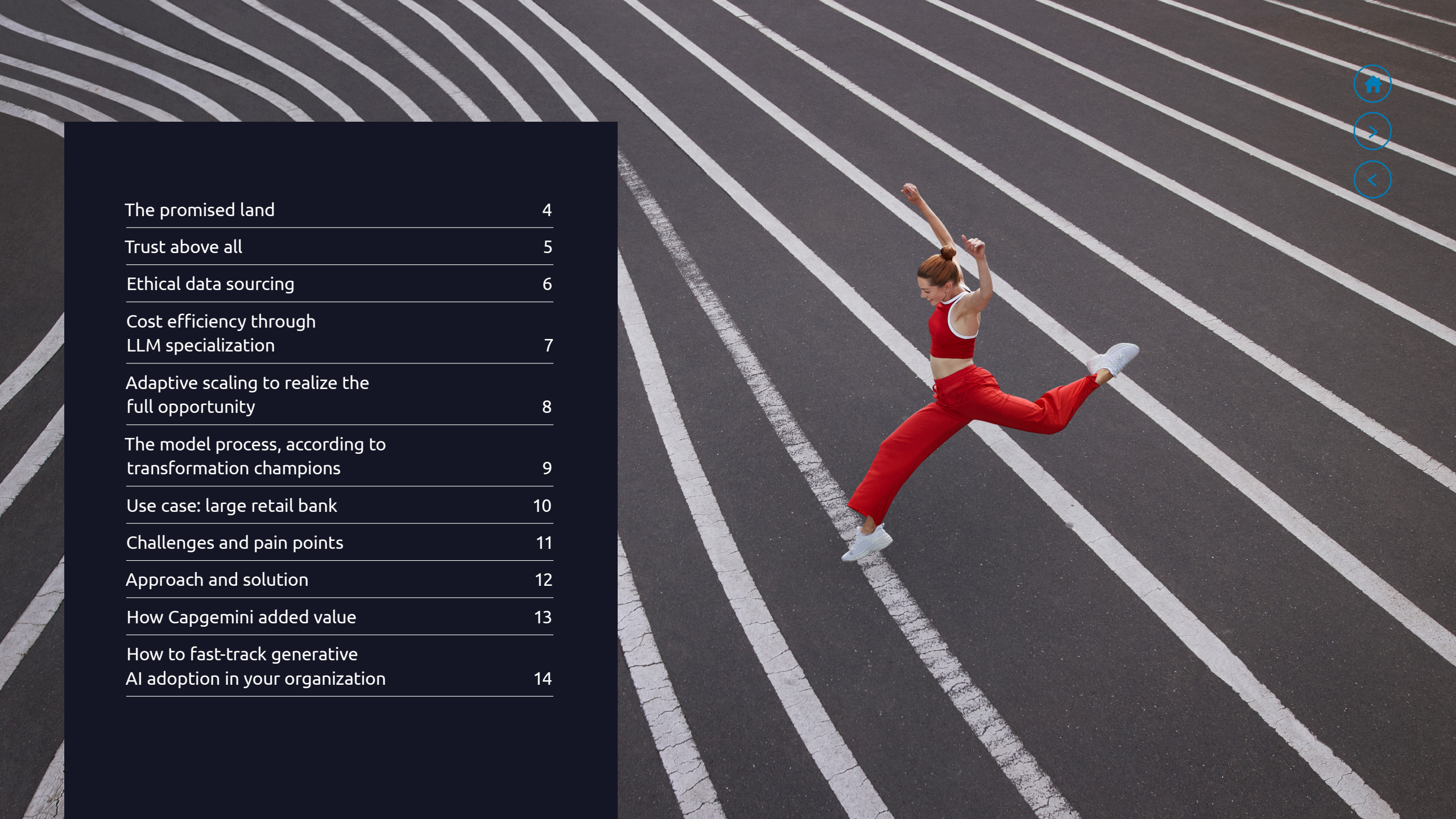
Riding the next *generative AI* wave

To be on the right side of the generative AI opportunity requires strategic planning for trust, cost, and scale.

Capgemini



California is once again the center of a gold rush. OpenAI, an artificial intelligence company with headquarters in San Francisco, caused the first wave of mania for generative artificial intelligence (or "Gen AI") when it launched ChatGPT in 2022[1]. Within two months it had two million users.

