WHITE PAPER

Generative AI in Banking

Six Myths You Need to Ignore

October 2024

By Rafal Cegiela, Michal Panowicz, Lukasz Rey, Stiene Riemer, Juergen Rogg, Robert Stanikowski, Michael Widowitz, and Leonid Zhukov

Generative AI in Banking

Six Myths You Need to Ignore

Amidst all the headlines, stock market superstars, and doomsday proclamations, how is it possible for banking executives to make objective judgements about Generative AI (GenAI)? For many, the complexity of the task means they tend to be cautious on investing in the technology. About two thirds of senior managers in a recent BCG survey describe their adoption maturity as low, little, or nothing.[1] But now as some forward-looking institutions start to use large language models (LLMs) more productively, decision makers need to get clarity on what the technology offers and how they should respond.

Across financial services, GenAI is increasingly a demonstrable driver of business growth. In our contacts with clients, we have seen benefits including a 30% increase in sales, a doubling of GenAI-driven retargeting conversions, and a 10% gain in assets under management. Some financial institutions have also achieved massive increases in efficiency, with as much as 95% of service requests managed by Gen-AI powered bots.

A good place to start when considering how best to engage with GenAI is to get transparency on what the technology can offer and what it cannot. In our many discussions with banks, we have identified six areas of misconception, or myths, which we believe are clouding understanding and preventing some institutions from moving forward:

- **Myth 1:** GenAI is mostly about process effectiveness and is mainly a cost-cutting tool.
- **Myth 2:** GenAI applications are limited due to a requirement in banking for transparent and predictable decisions.
- **Myth 3:** The risk of so-called AI hallucinations and uncontrollable outputs undermines customer-facing solutions.
- **Myth 4:** GenAI software should be bought off-the-shelf.
- **Myth 5:** GenAI is an extension of machine learning/ predictive AI and therefore brings similar implementation challenges.
- **Myth 6:** Data privacy and data residency regulations limit GenAI's potential in banking.

In this article, we show that banks embracing GenAI have managed to see through the six myths to develop a clear idea of what the technology offers. These institutions have shown that it is possible to create value right now and lay the foundations for future roll out at scale.

[1] IT Spending Pulse: As GenAI Investment Grows, Other IT Projects Get Squeezed, BCG, July 16,2024. https://www.bcg.com/publications/2024/it-spending-pulse-as-genai-investment-grows-other-it-projects-get-squeezed.

Truth 1: GenAI is not just a cost-cutting tool

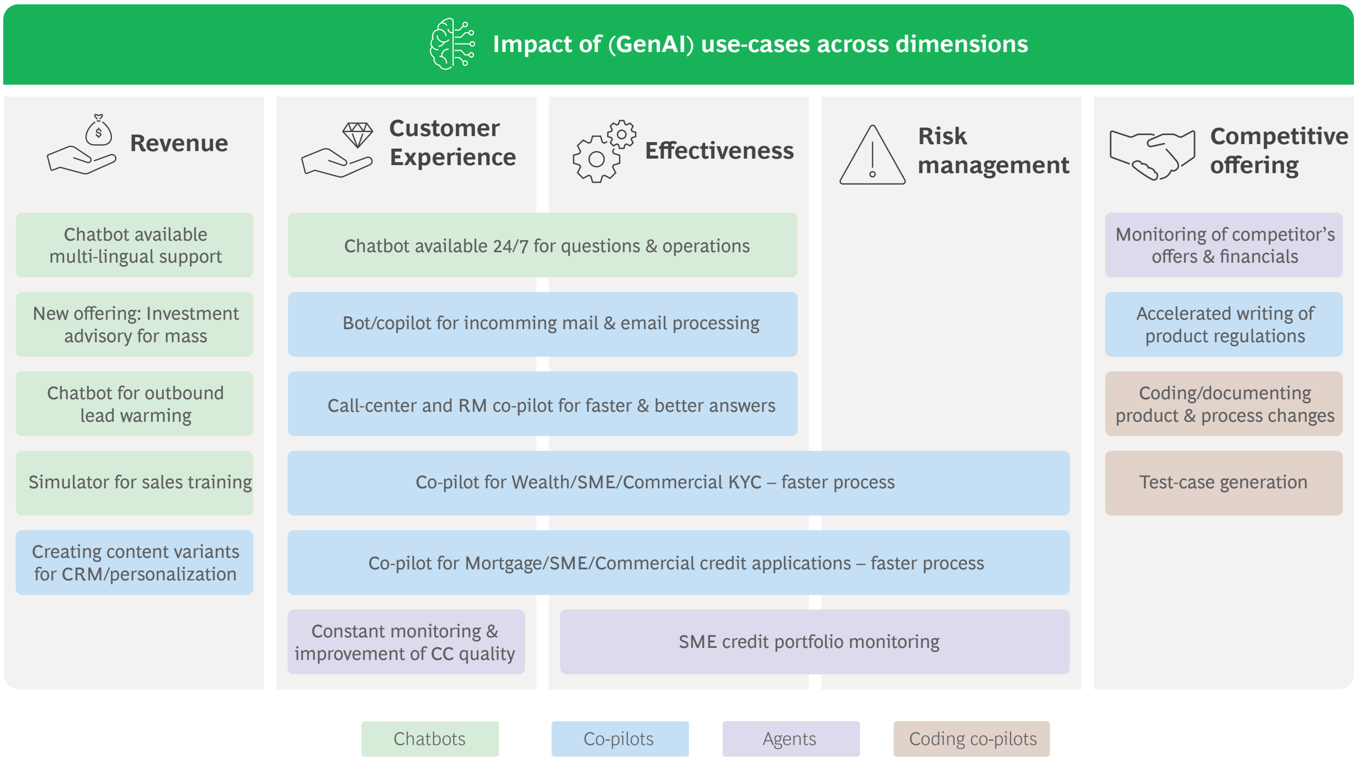
GenAI's capabilities mean it can be used to carry out tasks previously restricted to humans, but its potential goes much further than the abilities of the often-quoted parallel with "a large group of diligent and effective people." (See Exhibit 1.) This is because of the technology's ability to access a vast amount of data and at speed, meaning its answers are always synthetized from a wide range of inputs. Moreover, applications can be rolled out at a comparatively low cost, transforming multiple dimensions of operations:

