Banking on Generative AI: Maximizing the Financial Services Opportunity

Banking on Generative AI: Maximizing the Financial Services Opportunity

Technological change, like many evolutions, often happens slowly, and then all at once. Neural networks were conceived in the 1940s and natural language processing algorithms came 20 years later, when the first chatbot, Eliza, was created. There followed nearly 60 years of gradual development, until massive computing power helped create the models that have attracted global headlines over recent months.

In the financial industry, experimentation with natural language processing has also progressed in steps, with banks incrementally adding functionality such as chatbots and automated document processing. Fast forward to today. Massive computing power has helped accelerate experimentation and create generative AI (GenAI) models that are set to be game-changing for financial services.

GenAI is a catch-all term for a range of models, predominant among which are foundation models (FMs) and their large language models (LLMs) subset. (See "What Is Generative AI?")

WHAT IS GENERATIVE AI?

Generative AI (GenAI) is a set of algorithms, capable of generating seemingly new, realistic content from unstructured inputs such as text, images, or audio.

The term GenAI encompasses both foundation models and large language models:
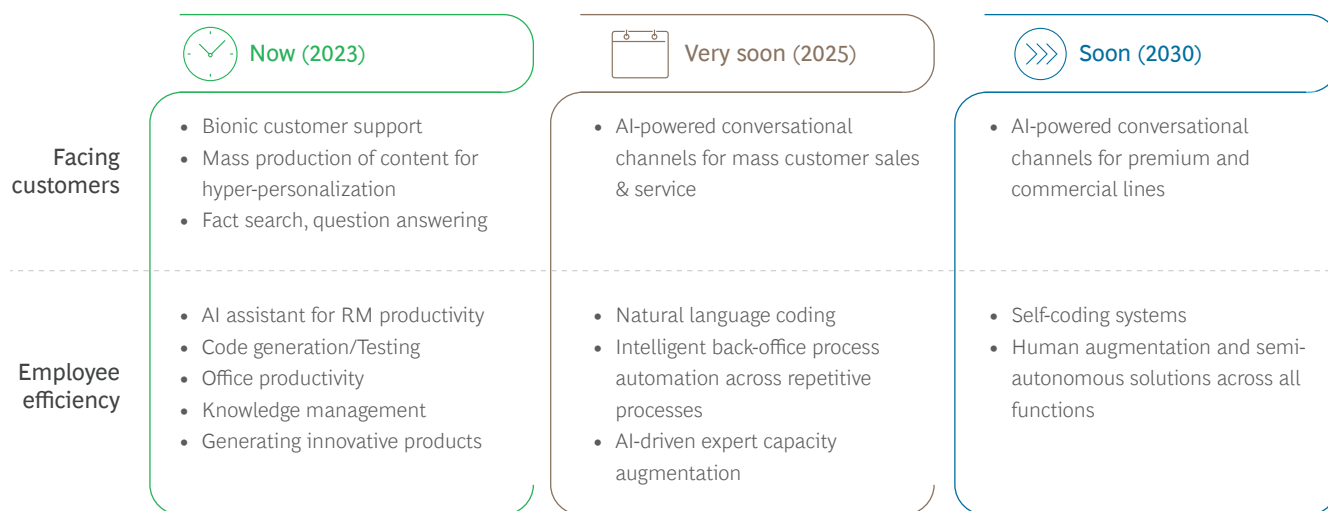
Foundation models are pretrained with large datasets and massive compute power, so they are ready to be used without additional training. They can be applied to many tasks (unlike traditional AI), including generating text or graphics, predicting, or classifying.

Large language models are a subset of FMs and can ingest and produce text. The terms LLM and FM are not interchangeable.

The promise of FMs is that they present shortcuts to resolving complex challenges. Their potential puts them on a different plane to other technologies that have been labelled game-changing over the past decade. However, they also bring risks, including leakage of confidential data, transmission of bias, and potential for low-quality outputs.

In the banking industry, potential FM use cases range from replacing mundane manual processes to revolutionizing digital customer interactions, creating highly personalized marketing content, and supporting investment decisions. If applied effectively, we expect these kinds of applications will boost cost effectiveness on a double-digit scale while significantly improving customer outcomes. Based on our experience working with clients, there is also potential to reduce back- and front-office process costs by 20-30% in the next two to three years and 50% or more in the following three years—freeing up staff to engage in more valuable and less repetitive activities. (See Exhibit 1.)