A Large Language Model (LLM) using generative AI can learn the properties and patterns of data and reapply them for a wide range of applications, from creating text, images, and videos in different styles to generating tailored content. Using deep learning techniques it can also produce valuable, actionable business insights, and improvements in processes for an entire enterprise. These can be integrated into business platforms to increase productivity, improve marketing content quality, enhance a customer experience, or in a production setting, execute quality control with radical effectiveness.

Similar to the sustained economic growth that followed the 19th century gold rushes in California and around the world[2][3], many expect generative AI to lead to a permanent global productivity surge[4]. But like the individual gold prospectors in 1849 who were left behind as mining companies industrialized gold prospecting, businesses that do not strategize for scale and cost, along with trust, will not realize the full value of this technology shift[5].

Where it took cloud computing and mobile technology around a decade each for their impact to be fully absorbed into mainstream businesses, with Gen AI, business intelligence provider Gartner forecasts it to become a fundamental business requirement within two to five years[6].

For these reasons, many senior business decision-makers want to have something in Gen AI, to not miss its advantages and not be left behind by competitors. In a 2023 Capgemini Research Institute survey, 96% of boardroom executives said that AI was a hot topic of discussion in their boardrooms[7]. They are motivated by its promises of enterprise value and operational efficiency.

1 Economist.com, How San Francisco staged a surprising comeback, 12 February 2024
2 Reeves, Keir and Lionel Frost, Charles Fahey, Integrating the historiography of the nineteenth-century gold rushes, 22 June 2010
3 Economist.com, Generative AI generates tricky choices for managers, 27 November 2023
4 Ibid.
5 Picture This: California Perspectives on American History, Gold Rush: 1848–1860
6 Gartner, Gartner Places Generative AI on the Peak of Inflated Expectations on the 2023 Hype Cycle for Emerging Technologies, 16 August 2023
7 Capgemini Research Institute, Generative AI Executive Survey, April 2023

