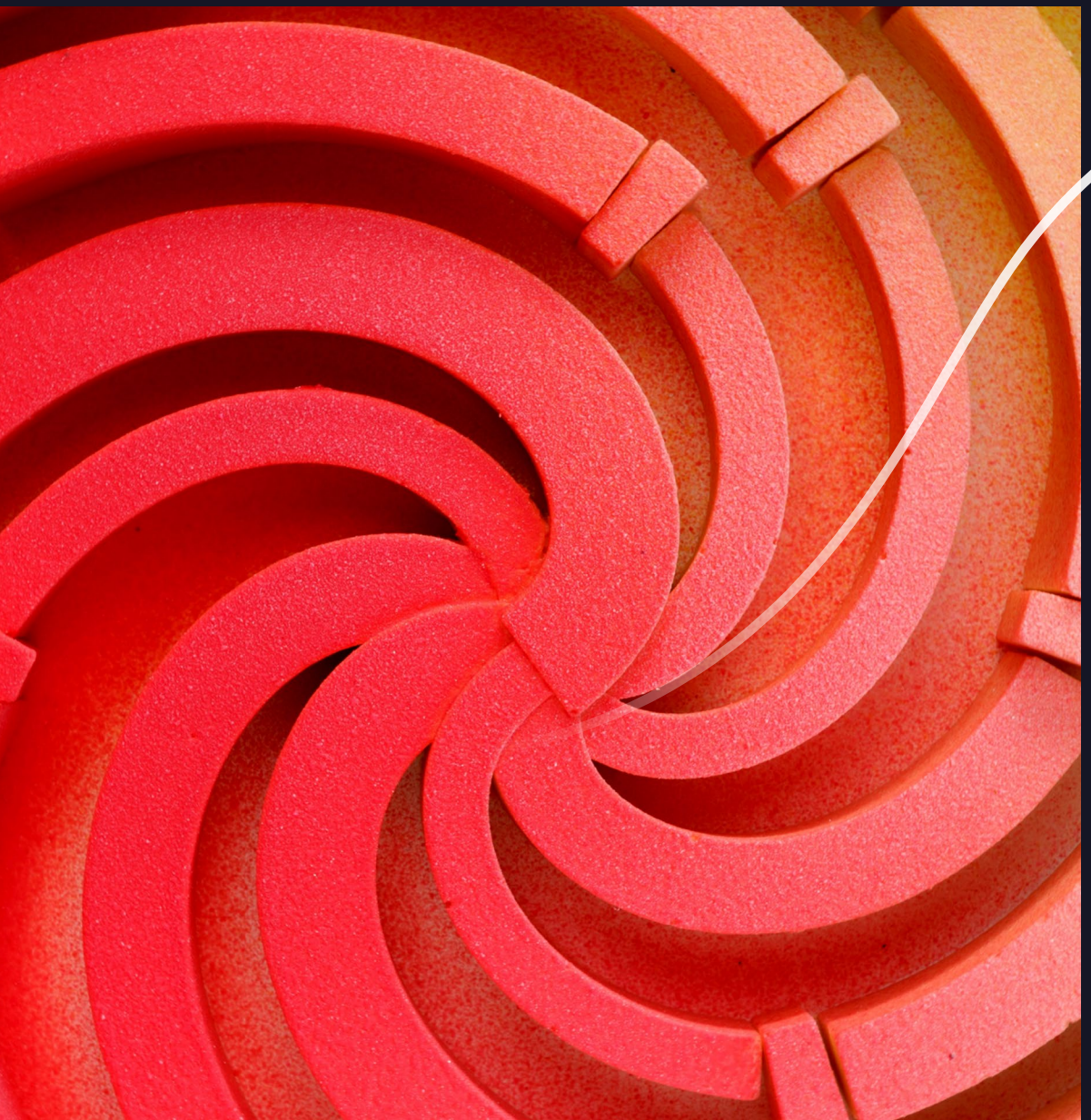


Enterprise-specific *AI agents* keep the Gen AI promise

Capgemini

Custom enterprise-specific Gen AI makes it possible to reach business outcomes faster with more insightful and original output. It can be the most effective way for a company to maximize the value of its own structured and unstructured data.

