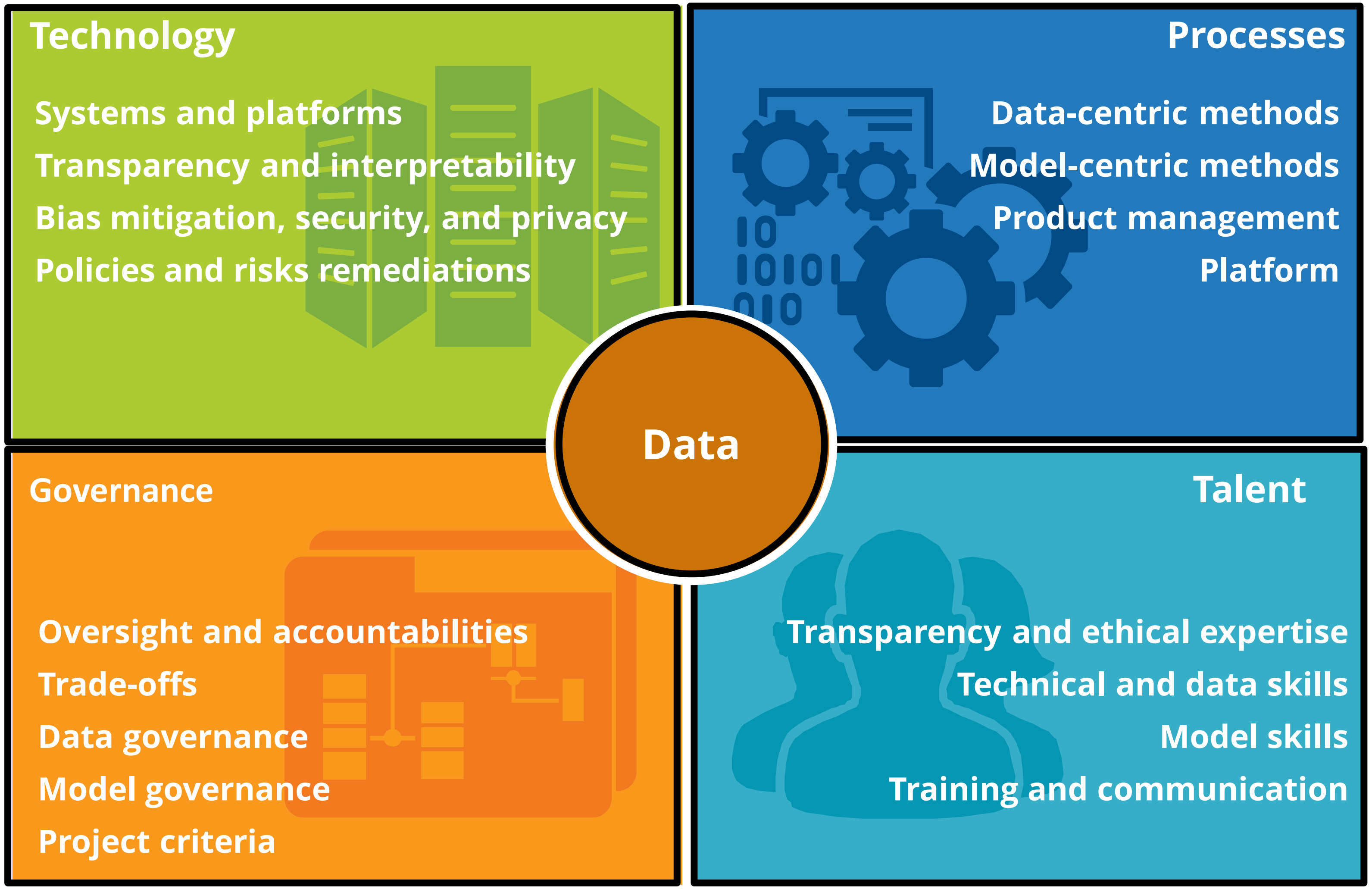
What type of transformation do you expect for your 🟦 AI software services and 🟧 AI lifecycle applications?