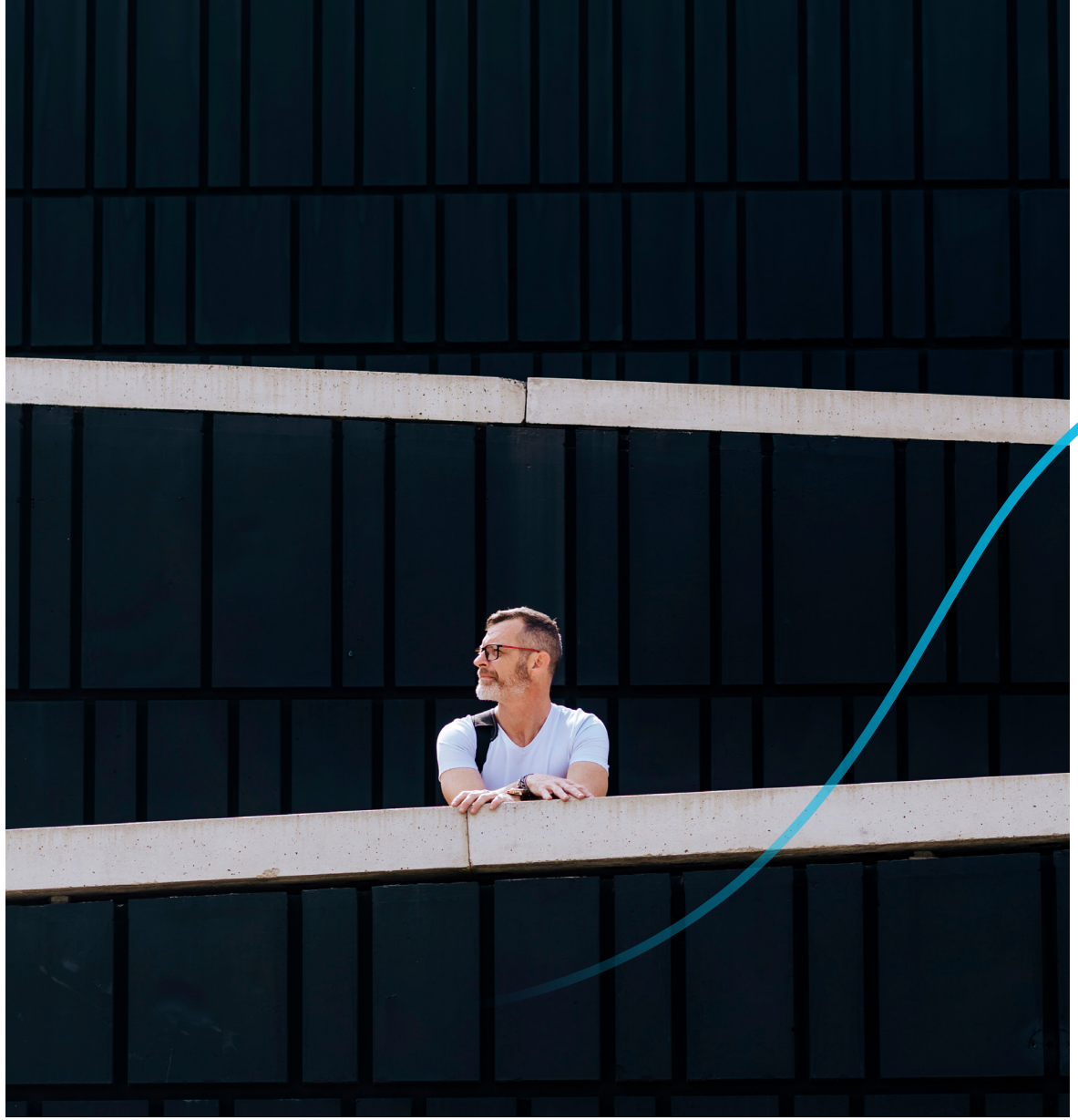
The promised land

The priority areas where CxOs see generative AI as having the most potential are:

1. IT development, by assisting with coding and testing in the software development lifecycle.
2. Sales and customer service, by optimizing sales support chatbots.
3. Product innovation and design, from creative brainstorming to faster drug discovery.
4. Marketing, to automate customer segmentation, or tailor content automatically to community profiles.
5. Manufacturing, using 3D modelling, or real-time QA process monitoring.
6. Operations and supply chain, by applying real-time analytics to logistics for optimization and regression.

How an organization sets the scale of its ambitions is an area where it could be at risk of making a misstep with Gen AI. It may have too broad a vision for how to bring it into their organization, opting for use cases beyond justifiable business need. Or it could have too narrow a vision, by not seeing the Gen AI picture in its entirety e.g. an engineering company that uses an LLM for content generation, but not for quality control.

The opportunities are there for the taking, but CxOs will each have questions about what generative AI will mean for their organization. CTOs and heads of legal and compliance may ask, “Can we trust it?” The CEO will ask, “How much it will cost and what will it add to our bottom line?” while a CFO might want to know what the risks are of unforeseen costs once the company is committed to transformation.



Trust above all

While traditional LLMs were for a specific use, their latest form offers one model to do many things. That's powerful, but if a model breaks, it affects multiple parts of a business, so effective governance is essential.

At the most fundamental level of an organization's generative AI, it must be robust, ethical, and reliable. Customers, clients, and regulators expect AI models to be fair, transparent, explainable, auditable, and free from bias. As customers become more aware of their digital rights, prioritizing ethics within generative AI becomes a competitive differentiator.

When generative AI breaks, as it can, its failure may not be obvious. It can give excellent or erroneous results (known as "hallucinations") with equal confidence. It is not linearly correct, and it is this variability in results that poses a unique, significant risk. When a model's adoption is scaled, an error multiplies exponentially. It is a high-risk, high-challenge technology that requires a well-structured, multi-dimensional framework that weaves in an integral trust layer.

Ethical data sourcing

At the heart of responsible, ethical generative AI use are the methods used to create LLMs. This means, how organizations source the datasets used to train them, and how they process the data of the companies and people who use them.

For example, in particularly sensitive situations, such as with medical data, a company may create a private LLM in a secure development environment, or use synthetic data. In other research contexts, "pseudonymization" is thought to be sufficient privacy protection for patient anonymity in data. In an AI environment without guardrails, it can be relatively easy to identify a prominent person, such as a political leader, by attaching details of their medical history with other publicly available information[8]. This necessitates the use of a secure development environment, and possibly synthetic data.

In a nutshell, guardrailing is about setting boundaries for an LLM. They are safety controls that monitor and dictate a user's interaction with an LLM application and should surround the entire LLM. They validate the inputs to an LLM — may a user make this request? — and vet them before further processing. Since we cannot be completely sure that the model's response will be secure and fully compliant with input constraints, there must be a guardrail at this point too. There are multiple possible versions of guardrails. Ethical guardrailing protects against unethical behavior such as bias towards certain groups. Topological guardrailing keeps the LLM on topic, where a business want to steer responses away from mentioning a competitor.

