

Adopting generative AI in banking



Adopting generative AI in banking

This white paper looks at how Generative AI (GenAI) can transform the financial services industry. It shows how GenAI can improve compliance, engage clients better, and manage risks more effectively while helping companies share expertise and bring new solutions to market faster. However, to fully benefit from GenAI, high-quality systems are needed, and it can be tough to tell different AI models apart.

The paper explains what makes an AI system good, including the quality of data, the complexity of the system, and cybersecurity measures. It also covers important tools like Retrieval-Augmented Generation (RAG) and Parameter Efficient Fine Tuning (PEFT), which help reduce mistakes.

A step-by-step approach for adopting GenAI is recommended, starting with simple tasks and gradually expanding to more complex uses to ensure a smooth transition. Key suggestions for the industry include using standard test data, setting up fair usage policies, creating AI testing environments (sandboxes), and simplifying regulations to make AI integration more effective.



Thomas Kaiser
CEO and Co-Founder, Kodex AI

Drives innovation in GenAI solutions for regulatory compliance within financial services. Formerly a strategy consultant at BCG, specialising in banking compliance projects, with leadership expertise in platform ecosystems and business development across Europe, Africa, and Asia. Based in Berlin.



Boon-Hiong Chan
Industry Applied Innovation Lead & Head APAC Market and Technology Advocacy, Securities Services, Deutsche Bank

Product builder and advocate specialising in technology and regulatory implications on financial services. Background in Computation/AI and Cybersecurity & Policy, based in Singapore.



Delane Zahoruiko
Founders Associate, Kodex AI

Specialises in corporate development and go-to-market strategies for GenAI solutions, with expertise in business operations, big data analytics, and consulting across the US and Europe. Based in Berlin.

Table of contents

| | |
|---|----|
| Foreword | 4 |
| <hr/> | |
| 1 Introduction, GenAI and its use cases | 5 |
| 1.1 GenAI: a powerful transformative tool..... | 5 |
| 1.2 Use in financial services | 6 |
| 1.3 Use case considerations..... | 7 |
| 1.3.1 Key differences..... | 7 |
| 1.4 Alignment of a problem statement with GenAI..... | 9 |
| <hr/> | |
| 2 Building a portfolio of GenAI use cases | 14 |
| 2.1 Three-stage approach | 14 |
| 2.1.1 Stage 1: apply GenAI's core text and language analysis capabilities | 15 |
| 2.1.2 Stage 2: chat-to-agent use cases | 16 |
| 2.1.3 Stage 3: chat-to-execution. The autonomous capabilities | 18 |
| <hr/> | |
| 3 Identifying excellence | 20 |
| 3.1 Accuracy and relevance (benchmarks) | 20 |
| 3.1.1 Benchmarks at the foundation model level..... | 20 |
| 3.1.2 Domain specific benchmarks | 21 |
| 3.2 Architecture and Process Factors..... | 23 |
| 3.2.1 Data..... | 23 |
| 3.2.2 Customisation | 24 |
| 3.2.3 Data and Training for language and cultural sensitivity..... | 25 |
| 3.2.4 System's creativity, reasoning and problem-solving | 25 |
| 3.2.5 Speed, performance and costs | 25 |
| 3.2.6 Support and upgrades..... | 26 |
| <hr/> | |
| 4 Factors that compromise quality | 27 |
| 4.1 Data risks | 27 |
| 4.1.1 Data risks mitigants | 28 |
| 4.1.2 Synthetic data to enhance training..... | 28 |
| 4.2 System risks | 28 |
| 4.2.1 Model Drifts..... | 29 |
| 4.2.2 Hallucination risks..... | 29 |
| 4.2.3 Feedback loop degradation: addressing user influence..... | 30 |
| 4.2.4 Mitigants | 30 |
| 4.2.5 Model evaluation techniques as mitigants..... | 31 |
| 4.3 Other risks | 31 |
| 4.3.1 Dependency risks | 31 |
| 4.3.2 Cybersecurity risks | 32 |
| 4.3.3 Sustainability risks | 33 |
| <hr/> | |
| 5 Implementing GenAI & Recommendations | 34 |
| 5.1 Regulatory considerations | 34 |
| 5.2 Explainability and transparency..... | 36 |
| 5.3 Data and Hallucination..... | 36 |
| 5.4 Synthetic data | 37 |
| 5.5 IP and copyrights | 38 |
| 5.6 Open standards and fair data practice | 38 |
| 5.7 Cross-border scalability | 39 |
| 5.8 Non-English benchmarks..... | 39 |
| 5.9 Expertise availability, jobs and reskilling..... | 40 |
| 5.10 Quantifying ROI and productivity gains..... | 40 |
| <hr/> | |
| 6 Conclusion | |
| Can GenAI thrive in the regulated financial industry? | 41 |

Disclosure: This white paper has been passionately authored by humans.

Foreword

GenAI is emerging as a transformative force in financial services, enabling efficiencies in compliance, client engagement, and risk management. They offer significant advantages to super-scale human productivity for more to be achieved with the same with higher quality. Other benefits include institutionalising expertise, raising strategic competitiveness and high speed-to-market when more business or non-technical users can access such systems to achieve better results faster. Social compacts between the firm and society can also be reinforced too, for example if a mature professional's domain experience is further extended as model trainer and evaluators.

A structured and incremental adoption roadmap is critical, beginning with language processing and automating routine workflows and scaling up to more complex decision-augmentation applications that can include Agentic AI in the future. Such a strategic approach ensures that benefits, learnings and risks can be optimally internalised at each stage before the organisation progresses; effectively allowing business and applications to mature together.

However, for such value creations to be realised, the GenAI system needs to be of a certain quality even if differentiating between GenAI systems can be difficult as all appears similar.

Hence, taking a business-technical approach, this white paper seeks to highlight quality determinants such as data, system components and cybersecurity of GenAI text-to-text systems. It also introduces key infrastructure components like Retrieval-Augmented Generation (RAG), Parameter Efficient Fine Tuning (PEFT) and Low Rank Adaptation (LoRA) that catalyses benefits like system adaptability, reduced running costs and to support data confidentiality segregation in more controlled fashion.

While Gen AI may look like a new technology because of many new terms, deeper examinations would reveal that many deemed challenges are familiar from previous technological advancements that the financial industry have successfully addressed. For example, Hallucination – a large word – is about inaccurate, unreliable or incomplete data from information retrieval technologies that have been addressed in previous innovations like early expert systems through improved data validation, user oversight and iterative model training. Premature responses to these risks believing they are novel to GenAI can only hinder and add unnecessary costs to an industry that needs new growth tools.

The paper concludes with key insights and suggestions for the future of adopting the GenAI system by the financial industry; including availability of industry test data, fair usage policies, upskilling, AI sandboxes and streamlined existing regulations relevant to AI systems for cost-effective adherence.

We hope you find this paper engaging and invites further discussions in this exciting field. Thank you for reading.