Exhibit 1 - Gen AI adoption is picking-up rapidly

	Now (2023)	Very soon (2025)	Soon (2030)
Facing customers	• Bionic customer support • Mass production of content for hyper-personalization • Fact search, question answering	• AI-powered conversational channels for mass customer sales & service	• AI-powered conversational channels for premium and commercial lines
Employee efficiency	• AI assistant for RM productivity • Code generation/Testing • Office productivity • Knowledge management • Generating innovative products	• Natural language coding • Intelligent back-office process automation across repetitive processes • AI-driven expert capacity augmentation	• Self-coding systems • Human augmentation and semi-autonomous solutions across all functions

Source: BCG.
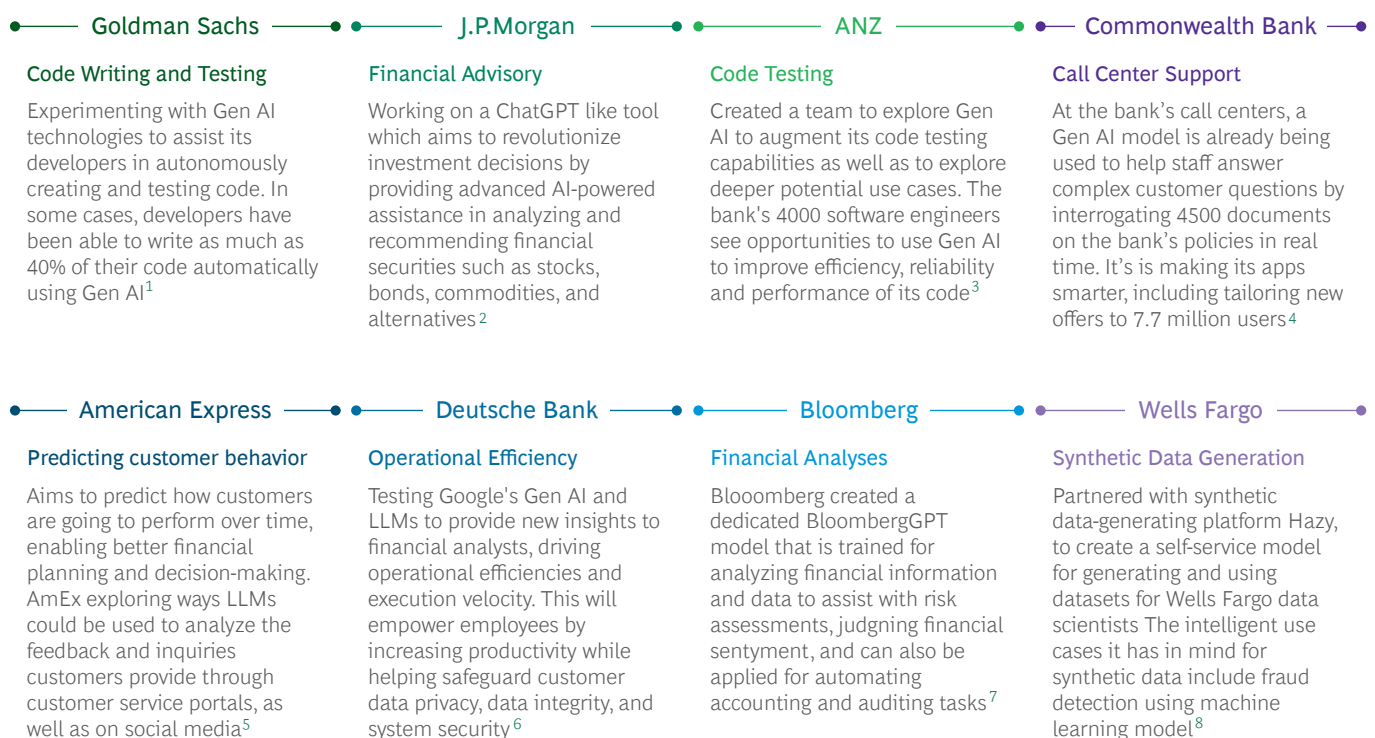
Banks Are Investing in New Use Cases

Amid a growing competition, the largest banks are investing heavily in FMs, leveraging external models and open-source offerings, such as Langchain—an emerging standard component also used by Google. Use cases range from more generic processes such as code writing and call center support to those specific to the financial services industry, such as financial analyses or delivering financial advice. (See Exhibit 2.)