Revenue generation through closer customer engagement: Banks can use chat bots to offer conversational interactions to customers who do not engage well with click-/tap interfaces – a group that includes both seniors and younger cohorts (who prefer talking to clicking). Internal bots, meanwhile, can serve as interactive sales coaches and sparring partners for RMs, as well as produce engaging and convincing content for CRM and marketing.

Transforming the customer experience: Due to a lower cost base and faster response times, Gen-AI supported agents can offer advice 24/7 and without the customer needing to navigate an interactive voice response (IVR) or wait in line. They can help humans respond about three times faster than if they relied on traditional CRMs and knowledge bases. And working alone, the technology can offer human-like interactions, while being more knowledgeable of product features and terms and conditions than any call center agent. Moreover, they can streamline back-office operations so that the bank can process documents, for example relating to loan applications, in minutes instead of days, again raising service levels significantly.

Exhibit 1 - Generative AI's impact goes far beyond cost effectiveness

Illustrative – not exhaustive



Impact of (GenAI) use-cases across dimensions

Revenue

- Chatbot available multi-lingual support
- New offering: Investment advisory for mass
- Chatbot for outbound lead warming
- Simulator for sales training
- Creating content variants for CRM/personalization

Customer Experience / Effectiveness

- Chatbot available 24/7 for questions & operations
- Bot/copilot for incomming mail & email processing
- Call-center and RM co-pilot for faster & better answers

Customer Experience / Effectiveness / Risk management

- Co-pilot for Wealth/SME/Commercial KYC – faster process
- Co-pilot for Mortgage/SME/Commercial credit applications – faster process

Customer Experience

- Constant monitoring & improvement of CC quality

Effectiveness / Risk management

- SME credit portfolio monitoring

Competitive offering

- Monitoring of competitor's offers & financials
- Accelerated writing of product regulations
- Coding/documenting product & process changes
- Test-case generation

Chatbots | Co-pilots | Agents | Coding co-pilots

Risk management: GenAI offers a step-change in the timeliness and exhaustiveness of risk management, for example equipping managers to systematically monitor credit portfolios instead of sampling them periodically. Equally, banks can use GenAI to track company reports or relevant news flow in real time.

Competitive offering: No product or strategy team can monitor market movements and competitive dynamics with the relentlessness of a GenAI agent. And when it comes to new products, GenAI's coding abilities are well known, enabling agile development and speeding up creation of product rules and sales manuals.