01

Introduction, GenAI and its use cases

In recent years, Generative Artificial Intelligence (GenAI) has evolved from a futuristic concept to a practical transformative tool that can reshape industries. For financial services industry professionals, stakeholders, employees and clients, it has reignited excitement in this field with broader interests to use GenAI systems for a huge range of use cases.

To this goal, the white paper aims to raise awareness of key factors involved in implementing GenAI systems and to provide guidance on approaches, success factors, risks and regulatory considerations. In the process, we also attempt to highlight what differentiates quality between different GenAI systems and propose some next steps that we believe can facilitate industry adoption of this extraordinarily powerful tool that can greatly empower users to achieve more with the same.

1.1 GenAI: a powerful transformative tool

The versatility of GenAI allows it to be deployed across a variety of use cases, and the ease of its use makes it a democratising transformative tool across industries and skill levels. At its most basic level, GenAI can be used for automating simple tasks, such as generating marketing messages, automating data entry, or creating boilerplate content with helper functions. These applications are already useful to augment productivity and speed, even by early-career professionals or those without the relevant background to achieve a measure of success. This is because at the core of GenAI is a superb human-computer interface that allows everyday language to be translated into precise computer instructions for machine execution. This allows more and different types of users to manage more complex tasks.

For example, an intern can use GenAI to create email invitation templates in different languages to a technology webinar, while a product manager can generate user flow diagrams without needing specialised design tools.

On the other end of the spectrum, GenAI powers complex and highly specialised systems such as Microsoft's Co-Pilot product launched in 2023. This AI-driven assistant integrates into productivity software to help users automate tasks like code generation, document drafting, and real-time collaboration, enhancing workflows for seasoned professionals. Co-Pilot can, for instance, assist software developers by generating large blocks of code based on minimal input, significantly reducing development time and improving efficiency. Similarly, in a legal setting, other GenAI systems can help attorneys draft legal documents by understanding context and offering suggestions, potentially transforming how professionals in high-stakes environments operate.

1.2 Use in financial services

Beyond these examples, GenAI's adaptability can be seen in its use within financial services. Financial institutions are increasingly leveraging AI for applications such as fraud detection, customer service automation via chatbots, and even risk assessment models that evolve as markets change. For instance, a compliance officer can use GenAI to query, without knowing SQL (the structured query language for manipulating data into relational databases), other machines for transaction anomalies in data sets to help flag potential compliance risks faster than in traditional ways.

Figure 1: GenAI models

| Type of models | What they are |
|---|---|
|  | Foundation model A broad, general-purpose model trained on diverse data. |
|  | Instruction-trained model A model based on Foundation Model but refined with specific instructions to perform particular tasks. |
|  | Fine-tuned model A Foundation Model or an Instruction-Trained Model that is further trained on specialised datasets to enhance performance for specific domain applications |
|  | Deployed model with prompt engineering The practice of users crafting inputs to guide the Fine-Tuned or Instruction-Trained model towards desired outputs |

This is possible because GenAI systems can be fine-tuned for specific tasks and specific domains to make them adaptable. Whether through low-code or no-code platforms, businesses can customise GenAI systems to meet their specific requirements in near real-time. A GenAI model fine-tuned for the financial sector, for instance, would be focused on understanding financial language and providing insights on the applications like market analysis or drafting financial reports.

1.3 Use case considerations

As GenAI and the range of AI technologies continue to evolve, their abilities to solve complex situational challenges would also expand and appear to be infinite. As expectation builds with each news of new capabilities or of another successful use case, GenAI systems risk becoming the silver bullet to everything that needs to be solved, which is unrealistic.

1.3.1 Key differences

Hence, in deploying GenAI, which is a highly powerful tool that augments and creates super productivity benefits when properly applied, it is important to understand what and where it should be deployed with grounded expectations. This would facilitate business case success and fit-for-purpose governance.

This paper examines the GenAI system, with Figure 2 highlighting key differences between GenAI, AI, and other comparable systems. While these systems may seem alike, each possesses unique capabilities, risks, and regulatory profiles.

Figure 2: Not all AI is the same

| |  |  |  |  |
|-------------------|---|---|--|---|
| What is this for? | Automate repetitive tasks and workflow. No "new" output | Pattern recognition, regression analysis/prediction, classification | Content generation (eg, text, images, code, etc) | Autonomous decision making and action |
| Main capability | Copy human interactions with systems. Does not create new methods of interaction | Analysis, application and prediction based on existing data/model. Arguably little real time learning | Output new data and generate output. Real-time learning, self-course correction | Interact with other systems, learn and act in real time |
| Learning | Imitation rule-based. Do not learn | Single algorithm to machine learning. More structured and constrained than GenAI. | Self-supervised, unsupervised, latent space representation | Reinforcement learning, unsupervised learning |
| Use case type | Task automation; data entry, process automation | Human enablement; risks management, customer segmentation, predictions | Human augmentation; Text, image, audio, code generation | Autonomous AI assistants and team |

GenAI as a human interface/integration into RPA, Traditional AI or Agentic AI

Source: SES Views, Deutsche Bank

1.4 Alignment of a problem statement with GenAI

When determining whether a problem statement is more appropriate for GenAI, several key criteria should be considered.

1. Problem variability:

is the goal to create new content with tolerance for variations, or to classify/predict existing data with the same results to the same queries?

GenAI systems excel at tasks such as generating text, images, audio, and code. For instance, in text generation, their capabilities include summarisation, extraction, sentiment analysis, inference, and applying one concept to another. They also adeptly connect related topics, akin to a mind map.

It is not suitable for quantitative data analysis, classification, prediction, or tasks commonly linked with 'traditional AI.' For example, if a business seeks to generate personalised client emails based on past interactions and understanding of their behaviour/buying criteria, GenAI can analyse the historical textual data and generate tailored responses in highly scalable ways.

Conversely, for tasks such as quantitatively analysing financial data to find correlations, classify information, or make predictions, a traditional AI system would be more appropriate than a standalone GenAI system.

2. Data sufficiency:

is there sufficient quantity and quality of data to train the model?

GenAI processes highly unstructured data, such as human queries, to produce desired outcomes. To achieve this, it relies on a substantial amount of high-quality data for training, fine-tuning, and generalisation. The quality and quantity of this data directly influence the accuracy, ease of generation, and relevance of the content produced. Additionally, the tokenisation strategy, which breaks down input data for model processing, can also affect these outcomes.

With comprehensive data for the domain and the right infrastructure and training, the model's ability to understand queries, context, and generate accurate outputs improves significantly.

To ensure effective solutioning, it is crucial to clarify data topics such as completeness, relevance, and balance. Addressing data quality involves considering synthetic data, data augmentation, resampling, and under-sampling. Additionally, Self-Play Fine Tuning (SPIN stands out as an advanced technique that enables large language models to enhance their capabilities by generating their own training data. Selecting an appropriate tokenisation strategy is also vital for LLMs, as it can help mitigate hallucination risks while impacting running costs.

Other related data topics include confidentiality and personal privacy treatment for data in transit, at rest and in archive. Remember, user prompts are likely to be retained for audit and investigative purposes and they can be regarded as business confidential data in which case, teams and vendors dealing in that GenAI system will need to observe banking confidentiality requirements; these prompts may also need to be retained for the duration of regulatory requirements.

