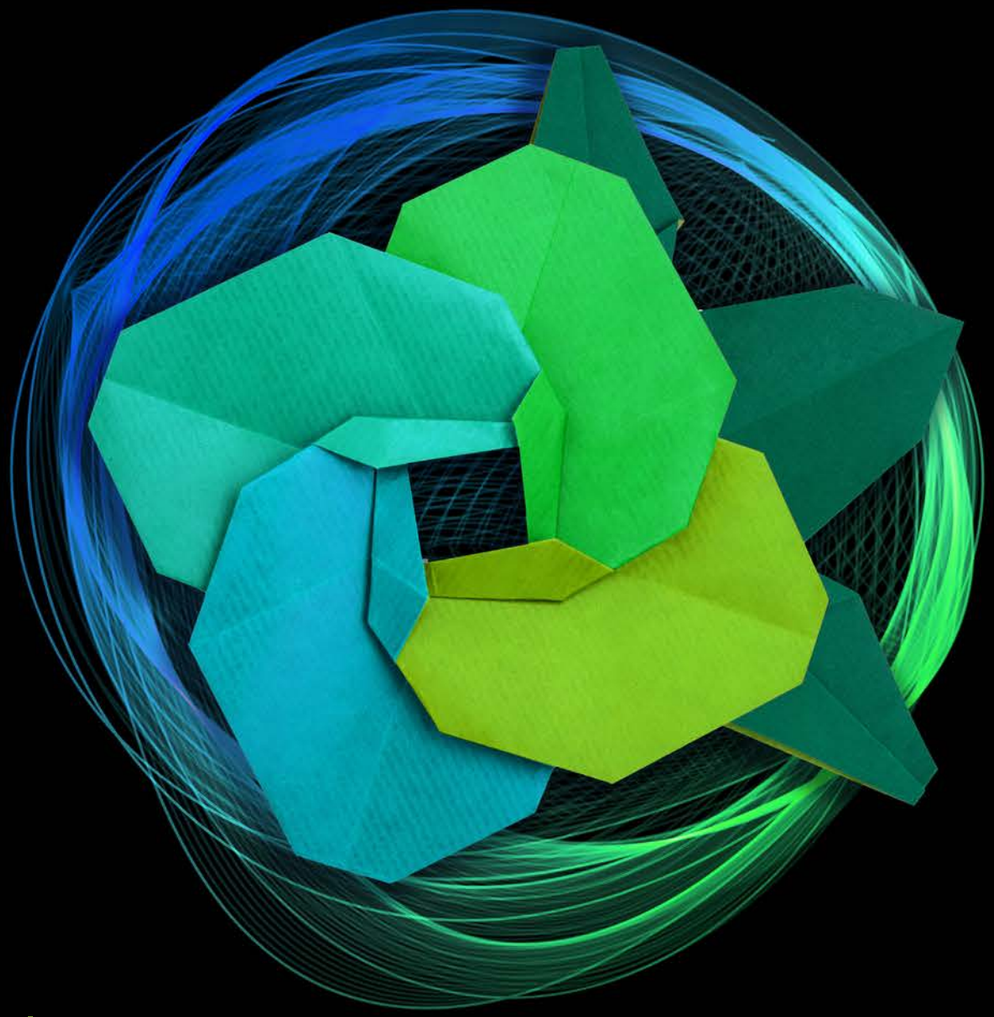
Deloitte.