Truth 2. Customer facing GenAI solutions are fully transparent

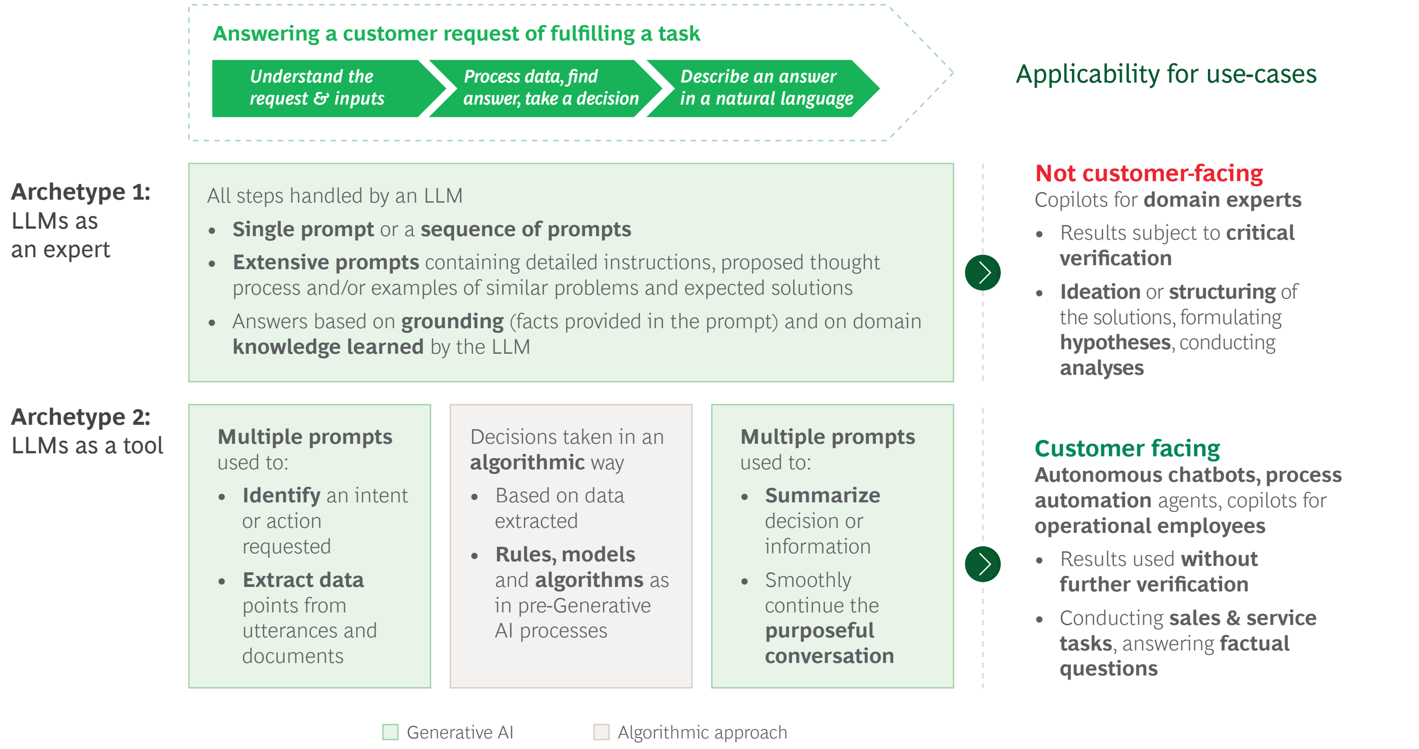
Large language models are a little like black boxes, which even data-scientists struggle to fully understand - in the same way a doctor does not fully understand the human brain. For some critics of GenAI, however, this means the technology can never be predictable or transparent.

In fact, GenAI is transparent in the context in which it operates, which is broadly to support two categories of tasks: (1) open-ended and potentially internal expert-facing tasks, for example focusing on new strategies, products, or marketing campaigns, and (2) repetitive and rule-based customer-facing tasks, for example undertaken by agents in sales and operations. (See Exhibit 2).

GenAI solutions supporting these two categories need to be shaped differently. For creative and open-ended tasks, LLM capabilities are akin to those of experts along the entire process – from understanding, through reasoning, and expressing answers. In addition, users generally have the time and expertise to verify model recommendations, while benefiting from the vast amounts of knowledge the LLMs can impart.

The second pattern is where GenAI is applied as a tool to support customer facing interactions and where there is no opportunity to verify the machine's answers. LLMs are used to translate requests and plain language inputs into data that can be processed by algorithms (the same rules that were applied in pre-GenAI era, for example to make lending decisions), and then to translate the outputs into conversational or plain language explanations. In this case, the LLM does the translating and wordsmithing, while the algorithm makes the decision. In other words, no LLM "knowledge" is required outside the grounding data used to build the model.

Exhibit 2 - In client-facing processes, Gen AI is used to understand/extract data and formulate human-like answers



Answering a customer request of fulfilling a task

Understand the request & inputs → *Process data, find answer, take a decision* → *Describe an answer in a natural language*

Applicability for use-cases

Archetype 1: LLMs as an expert

All steps handled by an LLM

- **Single prompt** or a **sequence of prompts**
- **Extensive prompts** containing detailed instructions, proposed thought process and/or examples of similar problems and expected solutions
- Answers based on **grounding** (facts provided in the prompt) and on domain **knowledge learned** by the LLM

Not customer-facing

Copilots for **domain experts**

- Results subject to **critical verification**
- **Ideation** or **structuring** of the solutions, formulating **hypotheses**, conducting **analyses**

Archetype 2: LLMs as a tool

Multiple prompts used to:

- **Identify** an intent or action requested
- **Extract data** points from utterances and documents

Decisions taken in an **algorithmic** way

- Based on data extracted
- **Rules, models** and **algorithms** as in pre-Generative AI processes

Multiple prompts used to:

- **Summarize** decision or information
- Smoothly continue the **purposeful conversation**

Customer facing

Autonomous chatbots, process automation agents, copilots for **operational employees**

- Results used **without further verification**
- Conducting **sales & service tasks**, answering **factual questions**

🟩 Generative AI ⬜ Algorithmic approach

Truth 3. Human-like solutions are feasible

Classic chatbots are disliked by many customers due to their awkwardness, while naive GenAI chatbots can easily be "jailbroken" by unusual requests – leading to unpredictable outputs or "hallucinations" - meaning that they don't know when they don't know. In this context, banks often resist using GenAI-based chatbots for customer facing interactions.