Leading IT solution vendors, including Microsoft, Salesforce, Pega, and ServiceNow, have created FM roadmaps and are offering new functionalities and products leveraging their own or third-party FMs. In parallel cloud vendors are unveiling environments for model fine-tuning and developing functionalities that will facilitate use cases. Cloud vendors are offering both proprietary FMs and FMs obtained through partnerships. Microsoft Azure is teaming up with OpenAI, Google Cloud Platform has unveiled its proprietary PaLM2, and Amazon Web Services offers both startup FMs (AI21 Labs, Anthropic, Stability AI) and its own Titan FM.

The power of FMs is centered on four key areas, each of which has applications in more than one area of banking:

- **Summarization** (including answering questions, fact finding) refers to the ability to ingest unstructured text and summarize it to the required level of detail. This may be used to boost customer service efficiency, identify customer needs and preferences from documents such as emails, create risk profiles from uploaded files, or document software source code.

- **Content generation** is the ability to elaborate on a given topic or create media content such as images or videos given succinct inputs. This can be used in customer communications or to generate software source code, among many other examples.

Exhibit 2 - Leading FI players are announcing GenAI use cases every day

Goldman Sachs

Code Writing and Testing

Experimenting with Gen AI technologies to assist its developers in autonomously creating and testing code. In some cases, developers have been able to write as much as 40% of their code automatically using Gen AI[1]

J.P.Morgan

Financial Advisory

Working on a ChatGPT like tool which aims to revolutionize investment decisions by providing advanced AI-powered assistance in analyzing and recommending financial securities such as stocks, bonds, commodities, and alternatives [2]

ANZ

Code Testing

Created a team to explore Gen AI to augment its code testing capabilities as well as to explore deeper potential use cases. The bank's 4000 software engineers see opportunities to use Gen AI to improve efficiency, reliability and performance of its code[3]

Commonwealth Bank

Call Center Support

At the bank's call centers, a Gen AI model is already being used to help staff answer complex customer questions by interrogating 4500 documents on the bank's policies in real time. It's is making its apps smarter, including tailoring new offers to 7.7 million users[4]

American Express

Predicting customer behavior

Aims to predict how customers are going to perform over time, enabling better financial planning and decision-making. AmEx exploring ways LLMs could be used to analyze the feedback and inquiries customers provide through customer service portals, as well as on social media[5]

Deutsche Bank

Operational Efficiency

Testing Google's Gen AI and LLMs to provide new insights to financial analysts, driving operational efficiencies and execution velocity. This will empower employees by increasing productivity while helping safeguard customer data privacy, data integrity, and system security[6]

Bloomberg

Financial Analyses

Blooomberg created a dedicated BloombergGPT model that is trained for analyzing financial information and data to assist with risk assessments, judging financial sentyment, and can also be applied for automating accounting and auditing tasks[7]

Wells Fargo

Synthetic Data Generation

Partnered with synthetic data-generating platform Hazy, to create a self-service model for generating and using datasets for Wells Fargo data scientists The intelligent use cases it has in mind for synthetic data include fraud detection using machine learning model[8]

Source: BCG.

- **Conversation** is the ability to provide human-like answers to customer prompts while considering the context and flow.

- **Data generation** (also provided by "small" and non-text generative AI models) is the ability to generate complex sets of data with specific characteristics. This is intended predominantly for test cases to compensate for small numbers of empirical observations, ranging from fabricated fingerprints to biometric identification, or realistic dialogues with customers to test chatbots.

These capabilities can be applied along the banking value chain, accelerating development of (and enriching) sales and marketing communications (such as in call centers), increasing sales and service efficiency and availability, and lifting software development productivity. (See Exhibit 3.) Still, for many financial institutions, the first step in using FMs is to support employees, for example in finding the right answer to an incoming query. These use cases are low-hanging fruit—which we estimate, based on client engagements, can deliver 10–20% effectiveness gains.

Exhibit 3 - Generative AI capabilities can be applied across banking value chain

Non-exhaustive

I. Marketing & Sales	II. Distribution & Onboarding	III. Product development	IV. Financial Advice	V. Servicing	VI. Risk & compliance	VII. Supporting corporate functions
Copywriting & creating visual collaterals	RM productivity: preparing for meetings	Generating creative/innovative products & features	Generate content for financial education	Composing personalized emails from RMs	Compliance monitoring and report generation	IT Code documenting/ generation/ review
Mass production of content for hyper-personalization	Credit approval: support/ automation	Contract & term generation	Composing personalized emails	Intelligent document processing	Data privacy & compliance checks	IT: Test case generation
Customer: Product search, fact search	Loan & other products application assistance	Configuring/ coding products in systems	Detecting trends and scenario/ portfolio optimization	Credit review support/ automation	Fraud detection with synthetic transactional data	HR: Copywriting recruiting/ employer branding content
Chatbots/ voicebots for lead warming & conversion	Individualized contract & term generation		RM productivity: rading gist of memos, fin. reports, interacting with analytics	Virtual assistant/ service chat/ voicebots		Memo writing
Sales trainings with simulated client conversations						Strategy: Competitor analytics
Data augmentation for model training						HR: Screening CVs
						IT Support: Knowledge base search
						IT: Test data generation