Contributed intellectual property, such as using reinforcement learning with human-in-the-loop as intellectual property, can greatly benefit from early discussions among AI engineers, business professionals, and legal experts. Engaging in these tripartite conversations ensures a comprehensive understanding and strategic alignment.



3. Results materiality:

is the problem to be solved mission critical and would the GenAI system directly interact with external users?

For mission critical applications where the outcome can directly impact business reputation and clients, allowing external users access to your GenAI system can be high risks due to the less predictable nature of its creative outputs. For example, to use GenAI to generate investment strategies based on end investors queries would be risky not least because GenAI per se is unsuitable for quantitative statistical analysis. Requirements that '98% accuracy' is not good enough can point to a non-GenAI as a primary solution too.

Hence, considerations in deploying a GenAI system to a problem statement include the criticality of the problem to business operations, the level of GenAI/AI governance maturity in the organisation and whether the generated outcome can be validated by inhouse expert humans before being used. GenAI can also be used as a computer-human interface to accept imprecise language as instructions to trigger other deterministic tools to generate results.



4. Extent of human decisions:

are there multiple decision-making stages, or a decision waterfall, in the problem statement?

A problem statement utilising a 'what-if-then-else' decision structure is often better addressed by non-generative AI systems, with humans validating or making decisions at key points.

Generative AI can still play a role as a computer-human interface, complemented by traditional AI systems that manage specific statistical analyses and decision-making tasks, with human oversight involved.



5. Cost of solution:

the running costs of GenAI

From an economic perspective when deploying a GenAI system, several factors that impact running costs should be considered to ensure the sustainability of the AI solution. These factors include compute power especially for real-time applications. Storage and memory, as part of AI specific infrastructure that includes vector databases, RAG architecture and knowledge graphs, can grow with larger parameter models with longer token sequences. Hence, assessing token usage and context length should be performed to manage this cost driver.

A token-based pricing model can drive expenses particularly for frequent lengthy chat-based interactions that retain prior chats as context which is valuable but can be expensive.

Data transfer costs are another consideration as LLMs can involve API calls and bandwidth for data inputs and outputs. Ongoing maintenance and fine tuning of the model should also budget for retraining to enhance performance or accommodate new data. Hence, these new cost considerations need to be managed to allow the maximum number of users to access the system, and therefore, the magnitude of benefits and strategic advantages.

6. Jobs:

questions on job concerns



When AI or GenAI emerges as a solution to a problem, discussions often turn to job security, particularly if the investment promises transformation and efficiency. Addressing these concerns early on is vital to avoid misunderstandings. Highlighting the significant benefits of AI/GenAI—such as enhancing roles, boosting productivity, and expanding human capabilities—can lead to a more fulfilling work experience for employees.

AI and GenAI augment human potential rather than replace it. While it may be possible that AI/GenAI systems could reduce the number of positions in the longer time horizon, the primary value of implementing AI/GenAI right now is not about cutting jobs but rather to drastically increase the efficiency and the capacity of human workers.

In experiments conducted by Deutsche Bank using a GenAI system (Aggie) in collaboration with Kodex AI, Aggie significantly reduced the time needed to summarise and write complex regulatory text from two hours, or 120 minutes, to just minutes, all while maintaining maker-checker governance. This time savings allow expert staff to spend more on client interactions than on keyboards.

Once the decision is made to adopt GenAI systems, clear communication, active employee engagement, and opportunities for staff to re-skill are essential. These efforts will foster an environment of innovation, growth, and transformation. Figure 3 outlines their relevance in understanding the fit of the problem statement with GenAI.

Figure 1: GenAI models

| Considerations | | Relevance to problem statement fit with GenAI |
|----------------|--|---|
| 1 |  | Create new content or to classify/predict? Assess whether the solution requires generating creative content, suitable for GenAI, or if it involves structured tasks such as classification or prediction, which may not necessitate GenAI |
| 2 |  | Sufficient quantity and quality of data The performance of GenAI largely relies on robust training data. The nature of this data, whether public or private, can significantly influence the time-to-market and complexity of the solution |
| 3 |  | Mission criticality or "I cannot accept 98% accuracy" For mission-critical tasks requiring absolute accuracy, a layered solution might be necessary. This could involve integrating GenAI with other technologies and implementing process governance |
| 4 |  | "What-if-then-else" decision structure Problems governed by structured rule-based logic are well-suited for traditional AI, whereas GenAI excels in handling complex, creative, and ambiguous scenarios |
| 5 |  | Business case benefits Is the key benefit centred on business non-technical users for productivity gains which GenAI can catalyse such benefits, or for specialised/tech teams for niche non-generative applications which other AI systems could better fit? |

Source: SES Views, Deutsche Bank

02

Building a portfolio of GenAI use cases

Even when a problem seems to require a GenAI solution, navigating the business case, governance, and compliance processes for GenAI can be significant, making a one-time effort inefficient. Instead, a strategic, composable approach to incrementally scale its applications across different use cases can better support the expected return on investment (ROI) from this powerful productivity and competitive tool. This method allows the organisation to learn and build trust in these tools with each successful application, enabling risks and controls to mature progressively and allowing for more accurate ROI estimates and realisation.

2.1: Three-stage approach

To harness its full potential, we suggest a three-stage composable approach to develop a GenAI roadmap of use cases, where each use case and stage builds on the prior ones yet will have tangible benefits to be delivered to the organisation.

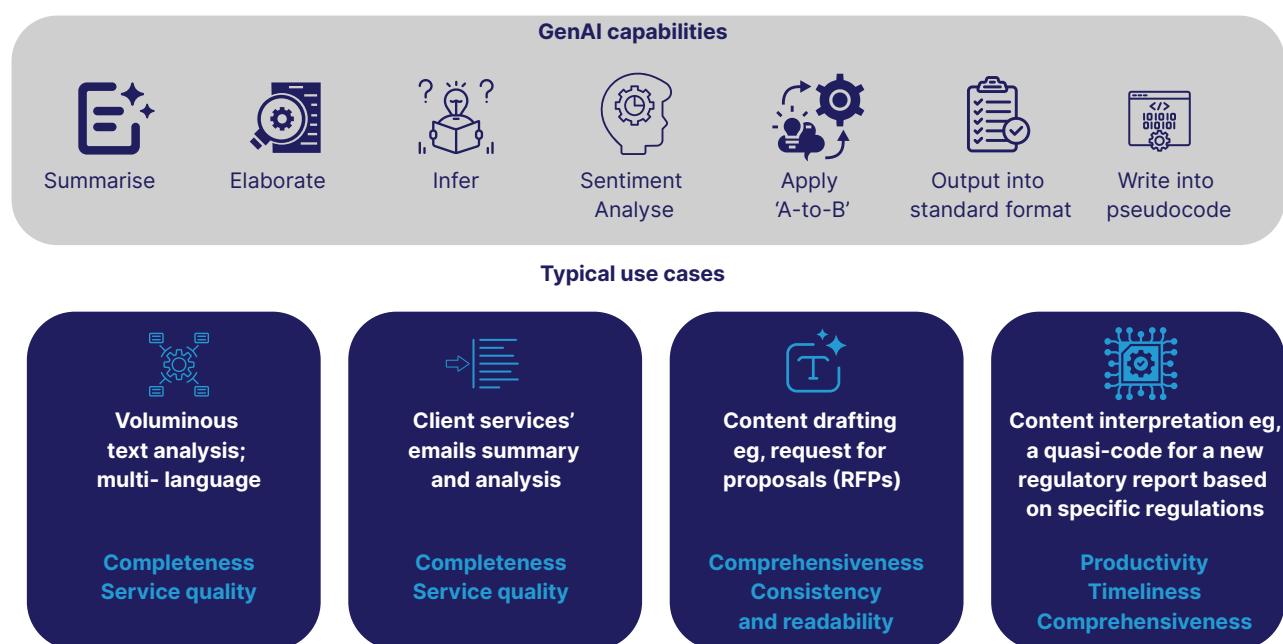
Figure 4: Composable portfolio for GenAI applications



