

WHITE PAPER

GenAI is about to transform banks front to back

February 2024

By Jeanne Bickford, Rafal Cegiela, Julian King, Anne Kleppe, Kevin Lucas, Neil Pardasani, Ella Rabener, Benjamin Rehberg, Stiene Riemer, Michael Strauss, Jon Sugihara, Mark Waehrish, Michael Widowitz and Leonid Zhukov

In financial services, GenAI is widely seen as a potential gamechanger. The technology offers a chance to transform customer relationships and widen offerings of products and services. Amid significant cost pressure, there is also a chance to reduce manual labour by a double-digit percentage. But amid slow progress on implementation, the task for decisionmakers is to turn GenAI ambition into operational reality.

GenAI could not come at a more critical time for the industry, which is battling rising competition from digital-only banks, neo-banks, fintechs, and hyperscalers. These digital natives are raising the bar on multichannel services, offering customers intuitive and streamlined experiences backed by automated operating models. As a result, their products and services are not only more appealing, especially to younger consumers, but also cheaper. Some banks are taking similar steps, reverse engineering digitized solutions into legacy systems, but many continue to fall behind.

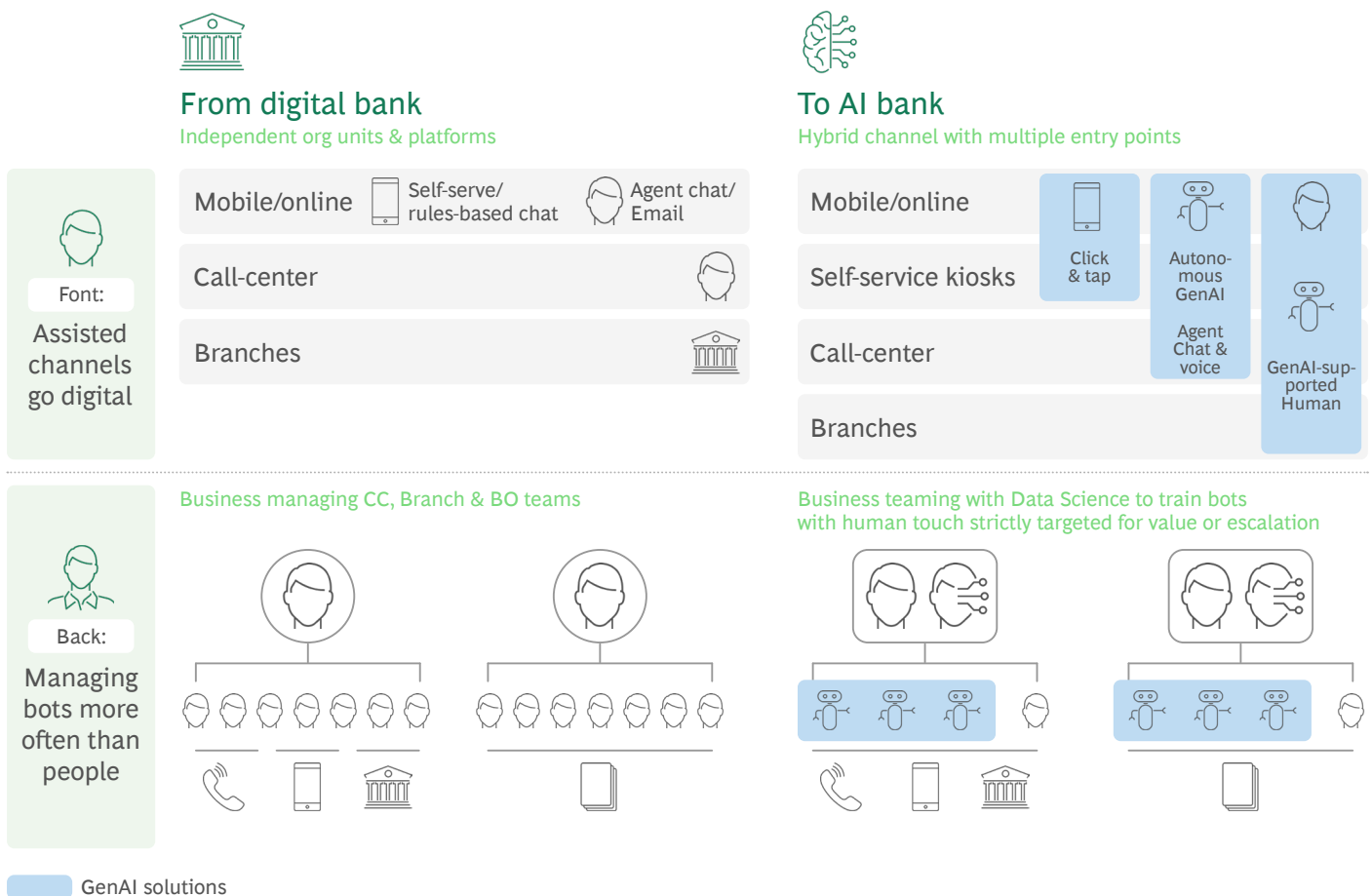
GenAI presents a chance for banks to accelerate modernization and tackle competitors in their own technology-defined back yards. However, for maximum impact, banking leaders need to go beyond an incremental or copycat approach, which tends to dampen GenAI's impact. Instead, they must embrace the opportunity. Consider a world in which customers use GenAI as a virtual assistant, from responding to requests in plain English to promoting financial fitness, automatically repaying debts, helping renew insurance, or looking for investment opportunities. These kinds of services are within reach but require strategic commitment and a robust approach to making change happen.