☐ Content generation   ☐ Synthetizing, question answering   ☐ Conversational interfaces   ☐ Data generation
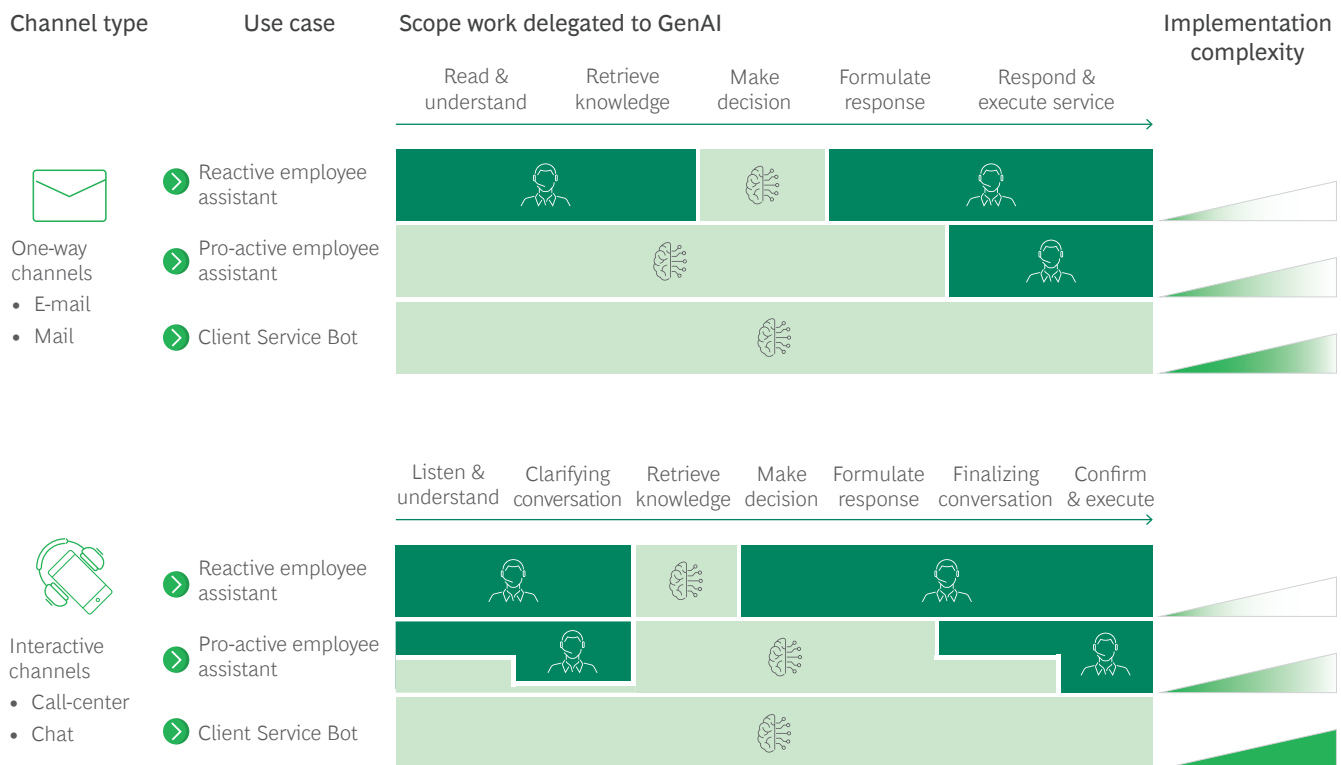
Source: BCG.

The next obvious step will be to directly read or listen to customer communications and recommend answers, while still relying on employees to lead the conversation and make final checks. These are more challenging tasks, requiring solutions that can get close to creating customer-ready outputs, and potentially leading to 20–50% productivity improvements.

The longer-term vision would be for FM-based agents to take responsibility for tasks end-to-end. This would include back and forth dialogues with customers—within clear guardrails and topic areas—and executing operations directly in internal IT systems. (See Exhibit 4.) In this context, there would be significant upsides in redeploying full-time employees to client-facing and higher-value work.