The roadblock here often comes down to chatbot design. Despite multiple new tools coming to market, chatbots are often produced simply by adding guidelines to a prompt or using a series of prompts to process customer requests. These simplistic approaches create bots that struggle to conduct structured, purposeful interactions. And while many can search the internet, or perform complex math calculations, they are unable to manage communications around more complex banking products or services.

To overcome these headwinds, forward-looking banks are building customer-facing chatbots with a backbone of core conversation stages: from identifying customer's intent, through collecting information, and reconfirming requests and execution in bank systems. (See Exhibit 3). With flow scripted or managed by an LLM-backed "meta-prompt" or "meta-agent," and LLM tools or agents applied to understanding questions and generating answers, banks can achieve a balance between a human feel and control over the conversation.

Exhibit 3 - Customer-facing chatbots balance control with a human-like experience

Not exhaustive

Approach	Flow-based (“IVR-like”)	Core-flow driven	Meta-prompt driven	Talking to LLM + guardrails	Talking to LLM (ChatGPT-like)
	All conversation turns and answers scripted	Key conversation stages & transitions scripted Answers generated by LLM under full control	Key conversation stages described with transitions managed by LLM Answers generated by LLM under full control	Conversation handled directly by LLM Guardrails preventing/ correcting unwanted utterances	Conversation handled directly by LLM
Control	■■■■■	■■■■■	■■■□□	■■□□□	■□□□□
Human-like conversation	■□□□□	■■■■■	■■■■□	■■■■■	■■■■■
Broad set of defined services	■■■■□	■■■■■	■■■■■	■■□□□	■■■□□
Ease of maintenance	■□□□□	■■■□□	■■■■□	■■■■■	■■■■■

Optimal balance between control and user experience

Truth 4. Real-life implementations do not come off-the-shelf

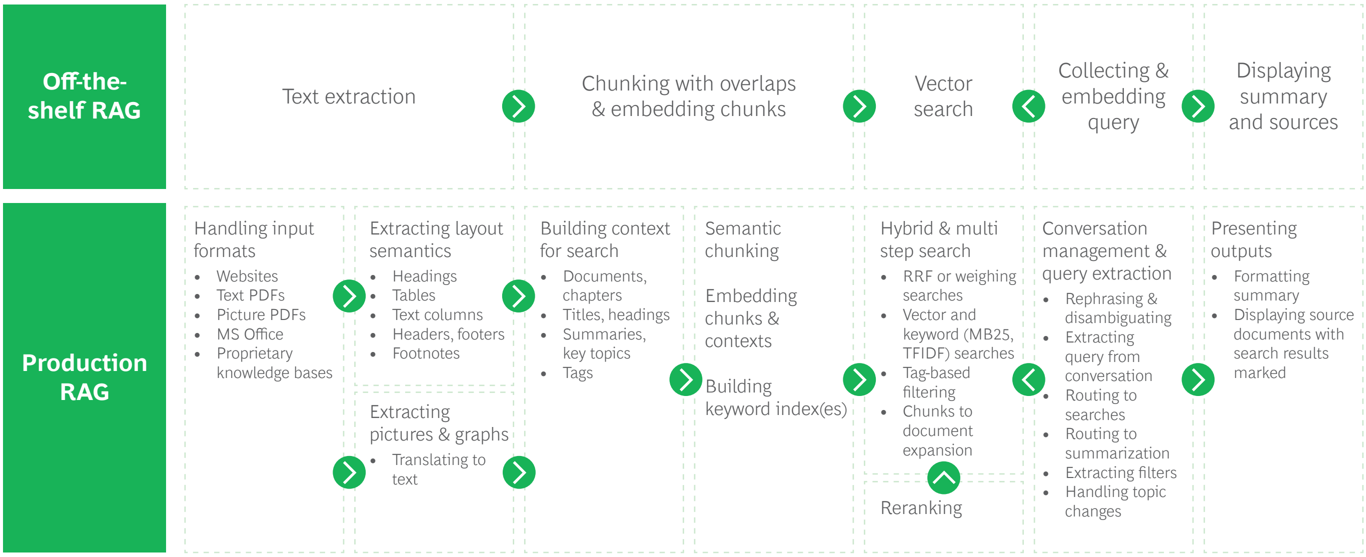
The supply of LLMs is constantly growing, amid smarter, faster, and cheaper solutions from leading startups and cloud service providers. In addition, open source and commercially available programming libraries and components are proliferating, albeit mostly in the absence of widely accepted standards. In general, if components themselves do not become de facto standards, they are abandoned and fade from view. Also, for many components, there are few support communities, meaning the opensource principle of continuous improvement is lacking.