N = 1,357 Source: QA17, QD11 Cloud Pulse 2Q23, August 2023 IDC

Five components of GenAI implementation

Checklist for managing risk



Technology

- Systems and platforms
- Transparency and interpretability
- Bias mitigation, security, and privacy
- Policies and risks remediations

Processes

- Data-centric methods
- Model-centric methods
- Product management
- Platform

Data

Governance

- Oversight and accountabilities
- Trade-offs
- Data governance
- Model governance
- Project criteria

Talent

- Transparency and ethical expertise
- Technical and data skills
- Model skills
- Training and communication

Start by assessing your maturity*

1	2	3	4	5
AD HOC	OPPORTUNISTIC	REPEATABLE	MANAGED	OPTIMIZED
AI Laggard	AI Apprentice	AI Practitioner	AI Transformer	AI Disrupter
AI is used in silos by select individuals or groups. There is no formal strategy or coordination as part of the broader view of AI potential.	AI is used for isolated projects. Data readiness, governance, skills management, and technology selection are one-off and limited to specific initiatives.	AI is used for multiple projects. Data readiness, governance, skills management, and technology selection are repeated across those initiatives.	Enterprisewide AI strategy aligned to business goals is in place. Data readiness, governance, skills management, and technology are consistent across rolled out AI initiatives.	Enterprisewide AI strategy aligned to business goals and redesigned business models repeatedly create new business value. Data, skills, governance, and technology usage maximize efficiency.

*While this maturity model was designed for traditional AI, it is consistent with GenAI.

Typical use cases

Three categories of GenAI use cases deliver increasing business value with correlated maturity requirements and risk

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 or Efficiency	• Task **productivity** • Operational **efficiency**	• **Limited talent** in-house • Limited **budget** • **Low risk** appetite • Early adoption • Limited/**poor data**	• Commercial applications with **embedded GenAI** • Native GenAI **standalone applications** (e.g., Microsoft Copilot, Jasper AI, etc.) • **Commercial** models	• Summarizing a report • **RFP** creation • **Code** production • Use case development
Functional	• Increased functional **effectiveness** • **Contextualized experiences**	• **Good data** • Available **talent in-house** • **Budget** available • **Medium risk** appetite	• Fine-**tuning open-source** models • Fine-tuning models available from model hubs and AI platforms • Retrieval-augmented generation **(RAG)**	• Hyper-personalized sales and marketing • Hyper-**personalized wealth and investments** management • Generative product design and prototyping
Industry or Transformational	• **New digital business models**, products, and services • Competitive **moats**	• **Strategic** differentiator • Talent in-house or partner • **Quality** and quantity of **institutional data**	• Fine-**tuning third-party or industry** models • Custom-built models (**BYOM**) • **Strategic partnering**	• Generative **drug discovery** • Generative **material design**