The awarding of the 2024 Nobel Science Prizes in physics and chemistry for artificial intelligence-related discoveries confirmed its significance for modern life. According to Nobel laureate Geoffrey Hinton, one aspect of AI related to generative AI - machine learning - will "have a huge influence. It will be comparable with the Industrial Revolution, but instead of exceeding people in physical strength, it's going to exceed people in intellectual ability."[1]

Nevertheless, generative AI has not been in popular use for long, with some well-publicized instances of unintended generative AI mistakes. Consequently, organizations are cautiously optimistic about the potential benefits of integrating it into their strategies, processes, and business models.[2]

While large learning models (LLMs) hold potential to re-shape business, they run into a performance ceiling when dealing with specialist areas that they have not been trained on. Users are cautioned not to trust their responses outright, especially for questions involving unique, organization-specific information. Simply training them with more public data and adding processing power is not enough to deliver domain-specific improvements and the expansion of capabilities needed to justify investment. They lack the specialized, high-quality data needed to operate with expertise in a company's specialism. This data is locked behind corporate firewalls, or otherwise unavailable for generalist LLM training. As a result, even top-tier chatbots like ChatGPT or Gemini, which are based on LLMs, can produce flawed answers, or "hallucinations".

Accordingly, there are sound reasons to view the generative AI phenomenon with some detachment, but its fundamental benefits to enterprise have not yet been widely or fully tapped in the economy. Enterprise-specific, i.e., custom private Gen AI, also known as agentic AI, can be improved to achieve higher accuracy using transparent, accessible sources and verifiable results. This form has an enhanced ability to focus on specific areas and produce context-sensitive results through training on refined data from specific business domains. Agentic implementation allows expanded use cases in more complex and sophisticated scenarios.

---

[1] *Babbage from The Economist,* The 2024 Nobel prizes: a triumph for AI, *Oct. 9, 2024*
[2] *Capgemini Research Institute* Harnessing the value of generative AI, *p. 27*

Specialize for better outcomes

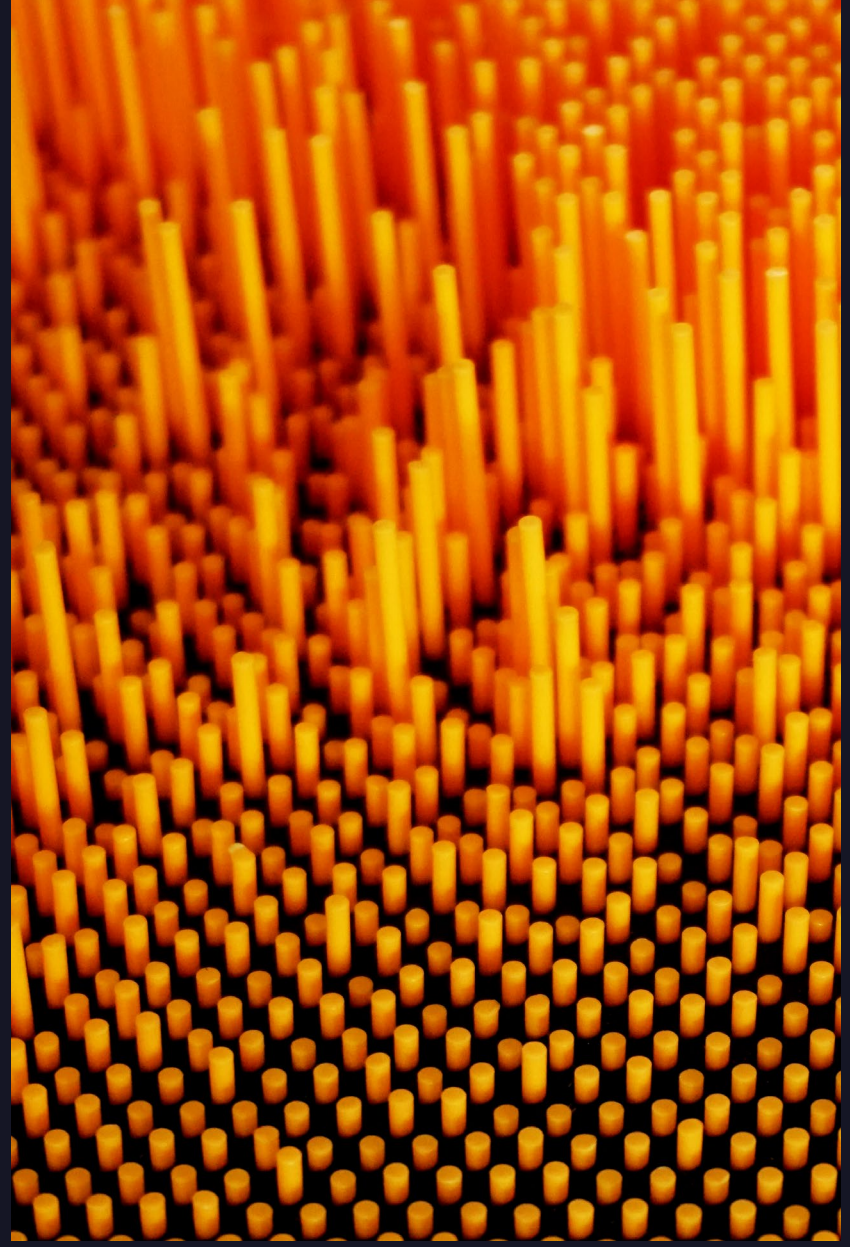
General purpose LLMs can be effective tools to increase productivity, but their lack of specificity reduces the relevance and quality of their output. They tend to produce generic results that do not deliver the full potential achievable from an organization's own intellectual property and data, which should lead to actionable insights and productivity gains.