What can GenAI change?

AI-powered chatbots will do more than simply streamline customer communications. They will reshape channel architectures, with assisted and digital channels merged into hybrids, and will serve the full range of click/tap, talk/type and mixed interactions. GenAI's conversational capabilities will be embodied in different forms: from a bank acting as a bot on other platforms to co-browsing with a customer in a mobile app or online service. Human-to-human interactions will be reserved for the highest value or most complex situations.

GenAI will also challenge organizational structures. Sales and service channel teams will work hand-in-hand with GenAI experts, training chatbots and voice-bots to refine their abilities and produce new or updated services. More human-like conversations and avatars will appeal to even the most conservative customers and brick-and-mortar channel will be increasingly automated, with humans will play a lesser role.

Exhibit 1: GenAI brings front to back revolution for customer service operations



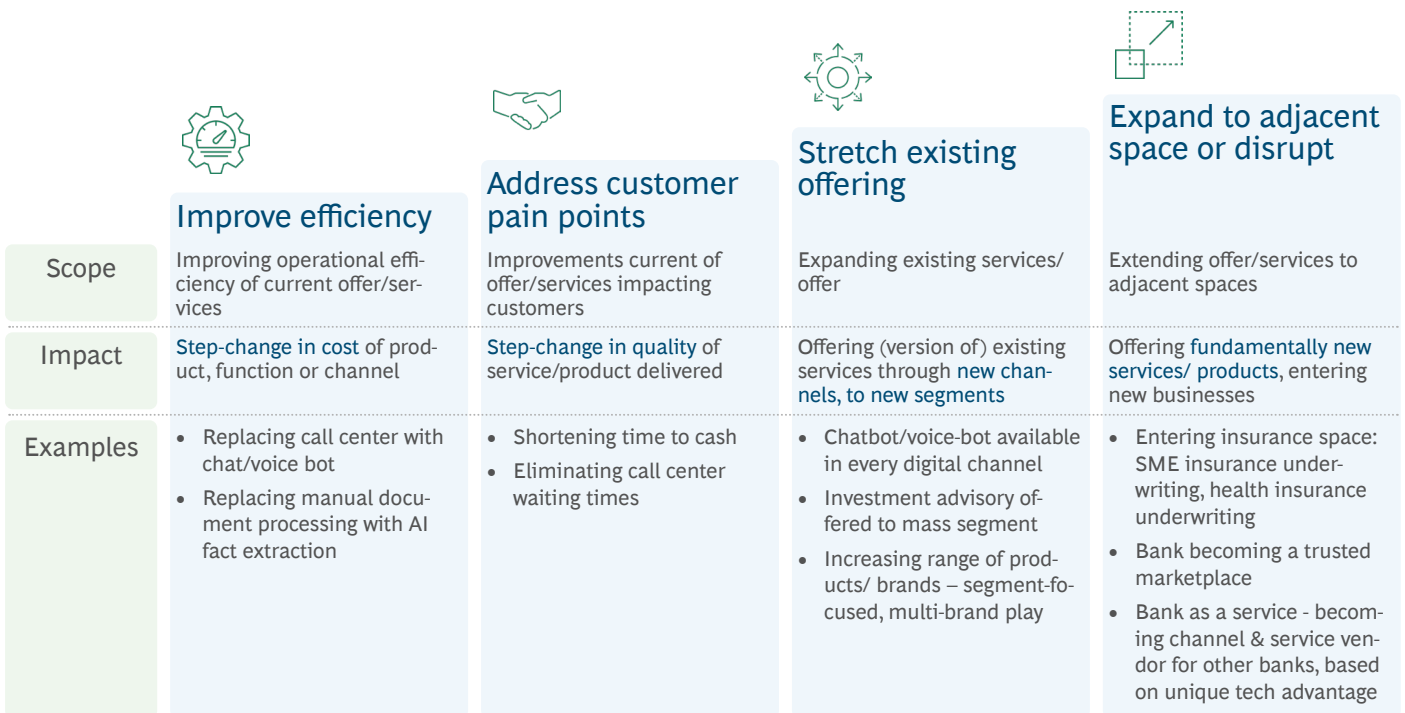
Source: BCG analysis

We expect that before long GenAI will front most customer interactions, including reading incoming correspondence and making assessments that are now made by humans. Credit decision-making and investment management will be streamlined, with human experts mostly supervising, tuning, and refining bots, as well as handling exceptions and the most valuable interactions.

Similarly, support functions including marketing, compliance, and HR will be equipped with powerful GenAI assistants, allowing them to make better-informed decisions and freeing up time to tune policies for bot management. GenAI will support risk management, creating insights across steering, and liquidity and capital management. Imagine a virtual Chief Risk Officer providing comments and insights at leadership level meetings. We expect this to be a reality in the next five years.

Finally, GenAI will have a significant democratizing effect, with chatbots offering all levels of convenience and advice to a much broader cohort of customers than is currently feasible. And GenAI will help banks speedily create new products (from coding to configuring) and maintain a wider range of products, including adopting marketplace models offering insurance, subscriptions, and investment opportunities.