Scaling Generative AI

13 elements for sustainable growth and value

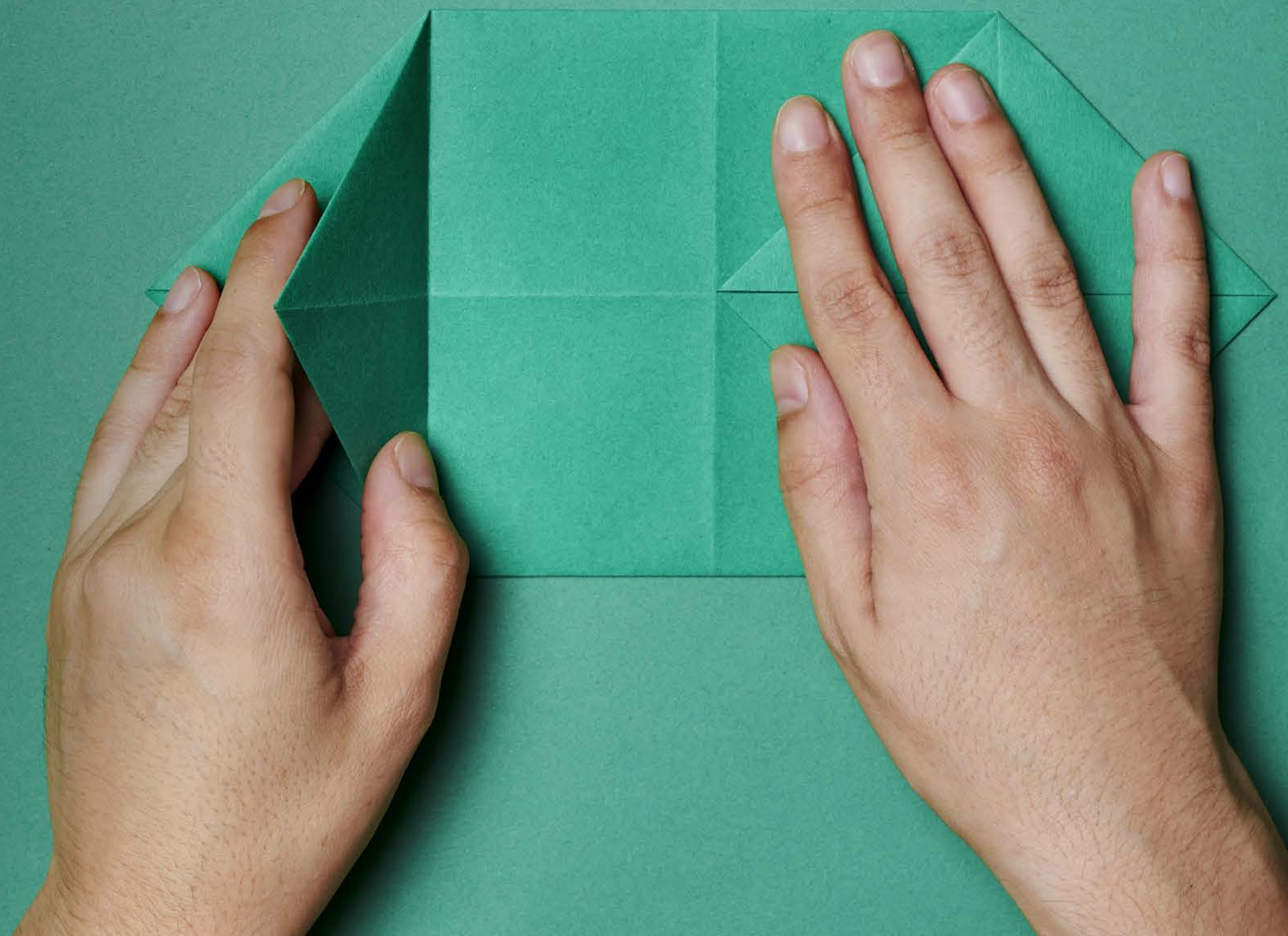
About the Deloitte AI Institute™

The Deloitte AI Institute helps organizations connect the different dimensions of a robust, highly dynamic and rapidly evolving AI ecosystem. The AI Institute leads conversations on applied AI innovation across industries, with cutting-edge insights, to promote human-machine collaboration in the "Age of With".

The Deloitte AI Institute aims to promote a dialogue and development of artificial intelligence, stimulate innovation, and examine challenges to AI implementation and ways to address them. The AI Institute collaborates with an ecosystem composed of academic research groups, start-ups, entrepreneurs, innovators, mature AI product leaders, and AI visionaries, to explore key areas of artificial intelligence including risks, policies, ethics, future of work and talent, and applied AI use cases. Combined with Deloitte's deep knowledge and experience in artificial intelligence applications, the Institute helps make sense of this complex ecosystem, and as a result, delivers impactful perspectives to help organizations succeed by making informed AI decisions.

No matter what stage of the AI journey you're in; whether you're a board member or a C-Suite leader driving strategy for your organization, or a hands on data scientist, bringing an AI strategy to life, the Deloitte AI institute can help you learn more about how enterprises across the world are leveraging AI for a competitive advantage. Visit us at the Deloitte AI Institute for a full body of our work, subscribe to our podcasts and newsletter, and join us at our meet ups and live events. Let's explore the future of AI together.

www.deloitte.com/us/AIInstitute



Near the top of every enterprise agenda is a question of how to leverage Generative AI (GenAI). With use cases proliferating horizontally across functions and vertically within business units, the next step is...**sustainably scaling GenAI for strategic business value.**

Generative AI, like origami, transforms a resource (data and paper, respectively) into a compelling output. Just as origami artists fold paper to resemble interesting shapes, Generative AI computes data to approximate human cognition and creativity.