Innovation that sets a company apart is most likely to come from an organization's own proprietary insights. According to business intelligence provider, Gartner, 68% of enterprises struggle to integrate AI into workflows that rely heavily on internal data.[3]

Custom private Gen AI makes it possible to respond faster to business opportunities, with more insightful and original content. It can be the most effective way for a company to maximize the value of its IP in various forms. Without the mass data interpretive power of Gen AI, much of this will lie dormant, siloed and forgotten. Meanwhile, the company reinvents the wheel, wastes time and budget, and misses opportunities.

To achieve generative AI's maximum benefit to enterprise productivity and creativity, a custom-designed language model can be trained on an organization's own dataset in its entirety using structured and unstructured data. This form of enterprise AI agent has a higher level of reasoning, producing more refined responses on specialist subjects.

The ideal generative AI model boosts productivity across the entire workforce. For instance, it can streamline the process a business development team follows to respond to RFPs, speed up selection of qualified talent for HR, or enhance the quality of customer interactions in a contact center.



[3] *Gartner,* **RAG in enterprise data strategy**

Safeguarding *Gen AI* data

Along with making Gen AI a more effective business resource, there are significant ethical and compliance reasons to prefer a custom private Gen AI system. Its closed loop structure allows tighter governance, security protocols, and continuous monitoring to keep it in line with a company's own ethical and security guidelines, and in compliance with data privacy and sovereignty, where applicable.

Knowing the methods used for data training are key to managing legal risks. These risks are real, with credible copyright and trademark infringement cases already underway.[4] Proprietary LLMs are safer due to their closely controlled, transparently-sourced training data.

By striking a balance between robust security controls and proactive safeguards for ethical AI performance, organizations can protect their critical assets while cultivating trust and strengthening operational resilience.

Data security has climbed higher up the board agenda due to tighter operational resilience regulation. A custom private Gen AI assistant allows companies to impose strict data security measures, retain full ownership of their intellectual property, and gain clearer insights into potential vulnerabilities.