Exhibit 2: Efficiency gains are just the beginning transformation journey



Source: BCG analysis

Given GenAI's transformative potential, what can decision makers do now that will make a difference to the bottom line? First and foremost, there is no value in delaying. There will be further gradual improvements in quality and cost, but the necessary technology capabilities are already in place. Thus, the onus is on decision makers to act.

The elements are already in place

The progress that banks have made in transforming their business models for digitization will serve them well as they roll out GenAI. Most banks already make extensive use of cloud services and understand the benefits of software-as-a-service, which offers both flexibility and cost efficiency in developing new offerings. All major cloud providers have regulatory approval and are signed off by internal legal and risk functions. Indeed, cloud offerings such as Azure, AWS, and Google Cloud Platform have become the preferred way to scale storage and compute, limiting the role of traditional data centers mostly to legacy workloads.

The good news for banks is that cloud services now also provide the building blocks for GenAI solutions. They offer access to large language models (LLMs) and multi-modal foundation models, alongside vector data bases, cognitive/semantic search components, and environments to help users fine tune models where necessary.

A parallel helpful trend is that GenAI providers are assimilating their solutions to existing software products. Parts of the Microsoft range, leading customer relationship management and business process management such as Salesforce and Pega, and chatbot offerings such as Kore.AI and Google Dialogflow are all starting to benefit from GenAI co-pilots.

The future is now

One of the arguments for not being a first mover in adopting new technologies is that better, cheaper alternatives will arrive in due course. In the meantime, it makes sense to observe and learn from other's mistakes. In the case of GenAI, this does not hold true. Across GenAI offerings, the technology is already advanced and ready for use at scale and at a relatively low cost.