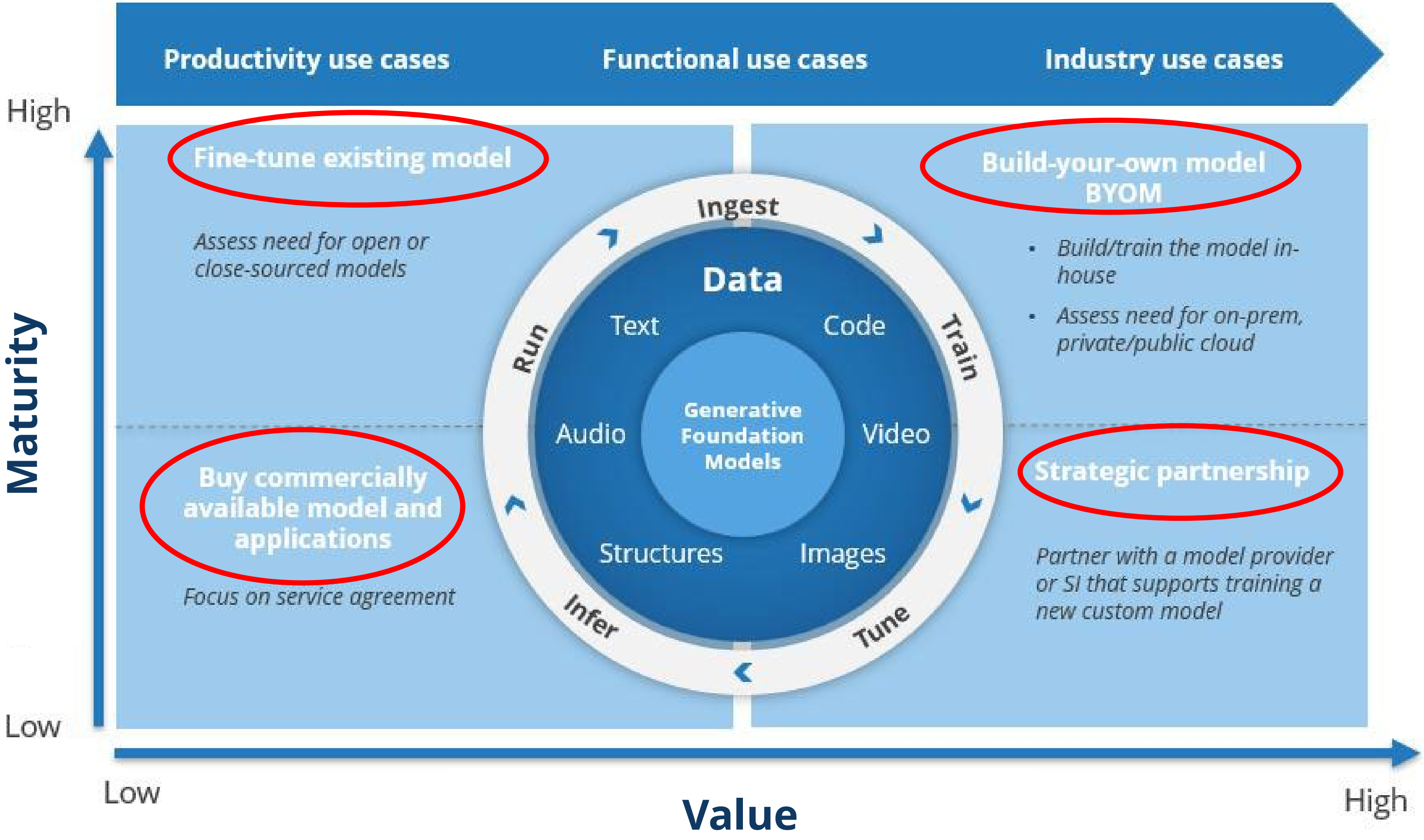
Increasing Maturity, Risk, and Benefit

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

Build versus buy

A false dichotomy – really build **and** buy

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



Fine-tune existing model

Assess need for open or close-sourced models

Build-your-own model BYOM

- *Build/train the model in-house*
- *Assess need for on-prem, private/public cloud*