Getting more GenAI into production

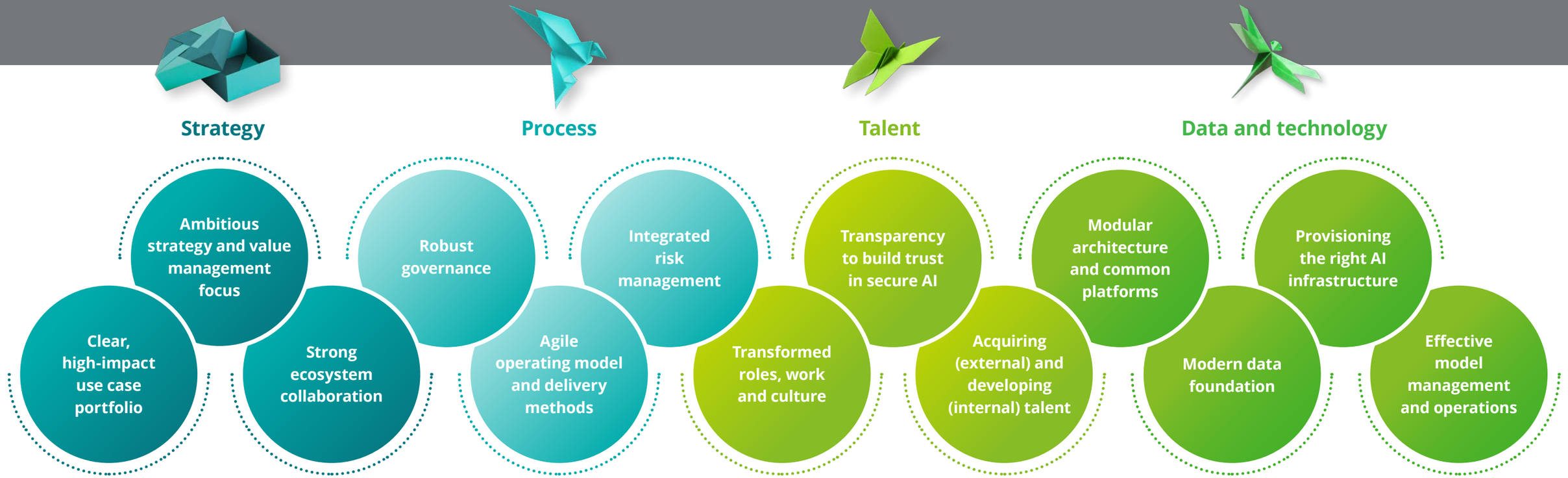
Deloitte's State of GenAI in the Enterprise Q3 report revealed that many businesses are encountering challenges when making the transition from GenAI proof-of-concept to scaled deployment.[1]

Seventy percent of surveyed organizations indicate that less than one third of their GenAI experiments have made it to production. This suggests that while enterprises are investing in GenAI, they are not yet seeing the full potential ROI. A common challenge is defining what is required to achieve GenAI scale at a practical level.

We define scale broadly as the ability of a system to handle a growing amount of work or its potential to be enlarged to accommodate growth with steadily decreasing unit costs. **For GenAI specifically, scaling also means moving from experimentation to implementation in a way that is sustainable, secure, and aligned with business goals.**

GenAI at scale generates more diverse and representative outputs, it can handle more complex tasks, and its speed, output quality, and accuracy are enhanced. As a result, operational costs become more efficient and business impact is governed, measured, and communicated.

At the highest level, **GenAI scaling factors can be grouped into the familiar areas of strategy, process, talent, and data and technology.** Each area presents challenges to be navigated and contains leading practices that help point the way to GenAI value realization.



Essential elements for scaling Generative AI initiatives from pilot to production

STRATEGY

Ambitious strategy and value management focus

An organization's GenAI strategy and vision need to be comprehensive, integrated with broader business objectives, and aligned with other existing AI programs. Executive buy-in and a top-down mandate are essential for aligning functions and decision-making. Leadership sets priorities and strategy, and without an executive mandate, it is difficult to coordinate change across multiple teams. A cohesive GenAI strategy defines business objectives, sets measurable goals, identifies valuable areas for application, and measures realized value. As a part of strategy development, inject waypoints that will show progress against short-term goals and inform any iterative improvements needed to the strategy.