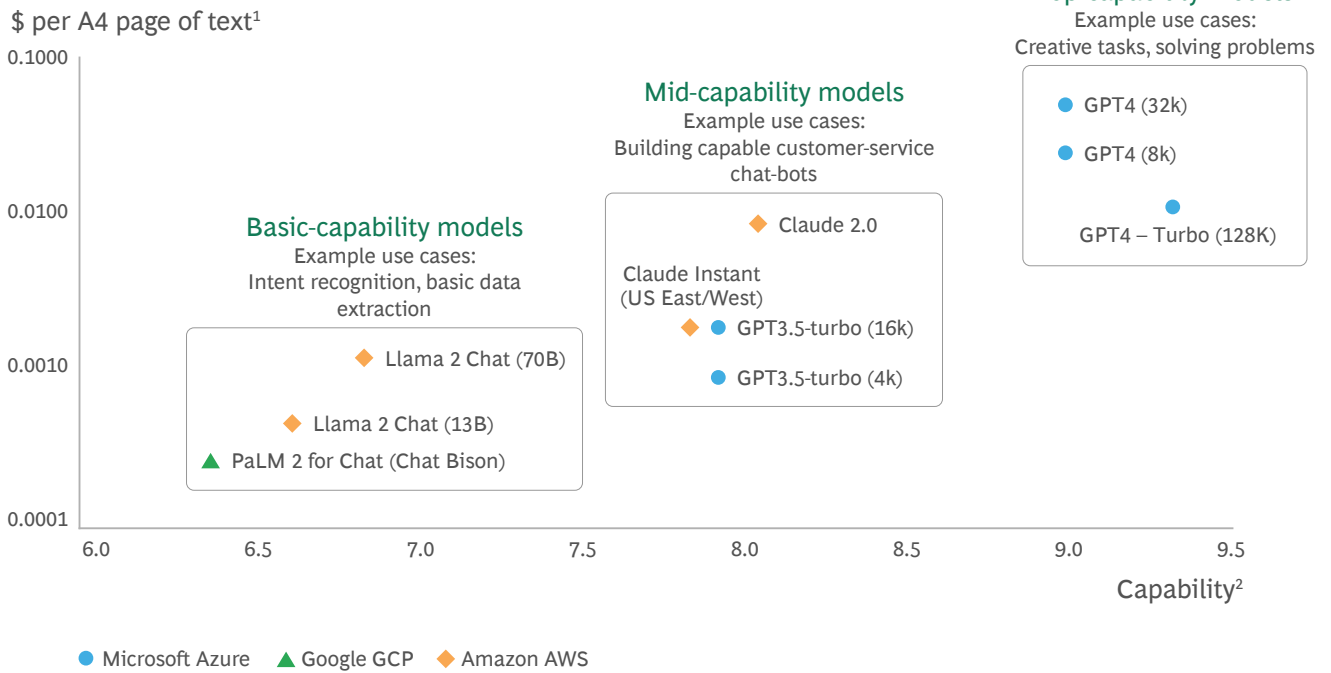
In addition, the law of diminishing returns applies. The cost of GPT4, OpenAI's most advanced model, is between 15 and 60 times higher than GPT3.5. Another leap of the same magnitude would put the price of processing a single page of A4 page at between 40 cents and \$1.50, which would be prohibitively high for even the most dedicated advocates. Moreover, the supply of graphical processing units and tensor processing units required to operate GenAI solutions is now barely sufficient to support large language models. Another quantitative leap in creating yet 10x or 100x more compute-hungry models is unlikely anytime soon.

Rather than further scale, the next generation of models is likely to see refinement of model architectures and slowing growth in internal parameters, resulting in incremental rather than quantum-leap improvements. The likely trajectory will be proliferation of smaller, and therefore cheaper and faster, models focused on domain specific languages and knowledge, offerings with guaranteed latency and response times, and products with enhanced privacy and security, for example in the form of virtually private models.

Already, most leading database engines are being equipped with vector search capabilities, which use so-called embeddings to help users to find solutions through vague "meanings" rather than specific words. There will also be rising numbers of toolkits for connecting components and integrating them into existing architectures. Still, these advancements are enhancements rather than anything more fundamental.

Exhibit 3: Diverse range of language models offered in trusted clouds

Non-exhaustive



1. Average of 400 words on a page, 1 token = 0.75 word, 50:50 input:output (prompt:completion) tokens 2. MT-Bench score
 Source: BCG analysis, December 2023, AWS, GCP, Azure, [2306.05685] Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (arxiv.org)

The clear implication for decision makers is that waiting will be suboptimal. Mid-sized models such as GPT-3.5, Google Bison and Anthropic Claude Instant offer an excellent cost/capability trade-off for bulk or mass-processing tasks and customer-facing solutions. That said, it will make sense to design use cases with the expectation that cheaper and faster, rather than smarter models, will eventually be available.