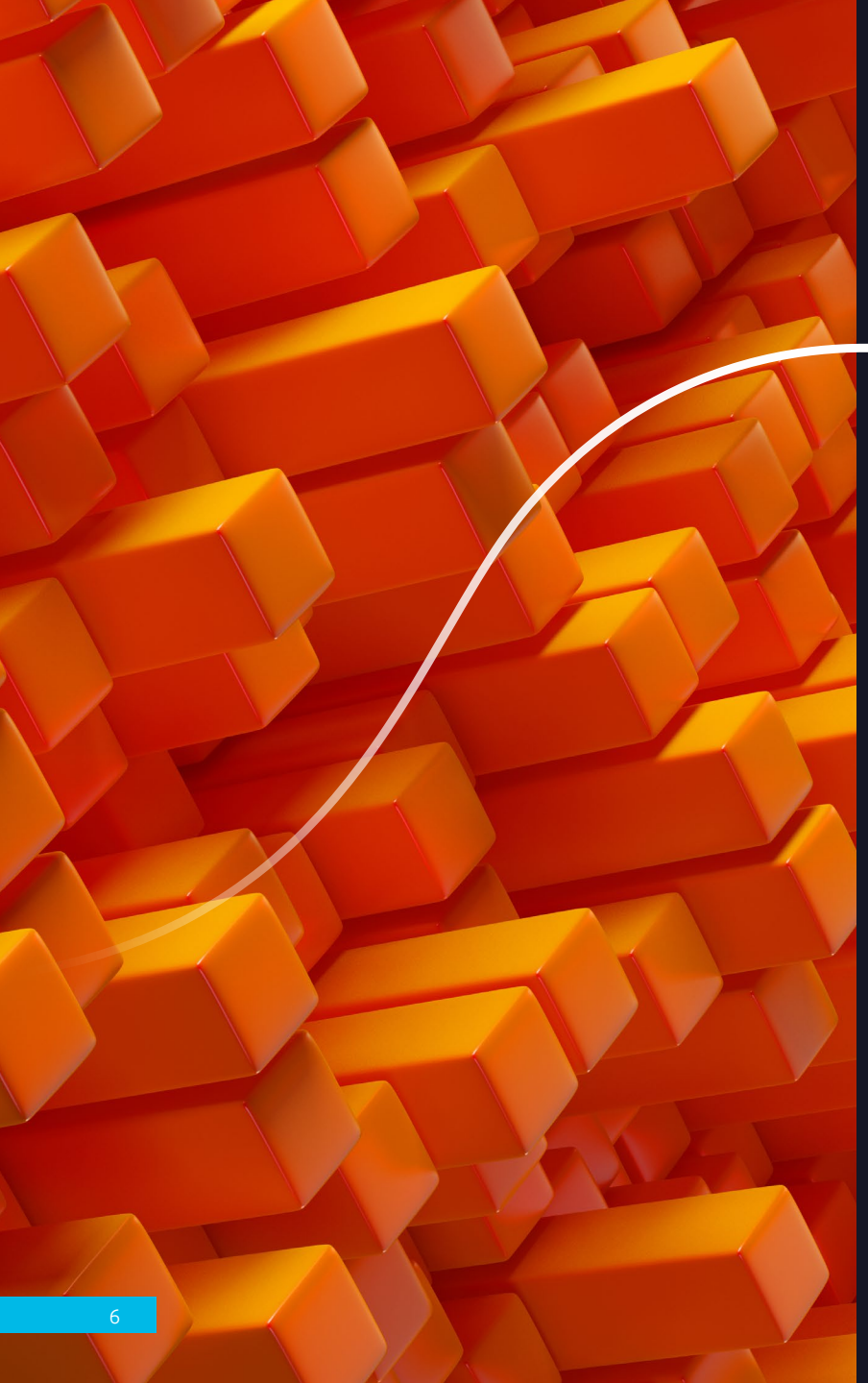
[4] *Reuters,* **AI companies lose bid to dismiss parts of visual artists' copyright case,** *August 13, 2024*

The *data ingestion framework* - the key to delivering full Gen AI value

The decisive component of a custom private LLM is its Data Ingestion Framework (DIF). The DIF is crucial for analysis of and attributing meaning to documents and other materials in a dataset and thus its productiveness. It extracts, organizes, and prepares data for future retrieval. It applies metadata for ontological purposes, ensuring that the model can access the right information at the right time. The aim is for better targeting of the required information for a query, making knowledge management and domain-specific assistance more productive. Queries are thus answered with data from the correct documents and the most applicable sections, a critical capability for domain-specific use cases and effective knowledge management.