Establish a comprehensive vision with a top-down mandate

STRATEGY

Clear, high-impact use case portfolio

There are six common macro archetypes for GenAI: Q&A-based search, summarization, content generation, content transformation, virtual agent, and code generation. In seeking value-driving applications, organizations should look across the archetypes for low-barrier, high-impact use cases for core business domains. These drive efficiencies and savings that can be reinvested in innovation. Other high-impact use cases may be more transformational and differentiating with enterprise-wide applicability. Whether deploying a proven application or striving for something novel, all applications require technical feasibility and a viable business case. What is more, existing processes will likely need to be reimagined to incorporate and leverage the capabilities of GenAI use cases in workflows.[2] At its core, the use case portfolio needs to be focused on answering business questions and meeting quantified goals. We see leading organizations create business cases that weave together the value GenAI can provide to multiple teams, rather than evaluating the value of individual applications. This is done most effectively by running a number of use cases in parallel. It makes efficient use of resources and allows for rapid portfolio management should a specific use case prove less compelling without sacrificing momentum of the overall Gen AI portfolio.

Explore low-barrier, high-impact use cases to drive efficiencies and savings

STRATEGY

Strong ecosystem collaboration

GenAI is maturing rapidly, with existing providers and new market entrants alike driving capabilities and lateral applications. The array of GenAI solutions and the speed with which they are evolving can make it challenging to select the appropriate tools and platforms that enable enterprise strategy. To reach target outcomes, enterprise leaders should build strategic relationships with technology and data ecosystem stakeholders and keep pace with GenAI development. By monitoring elements like product roadmaps, total cost of ownership, and labor delivery models, business leaders can gain an understanding as to how their GenAI programs should evolve and how ecosystem players can accelerate progress and results as strategic partners, rather than as transactional vendors. A framework can support a structured approach to evaluating solutions based on factors such as data readiness, AI maturity, risk appetite, and total cost of ownership.

Evolve with existing providers and new market entrants alike

PROCESS

Robust governance

Inconsistent processes can lead to risks and inefficiencies, while consistent governance processes help standardize workflows for data collection, solution engineering, output validation, and performance monitoring. Common delivery frameworks (e.g., LLMOps) bring together GenAI development and deployment into a unified, governed lifecycle that is secure and compliant. A common misconception is that strong processes can hinder speed and creativity. Our experience suggests the opposite. By understanding how work needs to be done and the accompanying guardrails, teams are empowered to explore ways to generate value without fear that they may be making a mistake. Clear boundaries allow freedom for bold action and innovation, while a lack of clarity may lead to more conservative approaches. Governance includes documented roles and responsibilities driving stakeholder accountability in decision-making across the AI lifecycle, and it informs the controls for risk identification and mitigation. Governance also standardizes how stakeholders identify, prioritize, and approve GenAI applications. As processes are amended, organizations need to be mindful about disrupting existing automated or manual controls and take steps to establish assurance in those amended processes. Even as the regulatory landscape is in flux, organizations should proactively establish governance processes that meet existing or likely regulatory requirements.

Create repeatable governance processes to help standardize work

PROCESS

Integrated risk management

For GenAI to reach its full potential business value and adoption, it must be trusted and secure.[3] Attempting to scale without accounting for trust in data and the machine that consumes it can have implications for regulatory compliance, finance and strategy, cybersecurity and privacy, adoption and change management, and brand reputation—the consequences of which can limit or even erase GenAI's intended value. Risk and trust need to be considered and addressed across the GenAI lifecycle, from design and development through deployment and scaled implementation. This includes validation processes and feedback loops for human oversight to manage solution performance and accuracy. It also includes guardrails to ensure privacy, drive ongoing compliance, and promote agility in proactively responding to emerging risks. Data security is particularly essential. Differentiating GenAI applications are fueled by sensitive, proprietary enterprise data, and training and usage can potentially expose or leak business-critical data and create risks to the organization. This is not a one-time event—organizations must make this part of regular work, rather than a separate consideration.