Buy commercially available model and applications

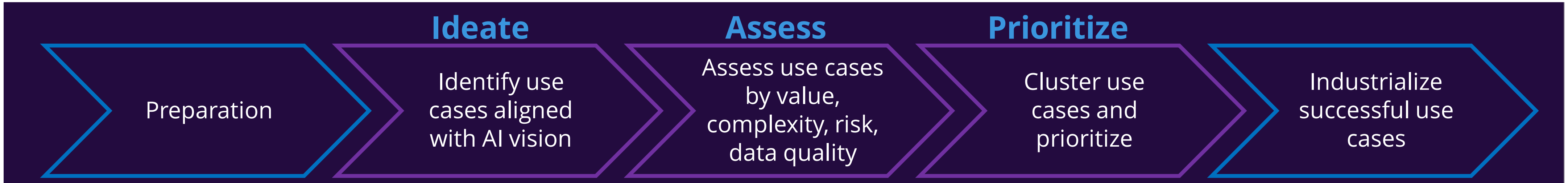
Focus on service agreement

Strategic partnership

Partner with a model provider or SI that supports training a new custom model

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

Balancing risk, value, complexity, and data quality



Ideate

Assess

Prioritize

Preparation

Identify use cases aligned with AI vision

Assess use cases by value, complexity, risk, data quality

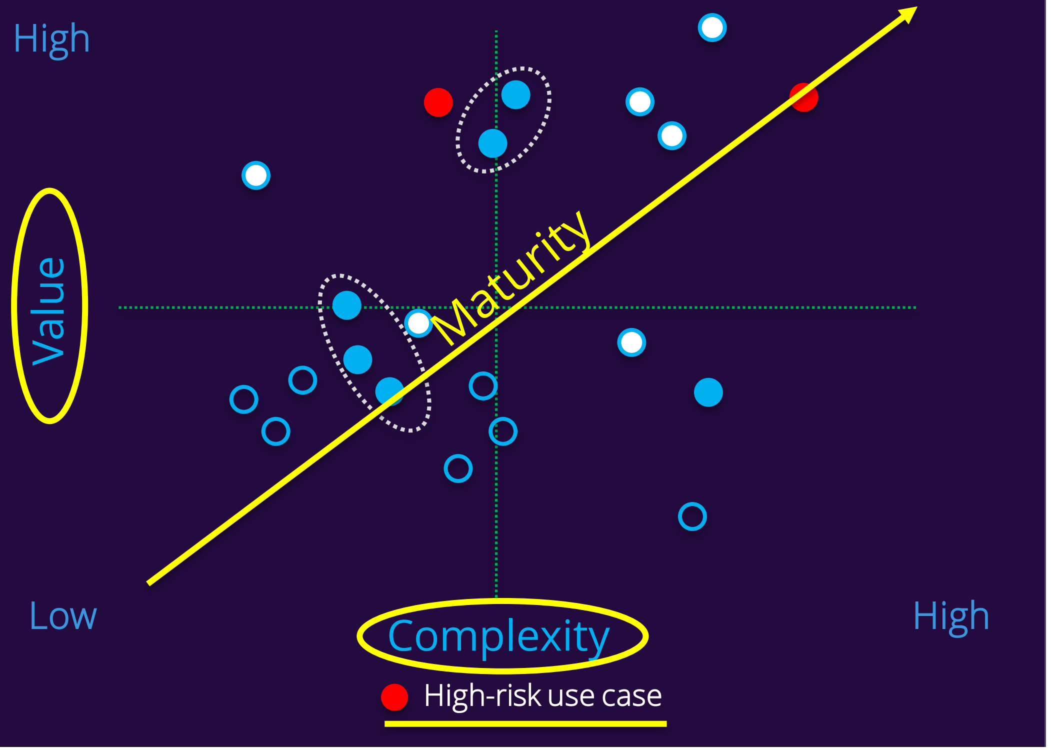
Cluster use cases and prioritize

Industrialize successful use cases

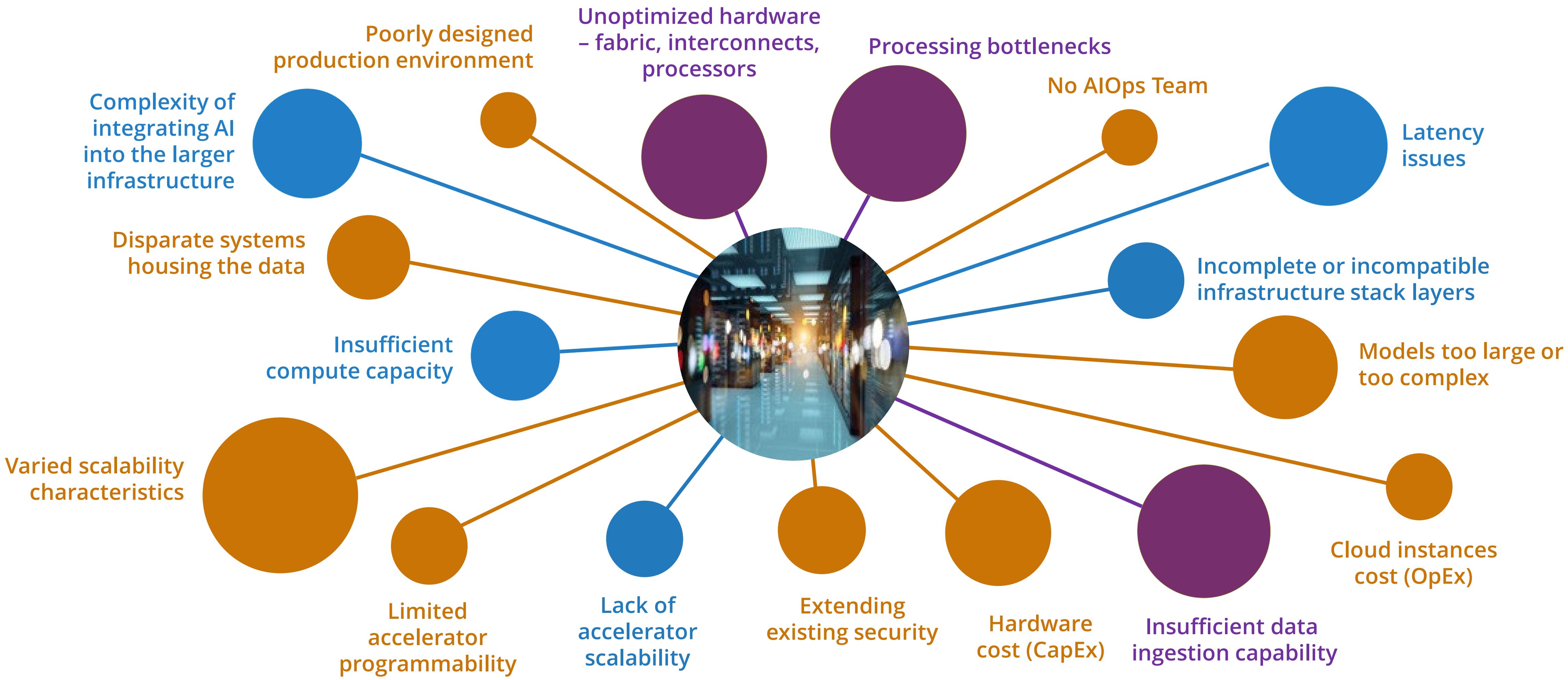
Dimension	
Value	Economic value
	Strategic alignment
Complexity	Data
	Algorithm
	Processing systems
	Required 'know-how'

Data Quality

Low ○ ········ ○ ········ ● High



Many AI projects fail because of improper attention to infrastructure



Complexity of integrating AI into the larger infrastructure

Poorly designed production environment

Unoptimized hardware – fabric, interconnects, processors

Processing bottlenecks

No AIOps Team

Latency issues

Incomplete or incompatible infrastructure stack layers

Models too large or too complex

Cloud instances cost (OpEx)

Insufficient data ingestion capability

Hardware cost (CapEx)

Extending existing security

Lack of accelerator scalability

Limited accelerator programmability

Varied scalability characteristics

Insufficient compute capacity

Disparate systems housing the data

Selecting the right infrastructure stack

Part 1 – What and how

Like any blanket recommendation: YMMV. These are guides, not rules

	Lean on-premises or collocation (Private Cloud) if	Lean off-premises (Public Cloud) if
AI Initiatives	On-going AI, consistent initiatives	Intermittent AI initiatives
System utilization	Keeping utilization rates high (keep expensive processors busy)	Limited/inconsistent utilization rates
IT Skills	In-house skills for complex AI deployments	Limited IT skills for AI deployments
Facilities	Sufficient datacenter floorspace, power, and cooling capabilities	Limited floor space, power, and cooling
Opex friendly options	System vendor can provide consumption-based pricing	System vendor can provide capital only pricing