Building on bank’s data resources and transformer models

Banks have access to an enormous volume of data within transaction streams, which provide deep insights into customers’ behaviors, short-term desires, and long-term needs. Experiments show that transformer networks, which underpin large language models, can be applied to fully utilize transactional data while replacing classic machine learning in producing risk scores and product propensity models. The resulting models will not have the form of language models but may take prediction precision to the next level and reduce the data engineering required to build classical models. Achieving these benefits, however, will require massive volumes of transactional data, equivalent to multiple years of history in a bank with 100 million customers. This could lead to a new competitive dimension between the largest banks and vendors, or prompt banks to explore federated learning.

Implementation challenges and benefits

In some respects, rolling out GenAI uses cases and scaling up the most successful will be easier than building earlier generations of artificial intelligence models from scratch. That's because large language models and foundation models are pre-trained, enabling them to step almost directly into human roles. However, in other respects, GenAI brings its own sets of challenges.

When banks started using machine learning and classical artificial intelligence just a few years ago, their biggest hurdle related to data. Many banks struggled to build effective data pipelines to feed high-quality models. At the same time, business teams and data scientists were hard pressed to weave modeling results into business processes. And banks were rarely organized for more comprehensive steering by means of machine learning.







Now, as banks consider how they can apply GenAI, the data issue is not so severe. Indeed, one of GenAI's most useful advancements is that it can use the same data and language that humans routinely use in their daily work. Moreover, GenAI offers an unprecedented ability to extract new data points from numerous documents and messages, giving banks the opportunity to both launch new solutions and refine their machine learning models.

Because GenAI is built to operate in similar linguistic environments as humans, the implementation process is likely to be more organizational than data focused. The biggest challenges lie in striking a balance between tasks delegated to GenAI and other algorithmic and human decisions, as well as guardrails, human oversight, and escalation/exception handling.

The benefits of getting these decisions right can be significant. In some functions, tools such as Chat GPT can lead to immediate 10-15% uplifts in productivity. Similarly, co-pilots retrofitted into applications will allow for even faster, more efficient execution of tasks by humans. On the other hand, research shows that entirely spontaneous attempts at adoption can sometimes surprisingly result in productivity declines.

Where GenAI can be applied to deeply transformational use cases, it is capable of producing 50-80% productivity gains. But these cannot be delivered by off the shelf components alone. Instead, banks will need to build tailored solutions on top of plug-and-play building blocks. This will require them to weave together classical software and data science solution engineering techniques, deconstructing problems into steps addressable by queries to large language models, equipping large language models with tools, applying retrieval augmented generation (RAG), sometimes building reasoning-acting (ReAct) agents, and in course of that creating multiple tens of prompts choosing from zero- or few-shot learning or chain-of-thought prompting.

Exhibit 4: High-responsibility use-cases require tailored solutions

	Ad-hoc use-cases	Embedded use cases	Base use-cases	Advanced use-cases
Functionalities	Direct queries to publicly available LLMs/FMs	Extending/retrofitted to existing tools	Solutions existing in public space, e.g., part of Langchain	Solution custom built/ tailored for purpose, beyond publicly available
Audience	Internal / employee facing	Internal / employee facing	Internal / employee facing	Employee facing & customer facing
Responsibility	<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="width: 40%; background-color: #e0f0ff; padding: 5px; text-align: center;">Increasing effectiveness of current ways of working</div> <div style="width: 30%; background-color: #e0f0ff; padding: 5px; text-align: center;">Transforming ways of working</div> <div style="width: 30%; background-color: #e0f0ff; padding: 5px; text-align: center;">Replacing human work</div> </div>			
Examples	<ul style="list-style-type: none"> • Chat GPT  • Midjourney  	<ul style="list-style-type: none"> • Microsoft copilot  • Github Copilot  • Google Infobot  • Kore.AI flow creation  	<ul style="list-style-type: none"> • Information search & question answering • Document summarization • Serial content generation 	<ul style="list-style-type: none"> • GenAI-based chatbots • Agent support/whisperer • Gen-AI-based process automation • Mass-content generation
Implementation	100% Change management	10% Deployment 90% Change management	30% Build 70% Change management	50% Build 50% Change management