Taking the simple example of a client chat on the contents of a document – say a product term sheet - the standard approach is based on so-called retrieval augmented generation (RAG). RAG means the chatbot retrieves document chunks from a vector database that are most relevant to users’ questions. The clauses are then presented to an LLM, which generates a final answer (called grounding an answer). This tends to work well with a few simple documents but less well when documents are more complex or numerous.

Making a RAG work in real-life is unlikely to be easy with an off-the-shelf solution. (See Exhibit 4). Instead, it requires engineering that will enable the RAG to select the right chunk from the organization’s resources, based not just on the chunk content, but also on context in terms of document, chapter, and neighboring chunks. In addition, the RAG decodes tables and figures, and tracks references between documents. Most off-the-shelf solutions are unable to offer this level of selectiveness or apply appropriate knowledge hierarchies. Moreover, in a multi-turn conversation they cannot handle nuanced discussions or topic changes.

Thus, real-live implementations generally must go one of two possible routes: either through an internal team on a made-to-measure pathway or with external vendor support but accompanied by knowledge transfer to the bank and well-documented source code that can be maintained and developed if the vendor is no longer involved.

Exhibit 4 - Real-life production RAGs have more functionalities than off-the-shelf components



Off-the-shelf RAG

Text extraction → Chunking with overlaps & embedding chunks → Vector search ← Collecting & embedding query → Displaying summary and sources

Production RAG

Handling input formats
- Websites
- Text PDFs
- Picture PDFs
- MS Office
- Proprietary knowledge bases

Extracting layout semantics
- Headings
- Tables
- Text columns
- Headers, footers
- Footnotes

Extracting pictures & graphs
- Translating to text

Building context for search
- Documents, chapters
- Titles, headings
- Summaries, key topics
- Tags

Semantic chunking

Embedding chunks & contexts

Building keyword index(es)

Hybrid & multi step search
- RRF or weighing searches
- Vector and keyword (MB25, TFIDF) searches
- Tag-based filtering
- Chunks to document expansion

Reranking

Conversation management & query extraction
- Rephrasing & disambiguating
- Extracting query from conversation
- Routing to searches
- Routing to summarization
- Extracting filters
- Handling topic changes

Presenting outputs
- Formatting summary
- Displaying source documents with search results marked

Truth 5. GenAI relies on an entirely separate tech stack, meaning no data platform is required

Many bankers think of GenAI as a building block in the same space as “classic” predictive AI / machine learning, for example used for credit risk scoring or CRM propensity models. On that basis, they believe a high-quality data layer and data governance are prerequisites to implementation. This impression is reinforced by, often unsuccessful, vendor attempts to extend data and machine learning offerings by adding GenAI capabilities.

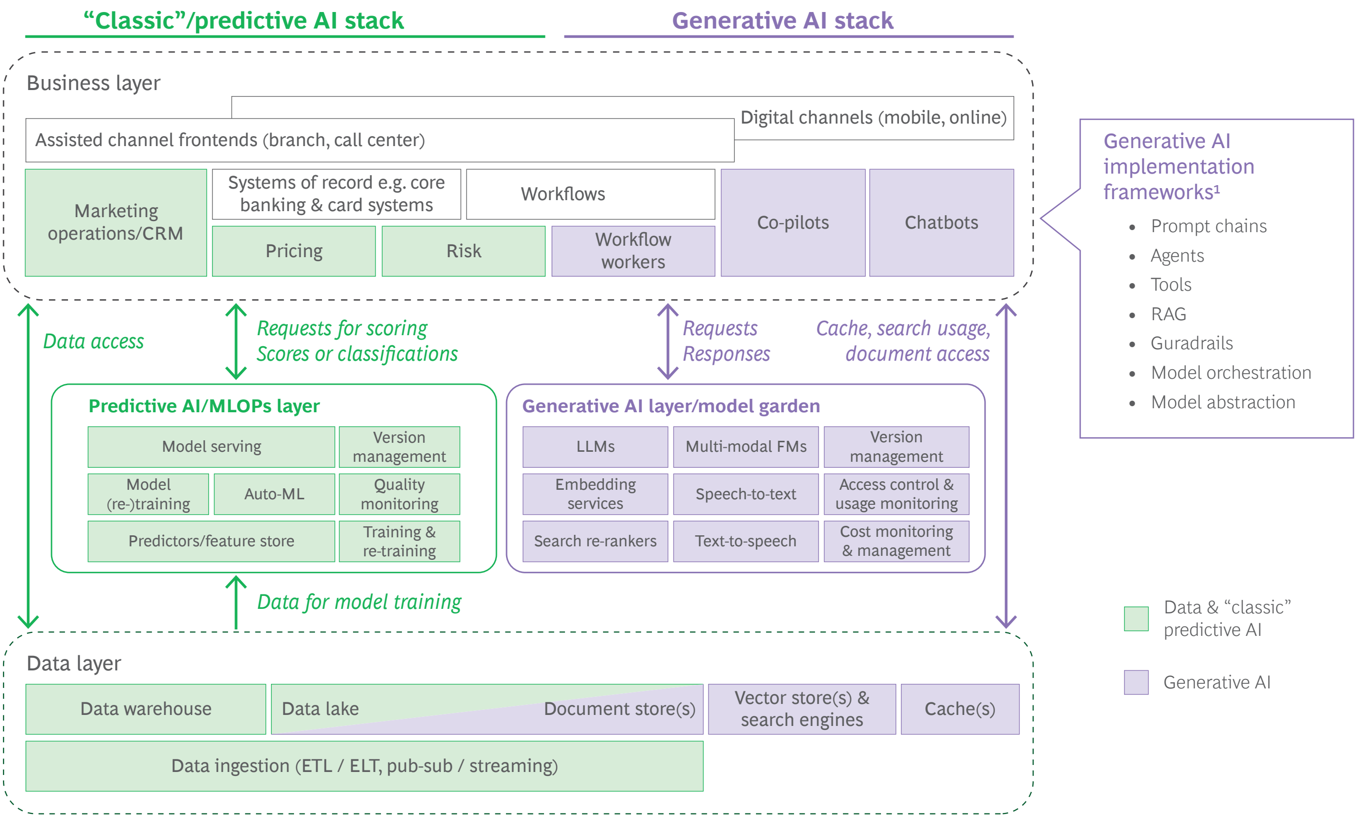
In reality, classic predictive AI and GenAI are distinct capabilities. (See Exhibit 5). Gen AI solutions depend on pre-trained large language and multi-modal foundation models (able to ingest and/or produce textual and graphical content), as well as specialized embedding and re-ranking models used to implement knowledge search use-cases.