Address risk and data security across the GenAI lifecycle

PROCESS

Agile operating model and delivery methods

The operating model impacts how the enterprise aligns technology, processes, and roles and responsibilities to create strategic business value. An integrated model connects the blueprint for value with AI business cases to inform how work is delivered and helps drive alignment across the enterprise. As the marketplace matures and new capabilities and risks impact AI lifecycles and governance, the organization needs to be agile in matching internal opportunities with the right technologies. To help, organizations may turn to technical experts or an AI Center of Excellence (COE) that equips decision makers with the insight to align the vision for success with the organization's AI maturity and ambition. This supports a cohesive approach to orchestrating the elements of GenAI development and application. It helps avoid AI and data silos and instead drive toward reusable building blocks, coordinated sourcing strategy, informed build-versus-buy decisions, and security and risk management.

Support a cohesive approach to orchestrating the components

TALENT

Transparency to build trust in secure AI

Trust in GenAI is essential to increasing workforce adoption and realizing benefits. With GenAI, employees may have existing biases, reticence, skills gaps, or even a fear that they could be replaced by a machine. Trust in GenAI grows out of transparency, where every stakeholder understands how the enterprise is pursuing GenAI applications, how they are intended to create value, and how the workforce can leverage these tools as efficiency and productivity enhancers. Transparency around the benefits targeted by GenAI solutions helps correct misinformation and creates an opportunity to improve the workforce experience. Trust is also important for external stakeholders, third parties, and customers, and a transparent approach to GenAI use includes consent for data collection, notification of how GenAI outputs may impact users, and documentation across the AI lifecycle to inform audits and compliance.

Help stakeholders understand the GenAI vision and how it creates value for them

TALENT

Transformed roles, work, and culture

Deployments at scale can disrupt the status quo, transforming employee responsibilities and how work is accomplished. As an enterprise reimagines strategy, processes, and technology to drive GenAI value, the workforce needs to be brought on the journey as value is created through individuals doing work differently. Organizations should nurture adoption by documenting and communicating responsibilities and process amendments to workflows. Poor communications may cause misunderstanding about GenAI's potential and limitations, leading to unrealistic expectations or resistance. Conversely, effective communications align stakeholders around the same vision for scale and value, including as they relate to governance, policy, IT security, risk, and funding. Topics to communicate include outcomes and lessons learned, the organization's AI roadmap, the impact on end users (e.g., customers or employees), and guidance to the workforce on how to balance day-to-day tasks with AI skills development. Ongoing adoption should be measured to identify optimization opportunities and internal leading practices. This should inform the overall use case roadmap and activation strategy. Simply put, upstream conversations should take place before continuing to build technical solutions that are underdelivering against expectations.

Nurture adoption by documenting responsibilities and process amendments

TALENT

Acquiring (external) and developing (internal) talent

Organizations deploying GenAI need to consider the skilled human talent required across the GenAI lifecycle. Skills mapping can reveal where the enterprise needs to expand or improve the workforce. Recruiting new talent is one avenue, such as by attracting new employees from educational facilities (e.g., universities). In reimagining work with GenAI, the organization may attract new leaders who are eager to use technology to deliver business value, as well as top talent seeking opportunities to learn and develop. Yet, most of a company's GenAI capabilities will grow out of training and upskilling existing employees, and as GenAI touches every part of the enterprise, the entire workforce requires training to adopt and use it. To this end, businesses may create overall AI literacy programs, training plans tailored to employee personas (e.g., technical, functional, sales, marketing, etc.), and opportunities (e.g., hackathons and digital playgrounds) for employees to apply new knowledge and build competence in GenAI application, management, and monitoring. A GenAI COE can help orchestrate human-centered continuous learning to promote adoption.

Balance talent acquisition with workforce upskilling

DATA & TECHNOLOGY

Modular architecture and common platforms

IT architecture needs to evolve as technologies mature and as the organization's needs change. Flexibility in modular systems includes leveraging microservices and APIs for tech stack integration, as well as techniques for improving output reliability (e.g., retrieval augmented generation, fine-tuning). This enables platform and model "lift and shift" and supports partnerships with hyperscalers that can provision on-prem or cloud-based environments via contracts that reward increased volume with lower unit costs. In prioritizing a modular architecture, organizations can facilitate user growth with a cost-per-user model, automate guardrails for managing GenAI risk, leverage GenAI capabilities in enterprise software platforms, and establish an internal marketplace where users can select models, access prompt catalogs, and leverage existing solutions. Modular architecture and delivery also accommodate low-code platforms for business users and provide a clear pathway to industrializing capabilities.