Source: BCG analysis

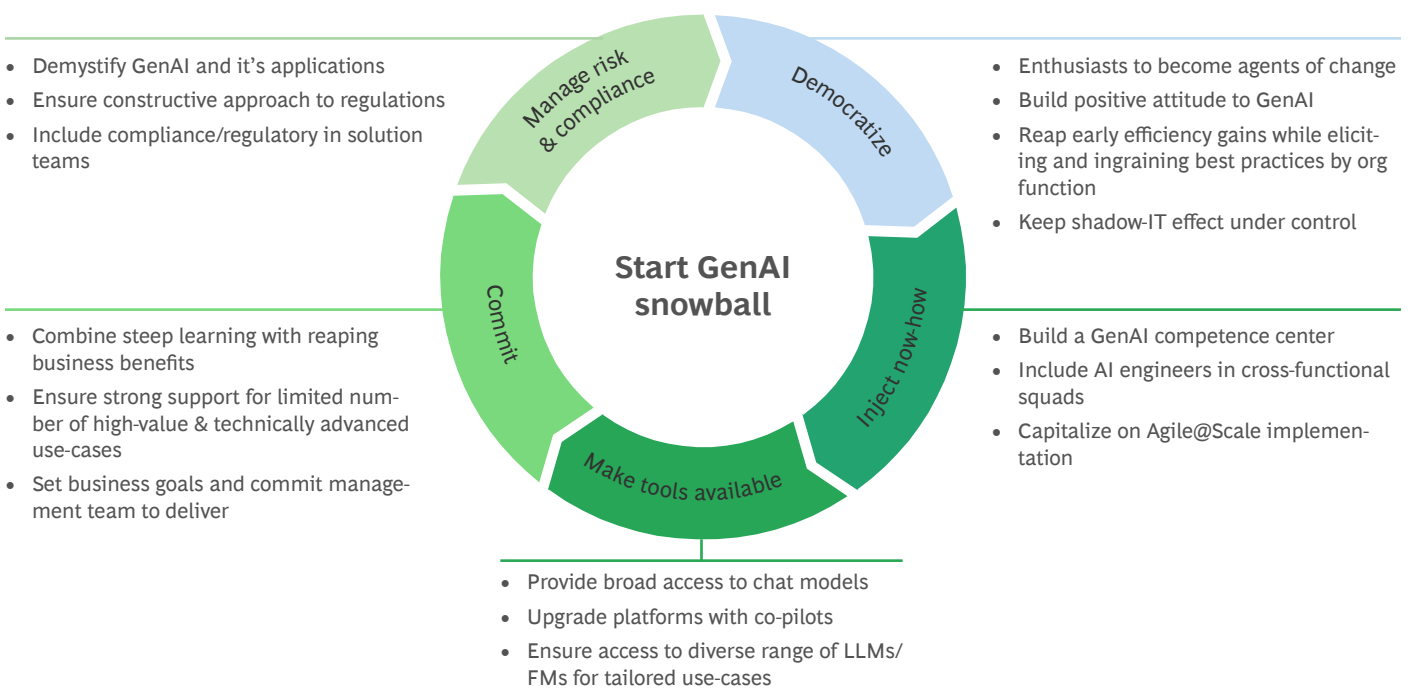
The most complex use case implementations, such as commercial lending underwriting or putting a GenAI chatbot in front of customers, will require considerable cutting-edge engineering. Innovation in design and implementation is where banks can create a competitive edge, as was often the case in previous technology rollouts, such as building digital channels and personalized digital marketing.

How to get started

On top of the complexities of mastering new engineering patterns and asking the right questions to large language models, banks will need to carefully manage the human and organizational aspects of GenAI implementation. They will need to acquire new skills and behaviors, while initial implementations may fall short of (inflated) expectations. Enthusiasts may stray into potentially ineffective and unsafe usages. Shadow IT, meaning systems deployed outside of scrutiny of IT department, can grow at pace. Managers will fear uncertain outcomes, high business or operational expectations, and potentially shrinking kingdoms. Employees will also be concerned about the need for re-skilling. Legal and compliance functions, meanwhile, may present conservative mindsets and tilt toward limiting risks rather than capitalizing on business or operational opportunities.

To overcome these challenges, decision makers will need to adopt an orchestrated approach that brings together stakeholders from across the business, data-science, technology and risk management – as well as AI engineers. Successful adoption will consist of five major components: (i) democratizing GenAI, (ii) injecting know-how into organization, (iii) enhancing the tech platform with new building blocks, (iv) committing to transformation, and (v) managing risks in the compliance and regulatory context (See Exhibit 5).