Whether used as software as a service or open-source and run on top of infrastructure or virtual machines, GenAI models do not require additional training. Indeed, even so-called fine-tuning is barely needed for the most recent generations of models. Best-in-class structures and data layer governance frameworks may not be prerequisites for GenAI implementation, but they can help across multiple dimensions, from integrating customer data with GenAI-powered chatbots to powering document processing.

Other components for GenAI use-cases include vector stores and caches. Vector stores implement search solutions and are available as separate items - as extensions to traditional relational databases or as part of search engines that also implement “classic” search algorithms. Caches are used to store assets such as histories of conversations with chatbots, enhancing conversation effectiveness, or to store prompts to LLMs and completions from LLMs, avoiding repetitive calls to LLMs with exactly the same or very similar prompts. The only major component shared with a “classic” data and AI stack is the data lake or document store.

Establishing the above building blocks does not require an effort on a par with building a data warehouse and does not need engineers to laboriously define predictors and train models, as it required in the “classic” AI space.

Exhibit 5 - Generative AI runs on a separate tech stack to “classic” predictive AI



“Classic”/predictive AI stack

Generative AI stack

Business layer

Digital channels (mobile, online)

Assisted channel frontends (branch, call center)

Marketing operations/CRM

Systems of record e.g. core banking & card systems

Workflows

Pricing

Risk

Workflow workers

Co-pilots

Chatbots

Generative AI implementation frameworks[1]

- Prompt chains
- Agents
- Tools
- RAG
- Guradrails
- Model orchestration
- Model abstraction

Data access

Requests for scoring
Scores or classifications

Requests
Responses

Cache, search usage, document access

Predictive AI/MLOPs layer

Model serving

Version management

Model (re-)training

Auto-ML

Quality monitoring

Predictors/feature store

Training & re-training

Generative AI layer/model garden

LLMs

Multi-modal FMs

Version management

Embedding services

Speech-to-text

Access control & usage monitoring

Search re-rankers

Text-to-speech

Cost monitoring & management

Data for model training

Data layer

Data warehouse

Data lake

Document store(s)

Vector store(s) & search engines

Cache(s)

Data ingestion (ETL / ELT, pub-sub / streaming)

🟩 Data & “classic” predictive AI

🟪 Generative AI

Another frequent misconception is the expectation that agents, guardrails, or model orchestration will be present as separately managed layers or components. In reality these logical concepts and frameworks are highly specific to use-cases and are often coded according to needs - either from scratch or by means of libraries (although libraries often suffer from instability and complexity). And while some vendors offer, for example, guardrails as separate, and often costly, components, these do not have the flexibility or functional richness needed for practical solutions - or the benefits of transparency and community around open-source libraries.

The platforms supporting GenAI are also different than those used for classic AI. For the latter, there are established as standard, often open source, components that are packaged into platforms and come with professional support and guaranteed quality checks. They are offered and supported by different vendors but are similar in the ways in which they operate. GenAI offerings, conversely, are much less standardized and can differ significantly from one vendor to the next, albeit with some convergence over recent months, for example in LLM APIs offered by different vendors.