Exhibit 4 - Use-cases with increasing responsibility entrusted to FM-based solutions
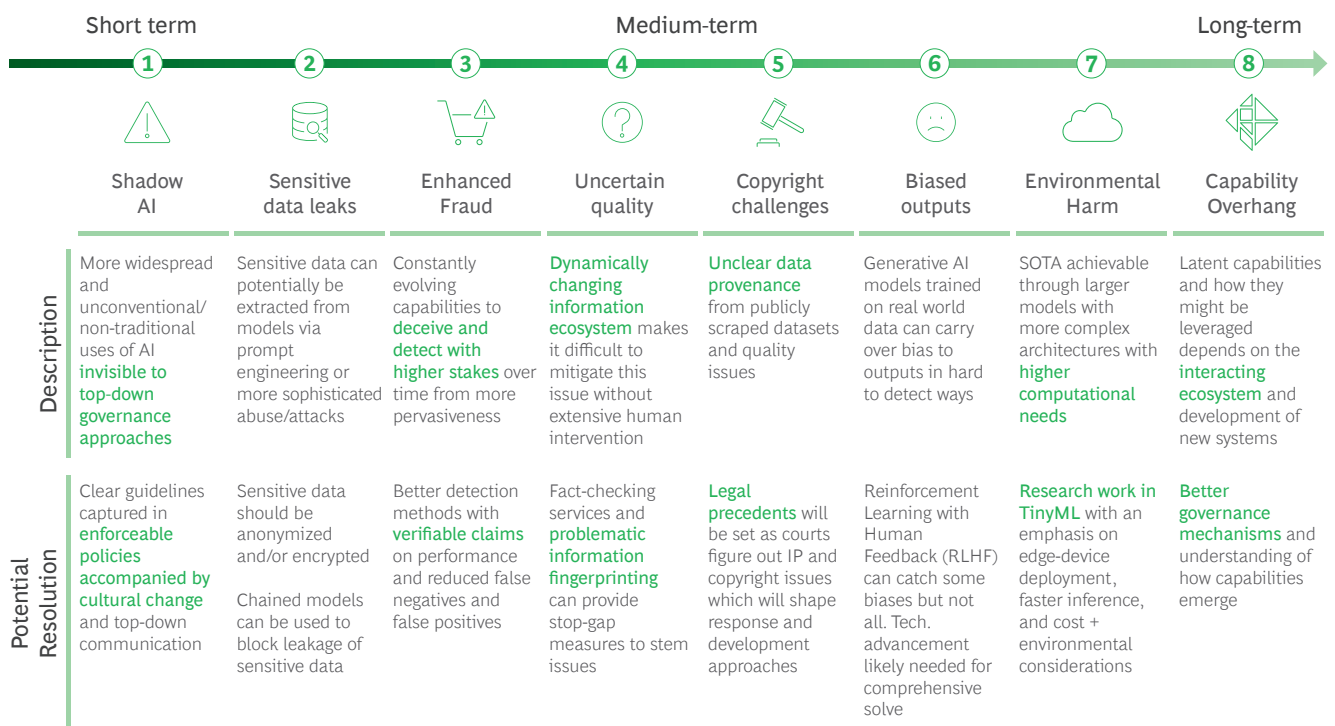

Source: BCG.

Risks Must Be Managed

The flipside of FMs' unique capabilities is the potential for new or exacerbated risks. Therefore, banks must take care in their use, including protecting themselves against malicious actors. Failures to do so will expose them to both potential regulatory sanction and reputational damage.

Risk that may play out in the short term include the emergence of shadow AI, or applications that are not overseen or managed by the organization, leading to potential compliance issues. Another short-term risk is leakage of confidential data, either through prompt engineering or through vendors using prompts for model training. To protect stakeholders, sensitive data can be encrypted, anonymized, or protected through permissions processes, but leakage can still occur through ungoverned usage. (See Exhibit 5.)

Looking ahead, a risk especially relevant to the financial industry relates to the potential use of FMs for fraud. The models may, for example, help criminal actors assume identities, forge checks, impersonate customers through voice imitation, or even mimic managers to persuade employees to bypass security measures (social engineering). There are also potential legal pitfalls, including infringements of intellectual property rights, for example when generating graphics that resemble copyrighted work.