2.1.1 Stage 1: apply GenAI's core text and language analysis capabilities

Core GenAI language capabilities allow the system to accurately understand and interpret natural language queries to perform text analysis tasks like summarisation, language and comprehension that Figure 5 illustrates.

Figure 5: Text generative capabilities (non-exhaustive)



Source: Project Aggie with Kodex AI GmbH and Deutsche Bank

This stage builds up the accuracy and capabilities of the natural language handling capability for the domain even as it is applied to solve domain problems like better client service summaries, drafting content that contribute to value creation. This first stage lays the foundation for the next stage “Chat-to-Agent” use cases, where the human natural language query is translated by an executing agent into a precise database and code commands for execution.

2.1.2 Stage 2: chat-to-agent use cases

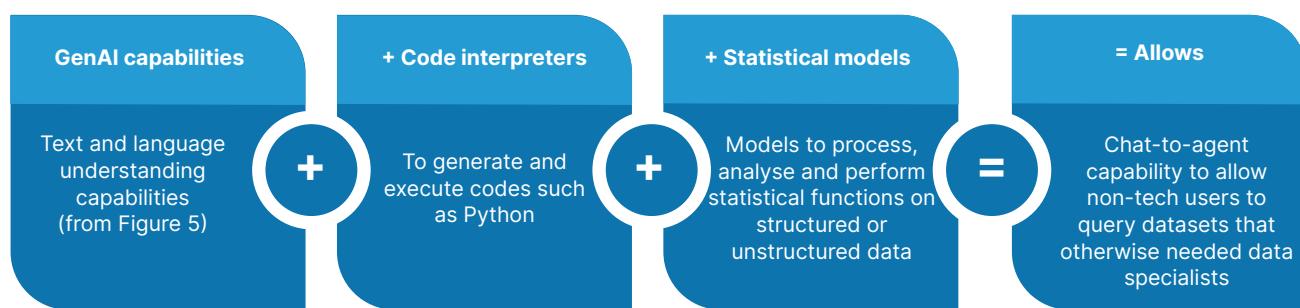
A ‘chat-to-agent’ solution is where human language queries would trigger the appropriate tools or models by an executing agent to perform specific tasks. An executing agent can be a Python program that receives the queries to write code and execute for results, acting as a pipeline where language models, code interpreters and other AI models would work in harmony.

While text analysis alone is based on language processing, use cases in this stage will see GenAI systems calling other libraries and AI models to allow users

to tackle more complex and multi-step tasks like generating codes, querying databases or automating workflows by using natural language as a primary human-computer interface.

A use case at this stage can include data analytics that business users can perform using natural language queries to understand data patterns and relationships.

Figure 6: linked GenAI-AI capabilities



Source: Project MILA 2023 with Jia-Wei Lee, NUS STEM intern with Deutsche Bank

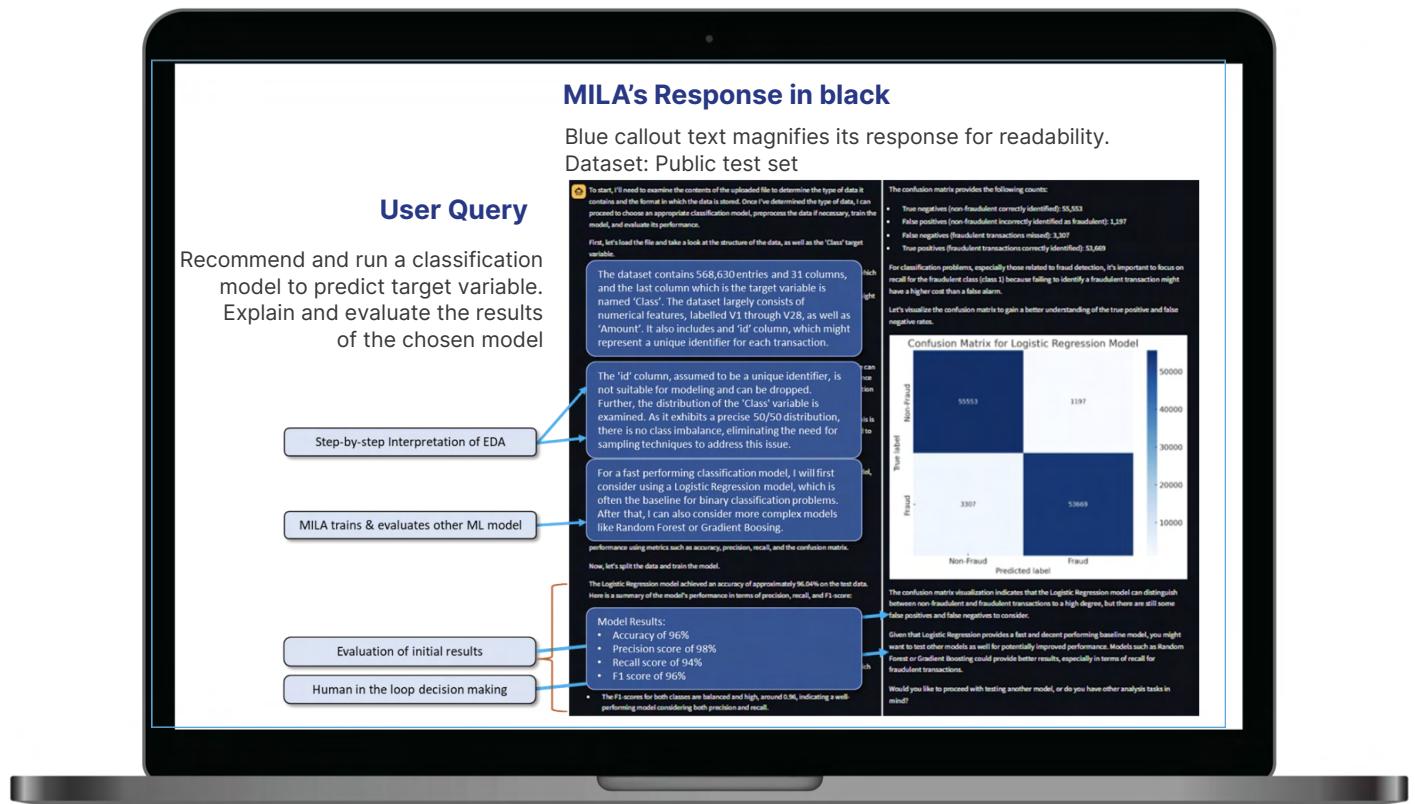
For example, in a 2023 wireframe experiment (Project MILA), a ‘chat-to-agent’ model directly helped business users to understand relationships between multiple factors with a comprehensive analysis that was also accompanied by visualisation. MILA was also integrated with ReACT (reasoning and acting) and self-reflection mechanisms to reason through complex problems, and provided transparency to users, which allowed it to decide when to call specific libraries or models based on the query received.