Truth 6. Data confidentiality and residency requirements can be satisfied in many ways

One of the most common misconceptions around LLM is that they risk leaking prompt data into the public sphere. The origins of this myth can be traced back to when startups used prompts to retrain their models. Nowadays, all the established providers offer guaranteed data privacy, while data retention and usage policies are strictly defined.

That said, choosing a vendor (the company that builds and trains the model) and provider (the company that exposes the model for usage) remains an important task – one that is often shaped by local regulation. (See Exhibit 6). The EU AI Act, for example, defines requirements related to disclosure of data summaries used for LLM training.

Selection of model provider should certainly take into account data privacy and residency requirements. Our experience suggests vendor offerings are useful for experimentation but production activities are best handled through Cloud Service Providers (CSPs), which are trusted business partners and fall within regulatory perimeters in most jurisdictions, meaning data privacy and residency can be fully controlled and trusted.

Exhibit 6 - Choice of deployment option is driven by regulatory requirements and availability of SaaS and virtual machines

	Vendor's SaaS	**Cloud Service Provider's SaaS**	**Cloud Service Provider's Virtual Machines**	**Local hardware (leased or purchased)**
LLM/FM deployment option	Using API of LLMs offered e.g. by OpenAI, Anthropic, Mistral under corporate agreement	Using API of LLMs offered by Azure, AWS or GCP	Deploying open-sourced LLMs to VMs offered by CSPs – optionally by means of CSPs' accelerators	Deploying open-sourced LLMs to local machines with GPUs (purchased or leased)
Range of LLMs	⊕ Best performing and most recent models	⊕ Best performing and most recent models	⊖ Open-sourced models only, very limited choice of multi-modal FMs	⊖ Open-sourced models only, very limited choice of multi-modal FMs
Data confidentiality	⊖ Contractually guaranteed, limited trust for execution	⊕ Contractually guaranteed, trusted provider	⊕ Full control inside the cloud tenant	⊕ Full control inside own data center
Data residency	⊖ Usually no choice	⊕ Clearly stated, growing choice	⊕ Fully controlled, growing availability of VMs with GPUs	⊕ Fully controlled, growing availability of leasing contracts
Management functionalities	⊖ Bare-bones, limited ability to industrialize	⊕ Version, access, cost management as for other cloud services	⊕ Access & cost management for VM	⊕ Implemented locally
Cost	⊕ Lowest cost until massive scale	⊕ Lowest cost until massive scale	⊖ More expensive unless huge scale, but can be scaled to peaks	⊖ The most expensive option
Typical application	Not suitable for banking	Best for countries with **supported languages** if available in **approved geography**	Best for countries where LLM needs to be customized for otherwise **not supported languages** OR **data residency** not supported by SaaS models	

Hybrid solution with locally deployed models for sensitive tasks (or sanitizing data) and SaaS for non-sensitive

In most cases, the preferred and cheaper option is to use LLM as a service. But it can make sense to deploy on local or virtual machines, for example where a high degree of customization is required, perhaps to manage a less popular language. Indeed, in some countries, CSPs do not offer local data residency, software-as-a-service LLMs, or even virtual machines with GPUs to run LLMs. Instead, they prioritize the US and EU for GPU expansion.

Probably the best option for countries without local SaaS would be to deploy LLMs on leased GPUs. But in practice, deployment models are often applied in a hybrid approach, with local models handling sensitive data and SaaS models outside data residency areas managing non-sensitive data.

Looking beyond the myths surrounding GenAI, the technology is on the cusp of transformative impact in the banking industry. That is, the benefits will extend well beyond cost reductions. In fact, we believe they will redefine the customer experience, boost competitive performance, and add value to the bottom line.

While there are concerns around hallucinations and controllability, carefully engineered solutions with structured conversation flows can create human-like and fully controllable customer-facing interactions. And real-world implementations require customized development rather than off-the-shelf components.

Our work with forward-looking banks shows that it is important to allocate a standalone stack to GenAI, reflecting different infrastructure needs and implementation approaches. In addition, concerns around data privacy and residency can be addressed through careful partner selection and deployment.

Taken together these myth-busting approaches, effectively implemented, can unlock new revenue streams, enhance the customer experience and, last but not least, help banks build long-term competitive advantage.