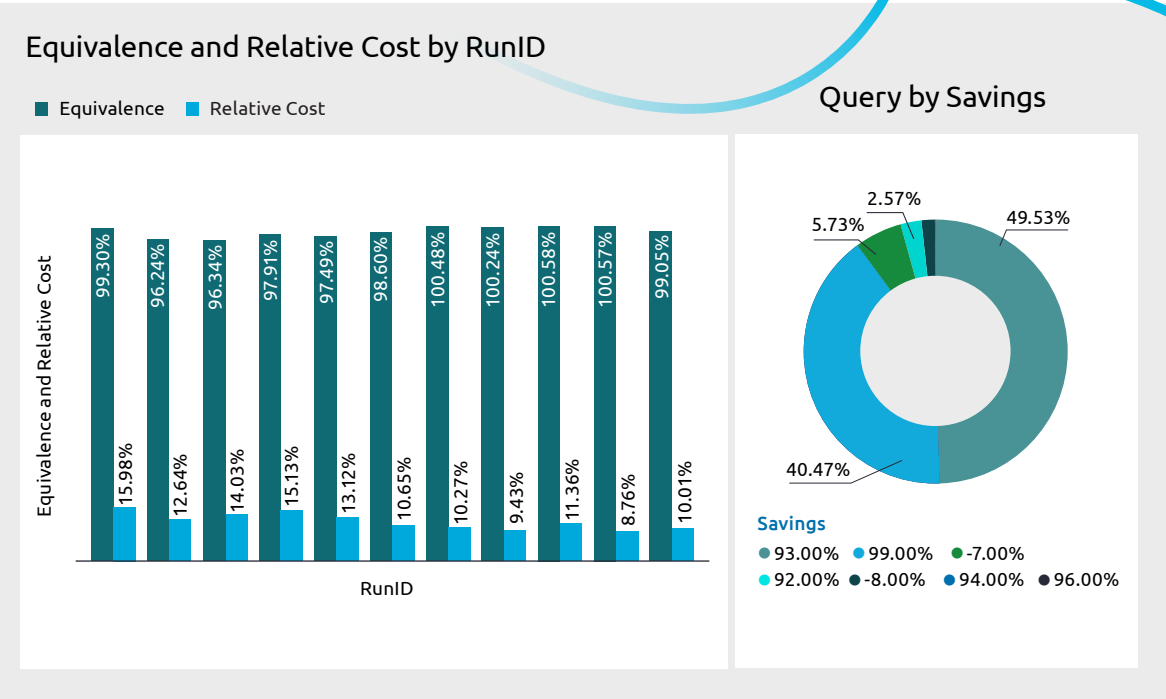
Some use of generative AI is more autonomous than legacy computing systems, requiring minimal human intervention. This, conversely, places greater importance on the role of human oversight as the last point of intervention before a generative AI decision takes effect. Errors could have serious consequences for customer trust, corporate reputation, or at the extreme, lead to systemic risk or loss of life.

8 BBC More or Less, The digital 'robots' unlocking medical data, 19th February 2024

Cost efficiency through LLM specialization

Generative LLMs are not all created equal, nor should they be if a business is to optimize costs. Using a cost management framework to choose specific, suitable LLMs, instead of going with a full-scale version by default, can generate a dramatic 60% cost saving. This saving could then be used for use cases elsewhere in a business, increasing further its total value. A full-scale LLM has its place for early-stage proof-of-concept. Smaller scale, lower cost LLMs, new versions of which are released frequently, can be combined in a cascade of LLMs. Each is chosen depending on the type of prompt to process and how it scores for fit and cost.

Other cost-saving tactics are available. A prompt-and-response caching solution checks for previous prompts with similar meaning and re-supplies the previous correct answer. A low-cost natural language processing solution re-writes prompts with more efficiency and effectiveness for the target LLMs.



Cascading LLMs: In testing on industry standard data, this approach demonstrated results equivalent to running a full-scale model at 96-100%, but at 9% to 16% of the original cost.

Adaptive scaling to realize the full opportunity

No matter how much drive or enthusiasm there is in the boardroom to adopt Gen AI, it will not deliver its true enterprise value without a strategy to scale. At this point in the technology's adoption, many companies across the global economy are at a relatively equal level of maturity with how they use open source and commercial LLMs. They have not yet found a fast, cost-optimized route to launch.

Early adopters have discovered that, if they did not begin with a vision for scale, the difficult reality of their approach dawned on them when it came to paying hyperscalers for the aggregated services. They have been left feeling the effect of expected business value evaporating.