Supporting a trained LLM with customized retrieval-augmented generation (RAG) improves the accuracy of responses. Where the LLM is an AI system trained on a large, known dataset, RAG is an information retrieval system that picks out specific, relevant data from available databases and documentation. An LLM is static, while the RAG model is dynamic. The LLM's strength is in specific, well-defined scenarios while RAG is capable of broad matching with up-to-date information, but with less precision, making guardrails critically important. [5]

[5] *Towards Data Science,* **The Practical Limitations and Advantages of RAG,** *April 15, 2024*



Multi-stage *guardrails*

RAG systems rely on guardrails at multiple stages to ensure quality and mitigate risks. Guardrails also keep Gen AI systems in alignment with organizational values by eliminating harmful and biased outputs.

The first layer of control begins with input filtering, which examines the specific content of user prompts for compliance and risk factors. For instance, a query requesting sensitive strategies or methodologies, e.g., how to redesign internal processes, could inadvertently expose confidential information. This is particularly important in regulated sectors like banking, healthcare, and defense, where strict compliance standards prohibit sharing sensitive data with external AI systems or vendors.

Intermediate guardrails then act as checkpoints during the retrieval process, validating selected data for alignment with policies and the context of the query.

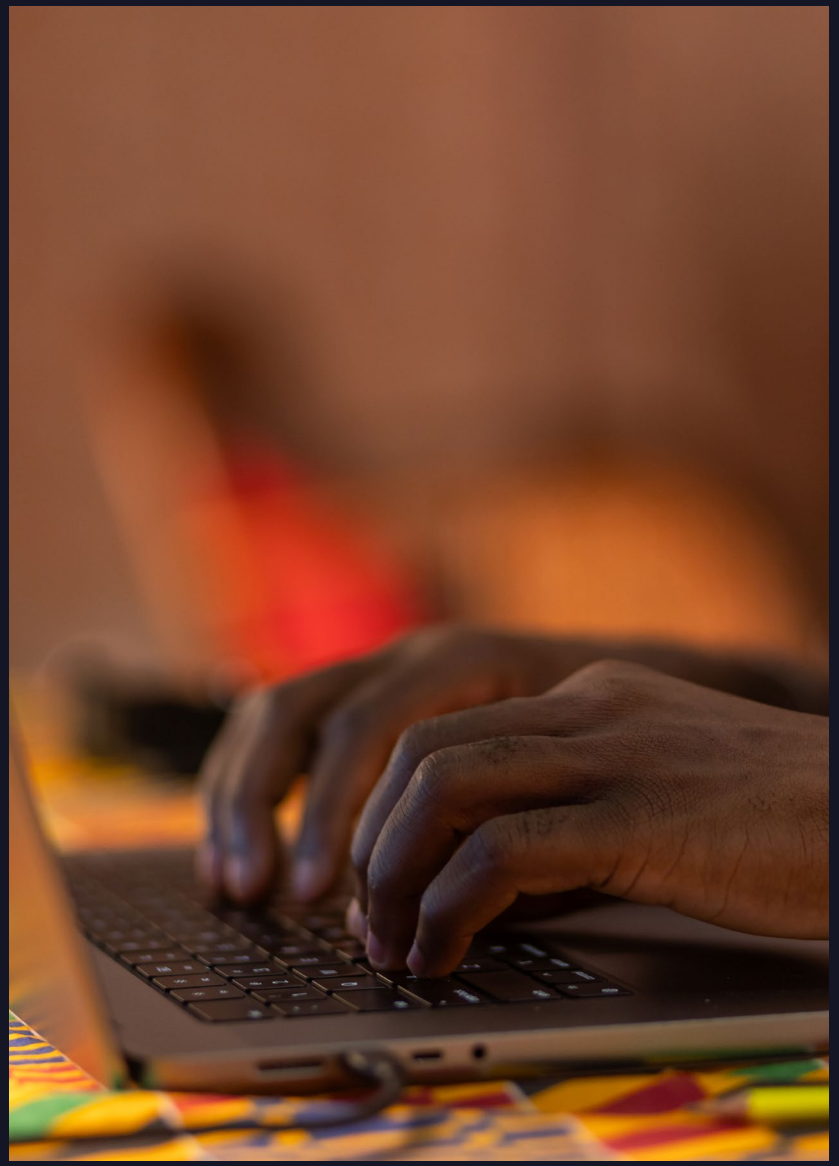
A final output filter evaluates the response once the information is retrieved, but before it is delivered to the user or downstream business processes. This filter ensures that the response adheres to key requirements, such as confidentiality, appropriateness, e.g., absence of toxic or harmful content, and compliance with company policies and regulatory standards. It also confirms that the output meets user expectations of accuracy and relevance, serving as a final quality control step.