Such basic self-reflection qualities enabled MILA to evaluate whether its output was aligned to the query, and to seek human feedback for next steps.

The experiment used a labelled structured public test data set with 31 features and about 520,000 entries. A user asked MILA to “analyse and help me understand the insights and relationships in this data set” as shown by Figure 7. MILA took the human query, translated it into Python code requirements and performed the following steps:

- 1 Data structure discovery with step-by-step interpretation of the exploratory data analysis;
- 2 Identified dataset features as possible unique identifiers or those that could be dropped from the analysis;
- 3 Analysed data balance/imbalance and proposed appropriate sampling techniques ;
- 4 Firstly called on more transparent and straightforward algorithms to perform statistical analytics;
- 5 Evaluated the initial results including Precision, F1 scores and Recall;
- 6 Asked the human evaluator if the results are satisfactory, which if not, MILA would call on other algorithms to retry; and
- 7 Output both statistical analysis with natural language explanations and analysis;

Figure 7: screenshots based on actual execution, no real data is used



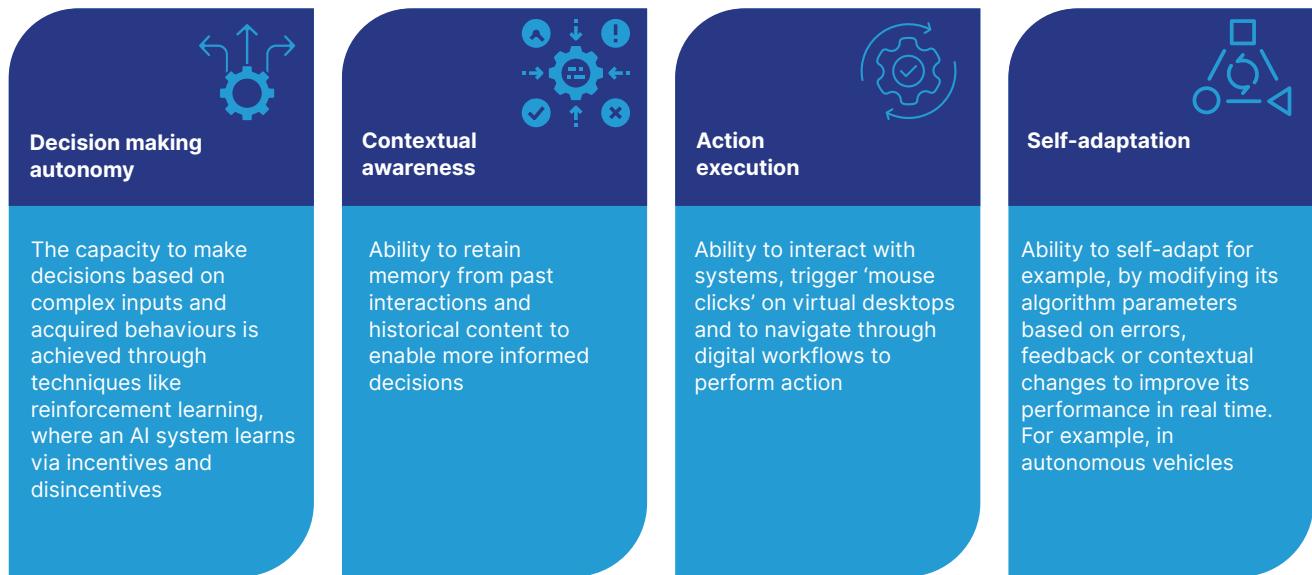
Source: Project MILA 2023 with Mr Jia-Wei Lee, NUS STEM intern with Deutsche Bank

It was empirically estimated that an experienced data engineer would require about 3 hours to perform the above seven steps that MILA took about three minutes to complete including asking the user to evaluate initial results.

Chat-to-agent empowers non-technical users to perform tasks that once required data scientists, through intuitive no-code queries. However, rather than rendering the data scientists obsolete, this shift should allow them to work on more advanced and interesting tasks such as fine-tuning systems for domain deployments. The data scientist roles evolve from executing manual data operations to developing more intuitive and accurate AI systems.

2.1.3 Stage 3: chat-to-execution. The autonomous capabilities

The transition from a “chat-to-agent” system to an Agentic AI system marks a significant advancement in autonomous digital capabilities. In a system like the MILA experiment, users make requests in natural language, which are then converted into specific actions, such as querying a dataset. The “chat-to-execution” stage elevates this by adding the AI system’s autonomy in decision-making, contextual awareness, and action execution capabilities to the foundational “chat-to-agent” framework (Figure 8).

Figure 8: Additional qualities as an agentic AI (non-exhaustive)


Source: crewai.com, DeepLearning.AI, various

As GenAI evolves from simply generating text responses to autonomously making decisions and taking actions, it introduces concerns about new risks and ethical challenges. The constant evolution of current transparency and explainability methods are crucial, improving systems logging to ensure accountability by human overseers. Appropriate data privacy and cybersecurity measures will continue to be necessary, especially if advanced GenAI systems have access to sensitive, commercial, or personal information. New practical accountability models can be required as autonomous AI agents can blur traditional lines of responsibility.

Integrating GenAI applications into a firm's operating model opens vast opportunities for innovation, automation, and competitiveness. However, it is crucial for firms and regulators to establish clear guidelines and adaptive sandboxes to learn, prevent preventable errors and plans for unintended consequences. This approach fosters a balance between innovation, growth, safety, and regulation, enabling the industry to effectively harness the power of GenAI systems.

Each GenAI system is unique, and the following chapters offer insights into key components used which differentiates the quality between systems.

03

Identifying excellence

With a growing range of available GenAI solutions, determining the quality of a GenAI system that looks to be similar but can perform differently should become a key determinant for organisations, particularly those in highly regulated sectors like financial services.

The investment into a GenAI solution is best justified when the GenAI system delivers on a range of crucial quality indicators that are aligned to the use case. Understanding what constitutes a high-quality system is essential for decision-makers, especially when implementing GenAI for financial industry activities.