Organizations must strike a balance between the opportunity for profitability from scaling, versus the costs of running these services. A properly established partnership with hyperscalers such as AWS, Microsoft Azure, or Google Cloud, is key to scaling successfully. This means being on a tailored and adaptative scale pathway that fits the organization, industry context and its specific business challenges.

The model process, according to transformation champions

A transformation plan that takes the above peer lessons onboard will identify the Gen AI business outcomes that matter most. It will then strategically prioritize investment and embed innovation into the delivery framework.

When an organization decides that generative AI has a role in its future, it typically starts with workshops for proof-of-concept (POC), strategy, operating model, and a use case roadmap.

There should be a measurement framework in tandem for transparent transformation management, including costs, with integrated processes and tools. For example, ongoing finetuning of foundation models will reduce computation costs and improve functionality.

Once in place, transformation champions continuously monitor execution, focusing on whether the investment is realizing its targeted value. It is orchestration that ensures that all activity, like costs and guardrailing, is captured and monitored consistently. It makes scaling possible, when allocating ressources can be challenging.

Planning for flexibility and modularity is also important and transformation leaders have the option to apply a multi-vendor approach where industry-specific applications are needed at the start or expected in the future. A modular approach makes possible the integration of existing infrastructure, such as data platform.

As a use case establishes its enterprise value, opportunities open up to scale across others. Doing this in the most economical way requires templates and frameworks, which also brings structural benefits, eliminating siloes which commonly lead to increased cybersecurity risk and cost inefficiencies. The alternative is a disconnected environment, which might deploy multiple dedicated graphics processor units for image generation beyond actual requirements. Or different applications could be using separate commercial application programming interfaces when these could be shared. And if content performance is not consistently measured, it is unlikely to be improved.

Use case: large retail bank

The client, a large retail bank, offered a customer-facing chatbot, but was unhappy with the maintenance it required and the lack of natural conversation the model was providing to customers. Customers also gave feedback indicating they would like a more naturally flowing dialog.

Challenges and pain points

The bank wanted to move to an LLM, but needed to train up the technical team on the services and applications most relevant to LLMs. They had an existing governance strategy and wanted to adapt it to include LLM governance.

Approach and solution

The project team began with a POC delivered via:

1. **A hackathon on application development with LLMs**
2. **A workshop on testing LLMs**
3. **Brainstorming sessions on governance, processes and budget management for a Gen AI project in this setting**

Once the POC stage was complete, we scaled up the LLM and developed a testing strategy and the chatbot's capabilities.

How Capgemini added value

The project focused on development of an LLM chatbot to increase user satisfaction and decrease maintenance and cost of a traditional chatbot. The proof-of-concept project phase had helped the client to understand the benefits and risks of LLM implementation. The governance and testing strategy, both part of the trusted AI offering, helped the client see how the model could be implemented with reliability. Adopting generative AI lead to improved trust in client chatbot communication; a reduction in overall testing time, cost and risk; and improved accountability and governance.

How to fast-track generative AI adoption in your organization

When properly implemented, generative AI gives businesses the ability to go further and faster with their data. It can transform organizations in many dimensions and deliver valuable, tangible benefits.

Capgemini has proven expertise in activating AI to its full potential for data-driven businesses. We offer an end-to-end operational accelerator for Gen AI — Capgemini RAISE (Reliable AI Solution Engineering). Capgemini RAISE is a powerful engine to make the most of foundation models in a secure environment with trust safeguards. It focuses on value use cases to scale custom Gen AI projects. Building on our partnerships with hyperscalers and AI startups, it enables a broad range of industry applications in many business domains.

About Capgemini

Capgemini is a global business and technology transformation partner, helping organizations to accelerate their dual transition to a digital and sustainable world, while creating tangible impact for enterprises and society. It is a responsible and diverse group of 340,000 team members in more than 50 countries. With its strong over 55-year heritage, Capgemini is trusted by its clients to unlock the value of technology to address the entire breadth of their business needs. It delivers end-to-end services and solutions leveraging strengths from strategy and design to engineering, all fueled by its market leading capabilities in AI, cloud and data, combined with its deep industry expertise and partner ecosystem. The Group reported 2023 global revenues of €22.5 billion.

Get the future you want | www.capgemini.com

Contacts

Mark Oost
Global Offer Leader AI, Analytics and Data Science
mark.oost@capgemini.com

Weiwei Feng
Tech & Solution Leader
weiwei.feng@capgemini.com

Bikash Dash
GTM & India Leader
bikash.ranjan-dash@capgemini.com