Exhibit 5 - Gen AI amplifies existing AI risks and introduces new ones – which can be resolved over time

	Short term		Medium-term					Long-term
	1 Shadow AI	**2** Sensitive data leaks	**3** Enhanced Fraud	**4** Uncertain quality	**5** Copyright challenges	**6** Biased outputs	**7** Environmental Harm	**8** Capability Overhang
Description	More widespread and unconventional/ non-traditional uses of AI invisible to top-down governance approaches	Sensitive data can potentially be extracted from models via prompt engineering or more sophisticated abuse/attacks	Constantly evolving capabilities to deceive and detect with higher stakes over time from more pervasiveness	Dynamically changing information ecosystem makes it difficult to mitigate this issue without extensive human intervention	Unclear data provenance from publicly scraped datasets and quality issues	Generative AI models trained on real world data can carry over bias to outputs in hard to detect ways	SOTA achievable through larger models with more complex architectures with higher computational needs	Latent capabilities and how they might be leveraged depends on the interacting ecosystem and development of new systems
Potential Resolution	Clear guidelines captured in enforceable policies accompanied by cultural change and top-down communication	Sensitive data should be anonymized and/or encrypted

Chained models can be used to block leakage of sensitive data | Better detection methods with verifiable claims on performance and reduced false negatives and false positives | Fact-checking services and problematic information fingerprinting can provide stop-gap measures to stem issues | Legal precedents will be set as courts figure out IP and copyright issues which will shape response and development approaches | Reinforcement Learning with Human Feedback (RLHF) can catch some biases but not all. Tech. advancement likely needed for comprehensive solve | Research work in TinyML with an emphasis on edge-device deployment, faster inference, and cost + environmental considerations | Better governance mechanisms and understanding of how capabilities emerge |

Source: BCG.

Spontaneous or underinformed use of FM-based functionalities may lead to unexpected results, including generation of so-called hallucinations. These are responses not justified by training data, which are often the result of the system misunderstanding the question. Similar risks could be manifested through use case design flaws, leading to uncertain or volatile outputs. Finally, the model may produce biased opinions, reflecting similar biases in training data. Potential antidotes include deeper fact-checking and information finger printing.

On a strategic horizon, potential risks relate to the evolution of business ecosystems, where business models and functionalities based on browsing the internet may cease, with pages replaced by conversational interfaces or on-the-fly generated content. Stakeholders will need to keep a close eye on how capabilities develop and put in place governance frameworks that are sufficiently flexible to keep pace.

Key Design Variables

Not all FMs are created equal, and financial institutions will need to make a range of selections depending on the intended use case, hosting preference, memory configuration, and other factors.

Initial applications often retrofit FMs into existing roles, adding them to tasks like co-authoring marketing messages, coding chatbot flows, or converting text to graphics. This approach allows users to do the same things but faster. Looking ahead, there is a strong argument for setting the bar higher. This would mean adopting a clean-sheet approach that puts the technology at the heart of the design process Potential examples include:

- Mass produce marketing content based on product characteristics and personas in the customer base, instead of manually describing each picture's details.

- A chatbot delivering human-like interaction, able to handle multithreaded, nonlinear conversations and accommodate new services without the need to code and maintain detailed conversation flows.

- Recommendations for lending and investment decisions based on a blend of structured and unstructured data, including detailed justifications for internal and external purposes.

Decisions regarding hosting of models should be aligned with the type of model employed. (See Exhibit 6.) General-purpose and specialized models (ChatGPT or OpenAI's Code DaVinci for coding) are usually adopted as a cloud-based service and accessed via secure API, under the guarantee that submitted "prompts" will not be used to train the model, cutting the risk of data leakage. For dedicated adaptors, which facilitate customization, or fine-tuned models, the range of options includes training and running by a vendor or internal data science and IT team (often within a cloud tenancy).

Exhibit 6 - Multiple design choices need to be taken for each use-case

GenAI solution layers	Design options						Tech components *Non-exhaustive examples*

Logical design & Tech Stack

Environment integration	Chat & voice interfaces/ channels	Data access	Executing actions in banking systems (RPA, APIs)				• Accessing CRM Customer 360, executing operations in Core Banking System or Card System • Technologies: Specific to bank's IT architecture
Prompting logic	Direct	Templated	Multi-step chains	Algorithmic	Agent-based		• Multi-step: Product-emotions-pictures • Agent based, eg. Intent-resolution-execution • Technologies: LangChain, LMQL, Microsoft Guidance
Short-term memory	None	Conversation history	Purpose-structured memory				• Buffers & structures with e.g.: task list of chatbot, working summaries of document parts • Technologies: LangChain, Meta FAISS
Long-term memory	None	Similarity-based retrieval	Search-engine based				• Vector stores containing e.g.: product information, processes, regulations, FAQs • Technologies: Chroma, Llamaindex, Pinecone