Prioritize a flexible IT architecture to facilitate enhancements

DATA & TECHNOLOGY

Provisioning the right AI infrastructure

GenAI infrastructure includes reusable assets, data pipelines, solution development environments, and a range of post-deployment management and feedback capabilities. Bringing the right secure infrastructure to the right place in the GenAI value chain is necessary for sustainable, cost-effective scale. Taking an AI Factory approach enables reusable components and data products while also integrating sourcing strategy, cybersecurity considerations, demand generation, prioritization, governance, and business outcomes. While focusing on speed to value and taking an agile, incremental approach to infrastructure development, organizations can look to iterative design and continual evaluation of cost mechanisms against a per-user or per-use model. One important consideration is that executives are likely to be more comfortable funding enhancements to existing capabilities, as opposed to building net-new systems. Using existing investments and approaching scale as building incremental capabilities can help encourage investments by overcoming a misperception that a GenAI endeavor is starting from scratch.

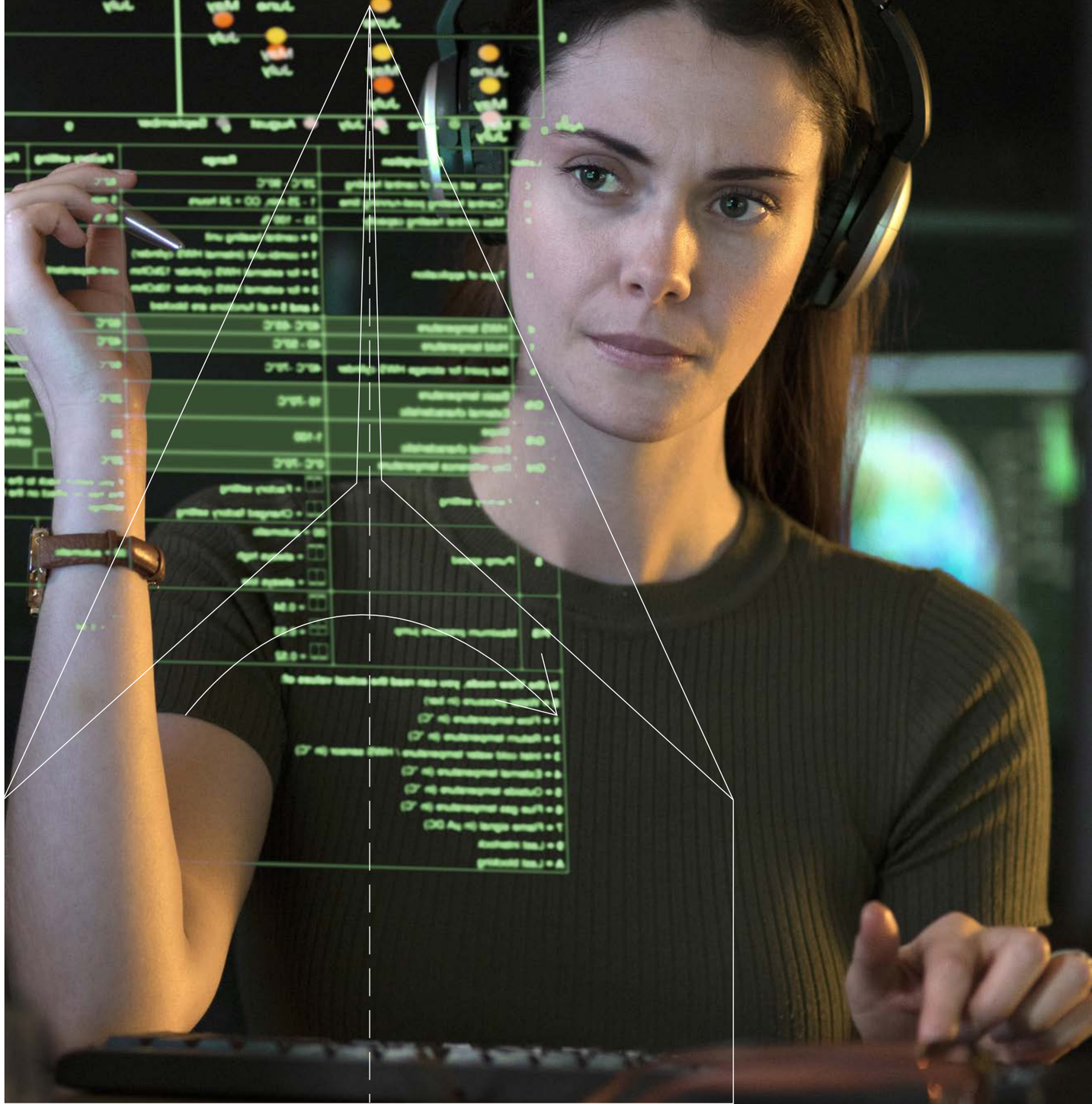
Take an agile approach to enable continuous improvement