3.1 Accuracy and relevance (benchmarks)

One of the best indicators of quality is the accuracy and relevance of the model's outputs. Benchmarks provide industry agreed metrics as a basis to compare different LLMs, indicating which model performs relatively better against a common minimum standard. Additionally, they reveal the progress of an individual LLM as it learns and enhances over time.

3.1.1 Benchmarks at the foundation model level

LLM benchmarks consist of meticulously crafted tasks, questions, and datasets that assess a language model's performance in standardised manners. These benchmarks can consist of diverse tasks, datasets, and evaluation metrics that test a model's capabilities across a range of areas such as natural language understanding, reasoning, and knowledge retrieval. By comparing performance across different LLMs, benchmarks provide standardised and objective measures of quality. Some of the most highly regarded general LLM benchmarks include:

GLUE (general language understanding evaluation)

GLUE is one of the most widely used benchmarks for evaluating LLMs. It consists of a variety of tasks that test a model's ability to perform sentence-level classification, sentence similarity, and textual entailment. High performance on GLUE reflects a model's general competency in understanding and processing natural language.

SuperGLUE

An extension of GLUE, SuperGLUE is a more challenging benchmark designed for models that have surpassed the performance limits of GLUE. It introduces more difficult tasks that require deeper reasoning and problem-solving, making it an essential bench-mark for evaluating cutting-edge models.

LAMBADA

Tests the ability of a model to predict a missing word in a narrative context, focusing on text coherence and contextual understanding while Winograd Schema Challenge (WSC) evaluates co-reference resolution capabilities and commonsense reasoning for text use cases.

SQuAD (Stanford question answering dataset):

Tests a model's ability to answer reading comprehension questions based on a passage of text. It is widely used to evaluate how well a model can extract relevant information from text and answer fact-based questions with precision.

MMLU

MMLU (massive multitask language understanding): MMLU tests a model's ability to handle a wide range of tasks across numerous domains, including STEM, humanities, and social sciences. This benchmark evaluates how well models generalise across different subject areas, which is crucial for assessing versatility and depth of knowledge. Also has a translation subset focusing on a wide range of languages, including Chinese, Indonesian, and other non-English contexts. It tests how well a model can translate complex text between languages, making it a good benchmark for cross-linguistic performance evaluation.

BIG-bench (Beyond the imitation game benchmark):

This large-scale benchmark focuses on testing models in diverse, challenging, and open-ended tasks. It includes complex reasoning, mathematics, and world knowledge tasks, providing a thorough evaluation of an LLM's advanced reasoning capabilities and real-world problem-solving skills.

Chat-to-Agent

For Chat-to-Agent use cases, Dialogue Natural Language Inference (DNL) measures a model's ability to maintain consistent, logical, and contextually accurate dialogue responses. Whereas MultiWOZ (Multi-Domain Wizard-of-Oz) is a dialogue dataset that spans multiple domains and intents, testing the ability of a model to perform complex goal-oriented conversations.

Chat-to-Execution

For Chat-to-Execution use cases, THOR Benchmark evaluates the ability of agentic models to execute actions and plans in a simulated environment based on natural language instructions. While ALFRED (Action Learning From Realistic Environments and Directives) measures a model's ability to follow complex, multi-step directives and interact dynamically with a simulated environment.

3.1.2 Domain specific benchmarks

After the foundation model has been fine tuned for domain specific applications (we refer back to Figure 1 on GenAI Types), the application should now be tested against specific standards which in this context would be the financial services. Some of these specific benchmarks are:

FinanceBench

FinanceBench evaluates models based on their ability to process and interpret financial data accurately, making it critical for assessing models that handle market analysis, risk assessments, or regulatory compliance reports. The detailed nature of the benchmark tasks ensures that the GenAI model can handle intricate financial datasets, such as balance sheets or regulatory filings.

FinQA

FinQA focuses on question-answering capabilities specific to financial contexts. It tests how well a GenAI system can handle fact-based queries drawn from financial reports, earnings calls, and other structured financial documents, ensuring the model provides not only accurate but also contextually relevant answers.

FNS

FNS (financial narrative summarisation) evaluates a model's ability to summarise complex financial narratives from dense data sets such as earnings reports or annual reviews. This helps organisations assess a model's potential for automating the generation of key insights from voluminous financial text data.

Financial benchmarks ensure that models are stress-tested in the environments they will operate in, reflecting real-world complexities and regulatory expectations. Another benchmark would be those for language translations for the financial industry for non-English working language markets. Without these specific financial tests, an AI model might perform well in general language tasks but can still fail to meet the high standards necessary for tasks like compliance reporting or financial analysis, leading to potential issues and risks.

Models that integrate retrieval-augmented generation (RAG) strategies, for example, enhance the relevance and precision of their responses by retrieving and synthesising external data sources in real time. The next chapter elaborates on main architectural and process components that influence the benchmarks.

3.2 Architecture and Process Factors

While the underlying core LLM is a crucial part of any GenAI application, it is only the tip of the iceberg. A robust and effective GenAI system consists of different interconnected components that work together to deliver meaningful and contextually relevant outputs. These components include data handling, Retrieval-Augmented Generation (RAG) strategies, fine-tuning methods, and pre- and post-processing techniques; and there are others too like Knowledge Graph, memory management and others. We highlight some of these factors below that play important roles in the functioning, reliability and trust of a GenAI – or indeed any AI – system.

3.2.1 Data

Data forms the bedrock of any GenAI system. There are two critical types of data to consider: training data and RAG (Retrieval-Augmented Generation) data. Curated training data is used to fine-tune the LLM to a specific industry, topic domain, or use case. On the other hand, RAG data is specific to real-time applications; it consists of the systems knowledge bases, often structured documents, that are queried during inference to provide more accurate and relevant responses.

RAG is an advanced technique that augments the generative capabilities of AI models by retrieving relevant information from external data sources. Different strategies can be applied here, from basic keyword search-based retrieval to more sophisticated semantic search mechanisms that leverage embeddings and vector databases. An emerging strategy is Knowledge-GraphRAG, which uses structured data stored in a knowledge graph to improve response accuracy and context relevance. This approach allows the GenAI system to tap into more complex relationships between data points, thereby providing richer and more meaningful outputs.

Fine-tuning a GenAI model involves adjusting its parameters to better suit specific tasks or domains. Several approaches are available, including Parameter-Efficient Fine-Tuning (PEFT), Low-Rank Adaptation (LoRA), and its quantised version, qLoRA. These methods enable effective fine-tuning with significantly reduced computational resources.