Selecting the right infrastructure stack

Part 2 – Model considerations

	Lean on-premises or collocation (Private Cloud) if	Lean off-premises (Public Cloud) if
Model Iteration	Many model training iterations	Fewer model iterations
Model Scaling	High scaling needs	Lower scaling needs
Model Accuracy	Highly customized	Limited customization
Model customization	Heavily customized	No API changes or customization
Model Performance	High performance requirements	Lower performance requirements

Selecting the right infrastructure stack

Part 3 – Data considerations

	Lean on-premises or Collocation (Private Cloud) if	Lean off-premises (Public Cloud) if
Data sensitivity	Sensitive data, strict data compliance requirements, proprietary data	Data is not proprietary, limited compliance requirements or sanitized
Data isolation	Model data cannot mix with public data, requires isolation	Model data can safely mix with public data, does not require isolation

Selecting the right partner when building an AI infrastructure stack

Hardware

Silicon

intel, NVIDIA, AMD, AMPERE, Google, aws, IBM, GRAPHCORE, SambaNova Systems, cerebras, SAMSUNG, Qualcomm

Compute

DELL Technologies, Hewlett Packard Enterprise, Lenovo, inspur 浪潮, SUPERMICRO, cisco, FUJITSU, NEC, HUAWEI

Storage

NetApp, PURESTORAGE, DELL Technologies, Hewlett Packard Enterprise, IBM, HITACHI Inspire the Next, FUJITSU, HUAWEI, inspur 浪潮

Networking

ERICSSON, NOKIA, cisco, JUNIPER NETWORKS, HUAWEI, ciena, Netcracker, Adtran, ZTE, NEC, FUJITSU, Hewlett Packard Enterprise

Power and Cooling

Schneider Electric, EATON, VERTIV, submer, EMERSON, ICEOTOPE, ABB, SIEMENS, MITSUBISHI, liquidstack

Software

Security

ARTHUR, AIShield Powered by Bosch, CRANIUM, ROBUST INTELLIGENCE, MINDGARD

Data Logistics and Protection

Deasie, TONIC, BROADCOM, PROTOPIA, VeraData, VERITAS, COMMVAULT, veeam, COHESITY

Software Defined Compute and Storage

Red Hat, vmware, SUSE, Microsoft, cisco, IBM, BROADCOM, NUTANIX, VAST

Systems Management

cisco, bmc, BROADCOM, Microsoft, splunk>, vmware, IBM, HITACHI Inspire the Next, Red Hat, servicenow, amdocs

Services

Cloud Infrastructure

aws, Microsoft, Google, ORACLE, Paperspace, Alibaba.com, Tencent Cloud, IBM, CoreWeave, Lambda

Colocation

EQUINIX, DIGITAL REALTY, NTT, QTS, CyrusOne, Cyxtera, FLEXENTIAL, switch, cologix

Connectivity

NTT, AT&T, verizon, vodafone, T, orange, BT, Telefónica, TATA COMMUNICATIONS, Telstra, américa móvil, etisalat e&, COMCAST, sky

Communications Platforms

twilio, RingCentral, VONAGE, sinch, infobip, Microsoft, zoom, webex by cisco, gupshup, 8x8, tanla, bandwidth, Alibaba.com

Source: IDC, 4Q23

For areas on which IDC publishes market share data, the top 3–5 market share leaders are represented. For areas on which IDC does not publish market share data, vendor selection is up to analyst discretion.

Some Takeaways

- A realistic understanding of AI maturity is key to implementation decisions
- If you don’t have or can’t build AI maturity, buy it with knowledge transfer
- A platform approach with integration to existing systems is key
- Know your data with confidence (or at least the data you will use)
- "Implementation” risk may arise from existing, licensed products incorporating GenAI
- Generative AI is a business transformer that happens to be a technology

Daniel Saroff

Group VP
End User Consulting and Research

dsaroff@idc.com
https://www.linkedin.com/in/daniel-saroff-9301991/

IDC.com

linkedin.com/company/idc

blogs.idc.com