DATA & TECHNOLOGY

Modern data foundation

As organizations increasingly shift to hybrid-cloud environments, data integration challenges may increase, with proprietary and third-party data sources existing on disparate platforms. In addition to master data, GenAI applications consume other forms of data (e.g., reference, unstructured first-party) that traditionally sit in the realm of knowledge management. Value creation opportunities from GenAI are blending knowledge and data management capabilities. Data quality and accessibility issues can limit value and potentially create a perception that scaled solutions are not viable nor valuable. A GenAI-ready data foundation includes the processes, philosophies, approaches, and approvals for data sharing and use. As a part of this, evaluate the organization's data findability, accessibility, interoperability, reusability, and storage. Rather than starting from scratch, the organization's existing data governance efforts can likely be extended and adjusted to accommodate unstructured data. Data should also be curated and integrated across departmental lines. Consider a parallel workstream for data readiness evaluation and progression focused on clean and organized data, efficient data pipelines, and robust data governance practices. By ensuring systems are secure and foundational data capabilities are aligned with the GenAI strategy and governance, enterprises can evolve data availability, engineering, and management to enable adoption and scale. At the same time, it is worth noting that interim value can be harvested, albeit at a lower potential, while comprehensive and foundational data modernization activities are underway.

Align data capabilities and processes with GenAI strategy to support quality and accessibility



DATA & TECHNOLOGY

Effective model management and operations

Trustworthy, compliant GenAI applications require coordinated solution management, including continuous monitoring for impartial output accuracy, waypoints for decision-making, and data feedback loops for continuous improvement. Cost management is also a factor. GenAI deployment raises questions around variable and fixed costs, and business leaders need visibility into managing and forecasting end-to-end costs for infrastructure, tools, personnel, maintenance, and models. Insourcing key functions may permit differentiation or better economics over time, and insourcing decisions need to be balanced against the cost to build a capability, the ramifications of moving to a fixed versus variable cost, and the expenses associated with capability management (e.g., hiring and training, oversight, technology acquisition, facilities).

Monitor for impartial output accuracy and focus on cost management

Measuring success with GenAI at scale

The value of scaled GenAI deployments is found in how they advance an integrated enterprise strategy and drive toward business goals.

Establishing realistic goals for quantitative KPIs (beyond productivity and efficiency metrics, such as hours saved) allows the enterprise to assess whether the scaled deployment is achieving its intended business impact. With a use case portfolio that balances cost- and revenue-oriented value levers, there are **key indicators that reveal whether the enterprise is on the right track:**

- **Increased speed to market,** from ideation to deployment
- A decline in proof-of-concept demand, as **demand shifts to low-code** environments available to business users
- A decrease unit cost for new capabilities/solutions, with technical solutions and code being reusable, thus **reducing development efforts**
- An increase in the number of foundational capabilities that help the organization **access GenAI advancements** as they emerge
- An increase in domain-specific models allowing for **more use cases and broader application** across the organization
- **Increased use** of capabilities and solutions, owing to a growing number of users in the enterprise
- An **increase in stated value** realization on a cumulative basis due to GenAI
- An **increase in internal certification/badging** of existing employees in GenAI capabilities, both functional and technical
- Use of GenAI to **redefine a business process,** rather than embedding GenAI in existing business processes

GenAI capabilities are improving and multiplying, and at this point, few organizations are likely to have achieved each element of scale to their greatest capacity. The leading practices, governed processes, and ecosystem of complementary technologies are still being developed and defined.

While change is inevitable, pursuing the elements of scale today positions the organization to go live with GenAI for business value as this transformative technology evolves.