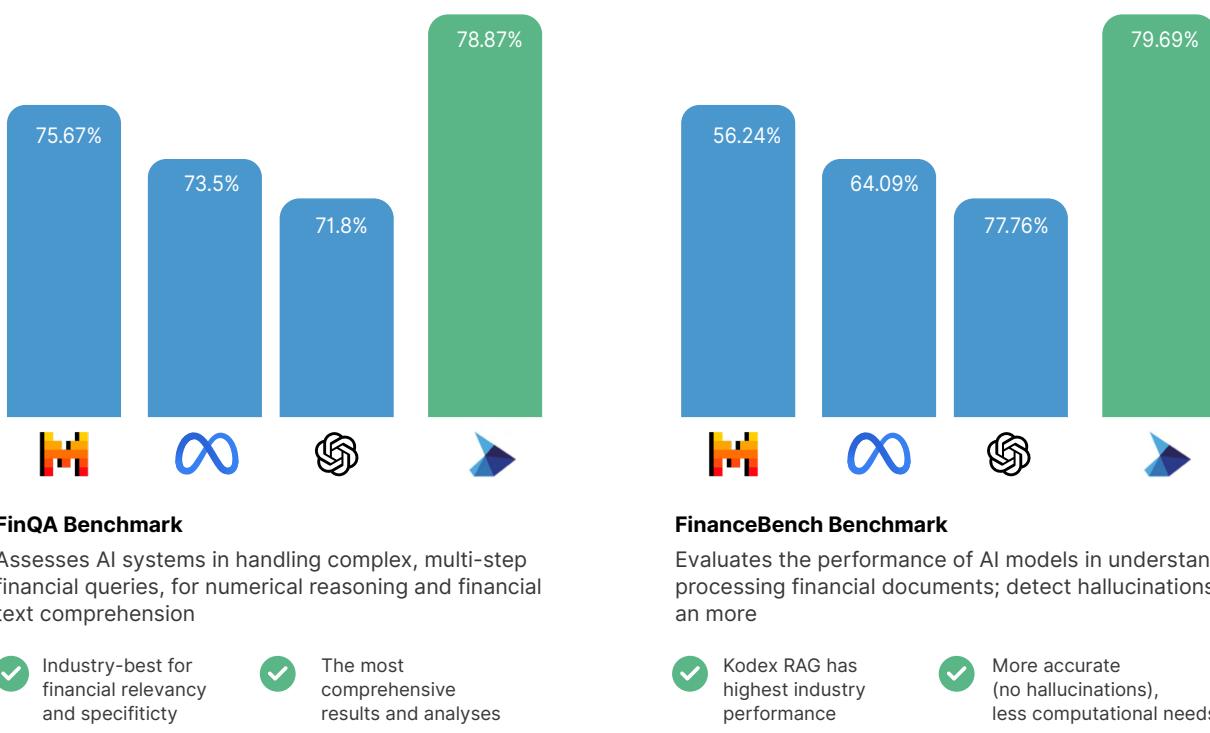
Another technique, SPIN (Selective Parameter Injection), focuses on injecting domain-specific knowledge directly into specific layers of the model. Each method offers distinct advantages, making it essential to choose the right strategy based on resource availability, desired output quality, and specific business needs.

Chunking, Parsing and Content Filters are pre- and post-processing steps that can be overlooked but which are vital to the success of a GenAI system. Pre-processing tasks involve parsing and chunking multimodal data to ensure the model handles different data types effectively. Content filtering is

another crucial step to remove unwanted or harmful content, ensuring outputs align with business and ethical standards. Additionally, question classifiers can be employed to guide the model toward relevant RAG data and tailor responses more precisely.

Meta-prompt templating helps structured model outputs for better readability and consistency. Source highlighting is a post-processing method that adds explainability and transparency by indicating the origin of the generated content, which is particularly important in enterprise applications where verification is crucial.

Figure 9: Financial benchmark performance: specialised versus generalised



3.2.2 Customisation

Whether through fine-tuning or model customisation, GenAI systems should be able to be deployed seamlessly with an enterprise's existing workflows and data environments.

High-quality models leverage techniques such as Parameter-Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA/qLoRA) allow the GenAI system to be customised without requiring massive computational resources. These features are particularly valuable in industries like banking, where data privacy is vital and localised, secure training can offer both customization and compliance advantages. Benefits that come from such techniques include faster responsiveness, more adaptive deployment and logical data separation to ensure confidentiality between groups of different users within a firm.

Planning also needs to consider if the system will need to access data from multiple different domains, whether such data is streaming or not and how the infrastructure can also support subsequent use cases.

3.2.3 Data and Training for language and cultural sensitivity

Language and cultural sensitivity ensure that output from the system—whether for internal or client-facing—can reflect precision, accuracy and intent of the culture that it serves. High-quality GenAI models, especially those tailored for financial services, should be capable of processing and generating content that reflects linguistic nuances, context, and cultural sensibilities across different use cases within that specific market. Having those abilities would reflect the intensity of training, the infrastructure that runs the system and the human expertise involved. The right type of training data set and subject matter trainers who are effectively bilingual are key success criteria.

3.2.4 System's creativity, reasoning and problem-solving

Beyond basic task execution, a high quality GenAI solution should exhibit an acceptable degree of creativity, reasoning and problem-solving ability. Reasoning ability includes ReAct and Chain-Of-Thoughts explanation. Whether it is generating insights from financial reports or developing new strategies for regulatory compliance, quality GenAI systems such as those in a Chat-To-Agent are those that can propose, reflect and test relevant solutions to complex queries with inputs from human decision makers.

3.2.5 Speed, performance and costs

Performance efficiency—measured through response time and system throughput—is another key aspect that differentiates quality GenAI systems from others. Speed influences the level of user and client experience, and real-time data processing, handling large datasets or providing outputs based on multi-turn chats in complex financial scenarios. If the model is optimised and quantised, this can also mean lower running costs which will be important for sustainable deployments.

Equally important is the cost-effectiveness of deploying quantised models like QLoRA, which significantly reduce the computational overhead and energy consumption compared to full-size models. By ensuring the model size is appropriate for the complexity of the task, organisations can optimise the trade-off between accuracy and resource usage, resulting in better performance without excessive costs. Speed and performance are also driven by hardware factors such as memory size, processor speed, and geographic proximity between users and cloud servers, which can influence latency and overall user experience.

3.2.6 Support and upgrades

Long-term quality also involves continuous support, updates and upgrades. Financial environments evolve rapidly, and GenAI models must adapt to new regulations, emerging market trends, and evolving user needs, and the potentials of retraining of the fine tuned model. Infrastructure will also need to be future-proofed to allow upgrades of model, vector database, RAG, and other components, thereby incorporating “no model lock-in”, “no cloud lock-in” and/or “no database lock-in” design principles.

Close collaboration between AI engineers and business users allow fast iterations to synergise system evolutions and fixes to address pain points and align with user expectations. Regular iterations based on user feedback can significantly enhance both usability and relevance, leading to higher adoption rates and better results. Additionally, a well-designed GenAI system should avoid model lock-in by supporting modularity and interoperability. This means enabling seamless migration to more powerful models and alternative frameworks without disrupting existing workflows, thus ensuring flexibility and the capacity to integrate future advancements as they become available.

04

Factors that compromise quality

The prior section has highlighted some components that drive a GenAI system's accuracy and trust, which in turn reflects the extent by which some of the key data and model risks of these components are addressed. This section highlights the associated key risks and their mitigants.