Cost controls

The cost of running LLMs can spiral if left unchecked due to their reliance on intensive computing resources and token-based pricing models. This is directly tied to the volume of prompts processed and the length of generated responses. This makes extensive or complex queries exponentially expensive. 65% of enterprises in 2024 reported difficulty in forecasting LLM usage costs, with some companies experiencing over 30% unanticipated budget overruns due to insufficient monitoring and control.[6]

A major benefit of using RAG with custom LLMs is the ability to shift computational burden to retrieval mechanisms, which are cheaper to run. RAG frameworks retrieve enterprise data in real-time, allowing LLMs to focus solely on generating summaries or context-specific outputs rather than processing entire datasets. This hybrid approach can cut compute expenses by 25–35%.[7]

Cost control measures in custom LLMs integrated with RAG systems with token usage monitoring enable enterprises to harness the power of generative AI without financial surprises. By focusing on task-specific applications and streamlined workflows, organizations can better forecast expenses and align their AI investment with business priorities.



[6] *Gartner,* Enterprise AI cost control report
[7] *Forrester,* Build Efficient And Robust GenAI Apps With Prompt Engineering And Advanced LLM App Architectures

Examples of Gen AI solutions using custom private generative AI agents tailored to domain and business function

Finance

- Automating data analysis across financial reporting, marketing, operations, supply chains
- Proactive data-driven recommendations to reduce manual analysis time

Banking and insurance

- Credit memo generation
- Covenant monitoring
- Suspicious Activity Report and other financial crime report filings
- Pitchbook generation
- Claims processing, e.g, legal case package creation
- Underwriting assistant

Healthcare providers

- Drafting longitudinal patient summaries
- Clinical trial package generation
- Personalized medicine
- Research assistant agent

Sales

- Identifying cross-selling opportunities by using a Gen AI agent as a portfolio sales executive
- Generating first-shot RFP/RFQ responses
- Customer intent/ insights agent

Research and development

- Synthetic data generation
- Advanced drug discovery and therapeutics
- Novel protein design
- Clinical trial and research facilitation

Use case - Streamlining an *insurance underwriting* workflow

A market-leading US insurance underwriter saw the opportunity to use generative AI to improve underwriting workflows. Their implementation of a custom private generative AI model analyzed historical claims, policy data, and external risk factors to draft underwriting recommendations. It also generated detailed explanations for its recommendations, enabling underwriters to then make informed decisions faster.

This reduced underwriting case turnaround times by 40%, improved risk assessment accuracy, and supported the creation of personalized policies. This ultimately enhanced customer satisfaction and profitability.

Our *experience*

Capgemini has proven expertise in activating data and AI to its full potential for data-driven businesses. Building on our partnerships with hyperscalers and the AI innovation ecosystem, we help our clients deliver value and generate competitive advantage with a portfolio of tailored, scalable Gen AI solutions.

We help maximize the value of your enterprise data by creating accurate, context-aware Gen AI assistants. They empower your employees and customers using your own data for specific business needs, while safeguarding data. These agents are typically used to streamline customer service, marketing, contract management, content generation, financial analysis, and more, at controlled cost of use.

Experts to contact

Pinaki Bhagat

AI & Generative AI Solution Leader,
Financial Services

Ashvin Parmar

Vice President, Generative AI CoE Leader,
Financial Services

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