Let's connect

Reach out for a conversation on scaling Generative AI

Lou DiLorenzo Jr.
US AI & Data Strategy Practice Leader
US CIO & CDAO Programs Executive Sponsor
Deloitte Consulting LLP
ldilorenzojr@deloitte.com

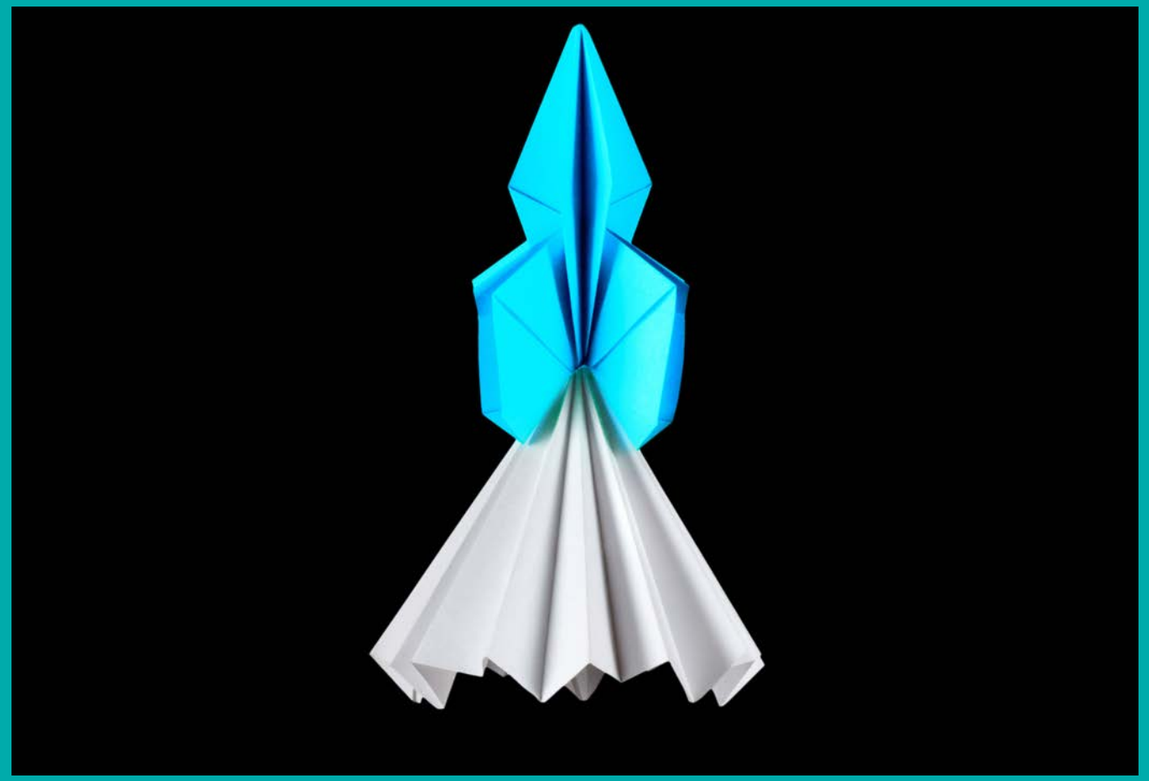
Edward Van Buren
Government & Public Services Leader – Applied AI
Deloitte Consulting LLP
emvanburen@deloitte.com

Rohit Tandon
US AI & Insights Practice Leader
Deloitte Consulting LLP
rotandon@deloitte.com

Sangha Pati
USI AI & Insights Practice Leader
Deloitte Consulting LLP
spati@deloitte.com

Aditya Kudumala
Life Sciences Global AI Leader
Deloitte Consulting LLP
adkudumala@deloitte.fr

Jenn Malatesta
Commercial Officer
Deloitte & Touche LLP
jemalatesta@deloitte.com



Acknowlegements

The authors would like to thank the following leaders and colleagues for their contributions to this effort.

Kevin Abraham, Beena Ammanath, Aniket Bandekar, Kevin Byrne, Ricky Franks, Justin Hienz, Kevin Hutchinson, David Jarvis, Carissa Kilgour, Lena La, Geoff Lougheed, Parth Patwari, Brittany Rauch, Jim Rowan, Kristin Ruffe, Baris Sarer, Dean Sauer, Laura Shact, Brenna Sniderman, Ian Thompson, and Saurabh Vijayvergia.

Endnotes

1 Jim Rowan, Beena Ammanath, Brenna Sniderman et al, "Now decides next: Moving from potential to performance, Deloitte's State of Generative AI in the Enterprise," Quarter three report Deloitte, August 2024.

2 Rowan, Ammanath, Sniderman et al, "Now decides next."

3 Deloitte, "TrustworthyAI™, Bridging the ethics gap surrounding AI," accessed 3 October 2024.

Deloitte.