Model & Data

LLM tailoring	General-purpose	Specialized	With adaptor	Fine-tuned	Custom-trained	Custom-built	• General: ChatGPT, Google PaLM, A21, Aleph Alpha • Specialized: Google MedPaLM, Code DaVinci • Custom: BloombergGPT • Model tuning: Google Model Garden, AWS Bedrock
LLM sourcing & hosting	Proprietary		Opensource				• Proprietary models: ChatGPT3, Google PaLM2 • Opensource models: Vicuna, ChatGPT2 • Model hosting: Google Model Garden, AWS Bedrock • Model repositories: HuggingFace
	Embedded	Public	Managed	Self managed			

No PII or banking information	Sensitive data & systems	Chatbot logic and short-term memory

Source: BCG.

Other design choices that will shape the models' impact include:

- **Prompting logic:** How will questions to the FM be prepared? Options include directly passing user requests, building an autonomous agent, or using tools such as internet search or mathematical calculators.

- **Short-term memory:** How will the "conversation notepad" be organized? Will it be a "raw" history of conversation or structured to perform a specific task, for example a list of outstanding customer requests.

- **Long-term memory:** How will the most relevant knowledge, such as frequently-asked question documents or email exchange history, be stored and retrieved to provide data for the FM?
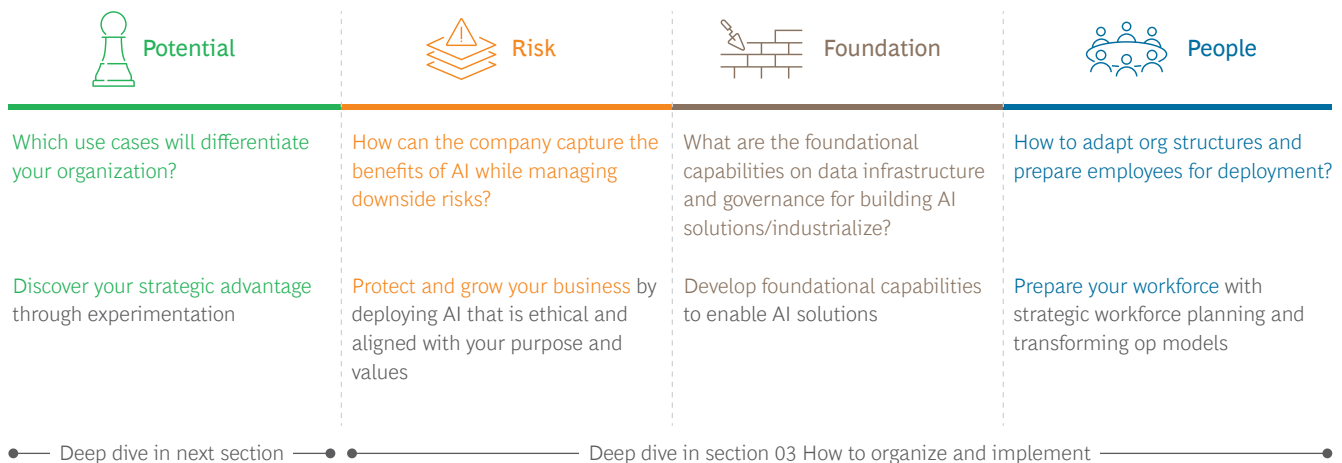
- **Environmental integration:** How will the solution be integrated with communication channels, data, and operational or banking systems? This "classic" IT architecture problem can easily become a bigger bottleneck on the way to adoption than FM-related challenges.

Four Pillars for Success

While FMs can create game-changing benefits, decision makers at financial institutions must take care in rolling out use cases and assessing impacts. There are both risks and opportunities. However, through detailed planning, many of the potential risks associated with the technology can be mitigated. In Exhibit 7, we identify four pillars of successful approaches:

- **Potential:** Where is there most potential to use FMs in the context of the bank's AI maturity? How do you build off existing data foundations? Which use cases will lead to most differentiation? And how do you create strategic advantage? Banks that have a clear vision to start with will be best placed to execute effectively.

- **Risk management:** What are the risks and how should they be managed? How can you protect and grow the business by making ethical choices that are aligned with your purpose and values? Decision makers must strike a balance between managing potential exposures and achieving benefits.