About the Authors

Rafal Cegiela, Principal, Data Science, BCG X Warsaw

Michal Panowicz, Managing Director and Partner, BCG, London

Lukasz Rey, Managing Director and Partner, BCG Dubai

Stiene Riemer, Managing Director and Partner, BCG Munich

Juergen Rogg, Managing Director and Senior Partner, BCG Zurich

Robert Stanikowski, Managing Director and Partner, BCG Warsaw

Michael Widowitz, Managing Director and Partner, BCG X Vienna

Leonid Zhukov, Vice President, Data Science, BCG X New York

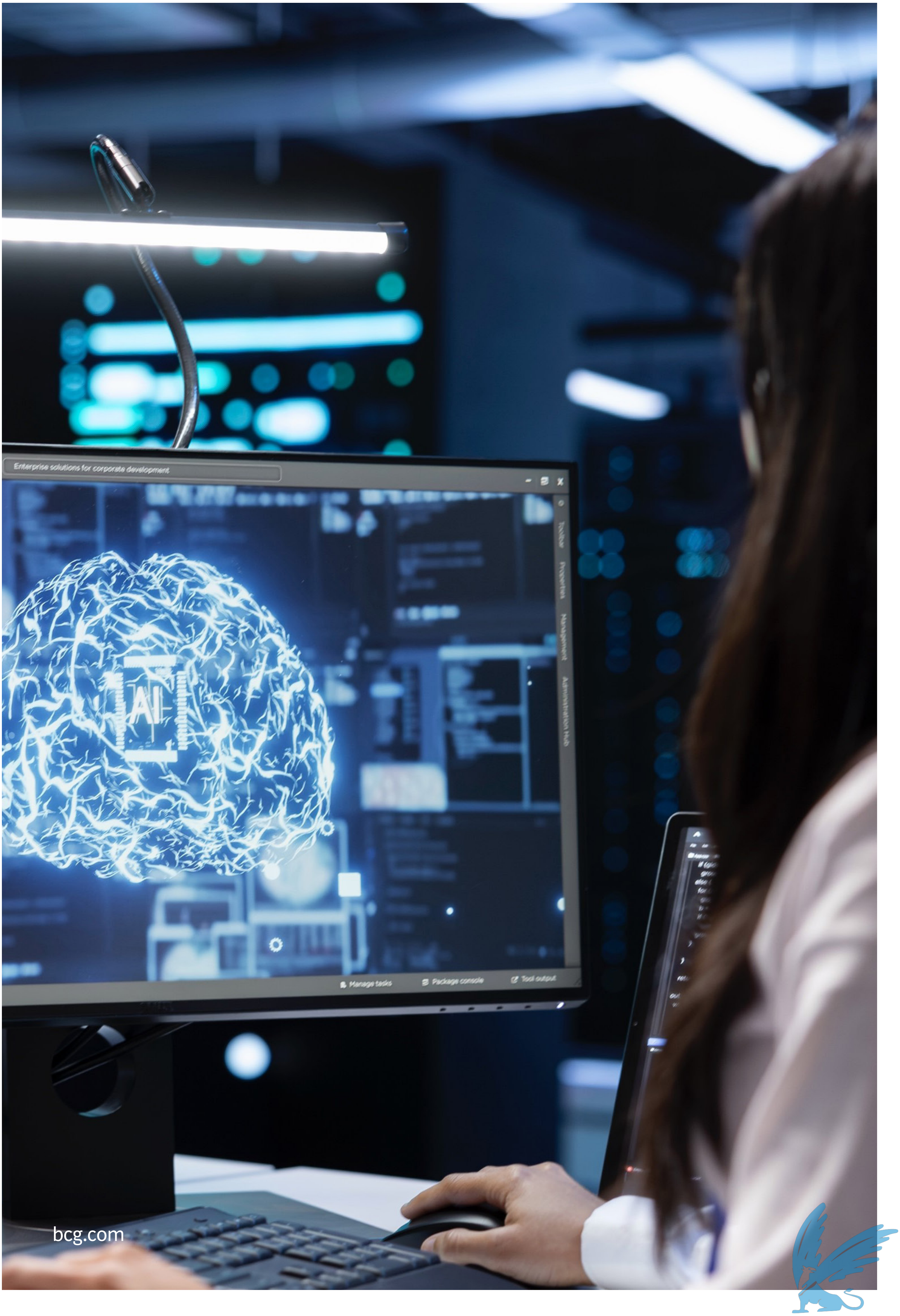
For Further Contact

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

Boston Consulting Group partners with leaders in business and society to tackle their most important challenges and capture their greatest opportunities. BCG was the pioneer in business strategy when it was founded in 1963. Today, we work closely with clients to embrace a transformational approach aimed at benefiting all stakeholders—empowering organizations to grow, build sustainable competitive advantage, and drive positive societal impact.

Our diverse, global teams bring deep industry and functional expertise and a range of perspectives that question the status quo and spark change. BCG delivers solutions through leading-edge management consulting, technology and design, and corporate and digital ventures. We work in a uniquely collaborative model across the firm and throughout all levels of the client organization, fueled by the goal of helping our clients thrive and enabling them to make the world a better place.

For information or permission to reprint, please contact BCG at permissions@bcg.com. To find the latest BCG content and register to receive e-alerts on this topic or others, please visit bcg.com. Follow Boston Consulting Group on Facebook and X (formerly known as Twitter).


Enterprise solutions for corporate development
AI
Toolbar
Properties
Management
Administration Hub
Manage tasks
Package console
Tool output

bcg.com