Exhibit 5: Multiple ingredients required to reap benefits



Source: BCG analysis




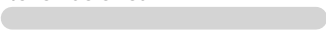

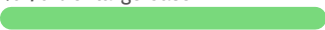



To build on early curiosity and spontaneous usage, banks should provide employees with a combination of encouragement, advice on best practice, access to tools, and policies for safe usage. This will foster enthusiasm, help people overcome fears, reap early performance improvements, and control risks around shadow-IT.

It will make sense to build a competence center, which can engage with business teams to share responsibility for designing and implementing tailored solutions. Here, organizations that have already adopted agile approaches at scale will benefit from their ability to operate effectively in cross-functional teams, enrich teams with new competencies, and incrementally improve MVPs. With time and accumulated know-how, competence centers will then become another unit of the agile organization.

When providing broad access to chat tools and co-pilots, the perspective of covering the investment with performance improvements and building up enthusiasm should prevail over a pure cost control perspective. We believe GenAI will in future be as pervasive as basic word processing tools are today, so early investment will likely pay dividends. As the offer of LLMs & FMs is expected to continue evolving in both capabilities and costs the best approach is to avoid premature lock-in with a single vendor.

It may be tempting to start cautiously, with low-cost proofs of concept and a willingness to accept best-effort results. However, given the frictions discussed in this article, this approach is likely to impede progress and raise the chance of disappointment. Conversely, if banks fully commit to a handful of challenging but potentially rewarding use cases, they can quickly accumulate expertise, generate business impact, and build confidence for further investment. The optimal approach will be to set “must have” targets with additional rewards for outstanding performance.

Exhibit 6: Committing to challenging use-cases guarantees impact & learning

Approach	Proliferating PoCs	Multiple easy/ mid-complexity use-cases	Focus on 2-3 transformative & challenging use-cases
Scope	Extending/retrofitted to existing tools	Solutions existing in public space, e.g., part of Langchain	Solution custom built/ tailored for purpose, beyond publicly available
Go-live	No up-front commitment	Expected	Full commitment
Transformation	Performance improvements w/o process & organization changes	Performance improvement & some process transformations	Transforming processes & roles/ organization
Engineering	Off-the-shelf tools and/or basic engineering patterns ¹	Off-the-shelf tools and/or basic engineering patterns	Ahead-of-the-curve engineering on top of available components
Key KPI	#PoCs conducted	Business impact	Business impact
Probability of success	High – easy task 	Low – limited commitment & dispersed attention 	High – full focus & commitment 
Business impact	None / deferred 	5-10% of small bases 	40-70% of large base 
Capability build	Basic skills, mostly engineering 	Practical, key patterns, engineering & deployment 	State-of-the-art, battle-tested, transformational skills 

Source: 1. Eg. Retrieval Augmented Generation, Prompt Engineering: Few Shot Learning, Chain-of-thoughts

To navigate potential legal and compliance hurdles, decision makers should bring relevant departments in from the get-go, ensuring they remain aligned and encouraging a constructive mindset. Contingent on achieving this sense of ambition and joint enterprise, banks can realize GenAI's potential and their own.

GenAI offers banks a chance to transform their services and reduce manual labour by reduce manual labour by a double-digit percentage. But amid slow progress on implementation, there is little benefit to be gained from waiting. To move forward, banks will need to adjust their channel architectures and recalibrate the organization to optimize the new capabilities. Along the way, they will need to consider factors including where and how best to deploy GenAI, guardrails, human oversight, and exception handling. Those that best orchestrate these variables, and achieve stakeholder buy in, will be most likely to build useful applications at scale and create the impetus for competitive advantage.

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.

Anne Kleppe is a Managing Director and Partner in the BCG X Berlin office. You may contact her by email at Kleppe.Anne@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.

Mark Waehrisch is a Partner and Associate Director in BCG's Frankfurt office. You may contact him by email at Waehrisch.Mark@bcg.com.

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

Leonid Zhukov is a Vice President - Data Science in BCG's New York office. You may contact him by email at Zhukov.Leonid@bcg.com.

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](https://www.bcg.com). Follow Boston Consulting Group on [Facebook](#) and [Twitter](#).

© Boston Consulting Group 2024. All rights reserved. 2/24