- **Foundations:** Establishing foundations means acquiring understanding. The aim should be to go beyond simple retrofitting and consider the skills, ways of working, and tools that will enable a transition to a new operating model. Among other things, this will likely mean creation and test cycles for marketing shortened from weeks to hours, and operations employees predominantly managing "AI assistants" instead of directly performing operational tasks. As in other technology decisions, the build/buy/partner dilemma will apply, and it will be incumbent on individual businesses to make choices that reflect their capabilities and direction of travel, achieving upside while avoiding vendor locks.

Exhibit 7 - Leaders must make choices across four key pillars

Potential	Risk	Foundation	People
Which use cases will differentiate your organization?	How can the company capture the benefits of AI while managing downside risks?	What are the foundational capabilities on data infrastructure and governance for building AI solutions/industrialize?	How to adapt org structures and prepare employees for deployment?
Discover your strategic advantage through experimentation	Protect and grow your business by deploying AI that is ethical and aligned with your purpose and values	Develop foundational capabilities to enable AI solutions	Prepare your workforce with strategic workforce planning and transforming op models
Deep dive in next section	Deep dive in section 03 How to organize and implement		

Source: BCG.

- **People:** Alongside technology choices, management teams must make capability-focused decisions. Budgets will need to be assigned to training, technology, and risk management. There may be a requirement to overcome resistance in some quarters, amid concern over the impact on roles and job security, while operating models and governance frameworks may need to be adjusted.

FMs offer banks a significant opportunity to boost workforce effectiveness and automate numerous processes, while creating a more streamlined and personalized customer experience. FMs can boost the quality and availability of customer-facing services, refine communications, and improve risk management and compliance.

Large banks are already rolling out use cases in areas including customer services and software engineering. However, these uses case are just the beginning. As the technology continues to evolve, banks must take a more holistic approach, focusing on the technical and environmental factors that will dictate model choice and model management. Not least, there is a strategic task to drill down into the potential of FMs, consider risks, and the lay the foundations that will promote security and accelerate the journey to scale.

1. Ryan Browne, "Goldman Sachs is using ChatGPT-style A.I. in house to assist developers with writing code," CNBC, March 22, 2023.
2. Hugh Son, "JPMorgan is developing a ChatGPT-like A.I. service that gives investment advice," CNBC, May 25, 2023.
3. Tim Hogarth, chief technology officer, ANZ, "How generative artificial intelligence can make engineers more efficient." bluenotes, May 22, 2023.
4. James Eyers, "CBA goes all in on generative AI," Australian Financial Review, May 24, 2023.
5. "AI in FinTech: 7 use cases market leaders pursue,' 8allocate, August 2, 2023.
6. Shritama Saha, "How Deutsche Bank is riding the generative AI wave," AIM, August 10, 2023
7. "Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance," Bloomberg, March 30, 2023.
8. "Synthetic data and the Wells Fargo-Hazy relationship," VentureBeat, March 28, 2022.

About the Authors

Jeanne Bickford is a Managing Director and Senior Partner in BCG's New York office. You may contact her by email at Bickford.Jeanne@bcg.com.

Rafal Cegiela is a Principal, Data Science in BCG's Warsaw office. You may contact him by email at Cegiela.Rafal@bcg.com.

Julian King is a Managing Director and Partner in BCG's Sydney office. You may contact him by email at King.Julian@bcg.com.

Kevin Lucas is a Managing Director and Partner in the BCG X Sydney office. You may contact him by email at kevin.lucas@bcgdv.com.

Neil Pardasani is a Managing Director and Senior Partner in BCG's Los Angeles office. You may contact him by email at Pardasani.Neil@bcg.com.

Ella Rabener is a Managing Director and Partner in BCG's Berlin office. You may contact her by email at Ella.Rabener@bcgdv.com.

Benjamin Rehberg is a Managing Director and Senior Partner in BCG's New York office. You may contact him by email at Rehberg.Benjamin@bcg.com.

Stiene Riemer is a Managing Director and Partner in BCG's Munich office. You may contact her by email at Riemer.Stiene@bcg.com.

Michael Strauss is a Managing Director and Senior Partner in BCG's Cologne office. You may contact him by email at Strauss.Michael@bcg.com.

Jon Sugihara is a Managing Director and Partner in the BCG X Singapore office. You may contact him by email at Jon.Sugihara@bcgdv.com.

Michael Widowitz is a Managing Director and Partner in BCG's Vienna office. You may contact him by email at Wido@bcg.com.

For Further Contact

If you would like to discuss this report, please contact the authors.