4.1 Data risks

In the development and deployment of GenAI systems, data quality and associated risks are critical factors that significantly impact the system's performance, accuracy, and reliability. For highly regulated sectors like financial services, where the margin for error is low, ensuring robust data handling processes and practices is crucial. Risks can arise from various stages of the model lifecycle—from training to deployment—and mitigating these risks requires a combination of technology tools and human oversight. The quality of data used to train GenAI models is foundational to the system's success. Poor-quality, biased, or incomplete data can lead to inaccurate outputs which can compromise the integrity of decision-making processes.

4.1.1 Data risks mitigants

Several techniques can be employed during the training phase to mitigate these issues:



Data cleaning and preprocessing: Before feeding data into the model, it is essential to clean and preprocess it to eliminate noise, redundancies, and inconsistencies. This ensures that the GenAI system learns from reliable and accurate information.



Bias mitigation: Addressing inherent biases in the training data is necessary. Techniques such as reweighting data samples and adversarial training can help reduce the likelihood that the model will replicate or exacerbate biases in its outputs.



Continuous monitoring: Once a GenAI system is trained, continuous monitoring of the model's performance on real-world data helps detect any drifts or deviations from expected outcomes.

4.1.2 Synthetic data to enhance training

One emerging solution to data scarcity and privacy concerns is the use of synthetic data. Synthetic data, generated through algorithms that mimic real-world data, enables organisations to train GenAI models on datasets that reflect real-world conditions without exposing sensitive or personally identifiable information (PII). This approach is particularly useful in financial services to navigate data protection regulations.

Examples of synthetic data in action include creating financial transaction datasets for fraud detection models or generating customer profiles to train recommendation engines. By leveraging synthetic data, companies can maintain high standards of data privacy while ensuring that their GenAI systems are trained on diverse and representative datasets.

Mitigating data risks and ensuring quality in GenAI systems requires a holistic approach that integrates cutting-edge data techniques, synthetic data solutions, and continuous human oversight.

4.2 System risks

Data and system are two closely interconnected and interdependent factors that drive a number of implementation details and safeguards. As GenAI systems become more integral to augment decision-making, understanding the possible type of model and system risks is critical to ensure sustained accuracy, reliability and trust. Model risks such as drifts, hallucinations, and degradation from feedback loops can undermine the system's precision and effectiveness. The following explains further.

4.2.1 Model Drifts

GenAI systems can also suffer from model drift, where performance degrades over time as its output starts to deviate from the data it was initially trained on. This happens when real-world data start changing leading to mismatch between the training data and the data the model encounters on a daily basis. Certain use cases are more susceptible to drifts, such as client service where daily client questions (behaviour) starts to differ because of product changes that the original data has not captured for the model's training.

To mitigate this risk, continuous monitoring with metrics such as prediction accuracy, error rates and consistency of answers can flag when the system's results are starting to drift from acceptable boundaries.

Automated alert systems can notify administrators when performance metrics fall below predefined thresholds. These alerts enable rapid intervention, ensuring that models are retrained or adjusted before their outputs lead to significant errors.

In conjunction with monitoring, regular retraining on updated datasets is essential to keep the model aligned with current trends and information. For example, a model fine-tuned for regulatory analysis and reporting should be considered for retraining when existing regulations or new AI-specific regulations are to be accurately interpreted for GenAI characteristics.

Model drifts and hallucination are related but they are different issues; the former relates to data match between real world data and training data, while the latter involves falsehood.

4.2.2 Hallucination risks

Hallucination risk is where the model generates outputs that are plausible-sounding but factually incorrect or irrelevant. The risk is related to the nature of GenAI but it does not mean it is chronic or cannot be minimised. Mitigants to this risk includes:



Source verification systems: One method for mitigating hallucinations is to implement retrieval-augmented generation (RAG) techniques, where the model cross-references external, verified data sources to ensure the accuracy of its outputs. This is particularly important in contexts where the model is required to generate responses based on complex or specialised knowledge, such as regulatory compliance or legal interpretation.



Human-in-the-loop oversight: Expert human oversight also plays a critical role, firstly in training, and then in identifying and rectifying hallucinations. By integrating human-in-the-loop (HITL) systems, organisations can have human evaluators review and correct model outputs, especially for high-stakes decisions.



Conservative model settings: For critical tasks, configuring the model to favour conservative outputs (where uncertainty is high) can reduce the risk of hallucinations. Instead of generating speculative responses, the model can be set to signal uncertainty or prompt human intervention when it lacks confidence to answer the query.

4.2.3 Feedback loop degradation: addressing user influence

The quality of GenAI systems can degrade over time due to feedback loop degradation, a situation where user interactions inadvertently reinforce undesirable behaviour in the model. This issue often arises in systems that rely heavily on user feedback for learning and optimisation. For example, if a GenAI system in customer service receives frequent but incorrect user feedback, it may learn to prioritise irrelevant or incorrect responses over time.



Feedback filtering mechanisms: To counteract this risk, systems must incorporate robust feedback filtering mechanisms that evaluate the quality of user inputs before using them to influence future outputs. Not all feedback is equal, and the system must be able to discern between valuable inputs and those that could degrade its performance.



Controlled retraining cycles: Rather than relying on continuous learning from user feedback, organisations can implement controlled retraining cycles. This allows time for proper evaluation and validation of feedback before it influences the model's behaviour. Controlled cycles help ensure that only high-quality data is used to update the model, maintaining its integrity.



Diverse feedback sources: Another method for mitigating feedback loop degradation is to introduce diversity in feedback sources. Relying too heavily on a small set of users or a specific subset of interactions can lead to overfitting and degradation. By integrating feedback from a broad range of users and scenarios, the model can maintain a more balanced and accurate output profile.

4.2.4 Mitigants

There are also several mitigants that can be deployed to address these risks. These include



Human-in-the-loop (HITL) evaluation: While automated systems can handle vast amounts of data, human oversight remains a key mitigant for ensuring model quality and ethical decision-making. Human-in-the-loop (HITL) methodologies involve human evaluators at various stages of the model lifecycle, particularly in the areas of:

2

Model training and validation: Before deployment, human experts validate the outputs of GenAI systems to ensure that they align with industry standards and ethical guidelines. This is especially important in financial services, where errors can lead to regulatory violations or financial losses.

3

Ongoing feedback loops: In live environments, HITAL systems allow for continuous evaluation, where human feedback is incorporated to refine the model's outputs over time. This iterative process ensures that the model adapts to evolving conditions and remains aligned with the organisation's goals.

4

Robust model governance framework: includes version control, continuous validation, and anomaly detection systems that flag irregularities in model behaviour.

4.2.5 Model evaluation techniques as mitigants

To ensure that GenAI models perform optimally, a variety of evaluation techniques that are in addition to benchmarks can be applied. For example,



Cross-validation: During model training, cross-validation techniques are used to assess the model's performance across



Performance metrics: Models are evaluated using performance metrics such as precision, recall, and F1 score, which measure the accuracy and relevance of the outputs. These metrics are especially important in compliance and risk management applications, where high precision is critical.



Scenario testing: In the financial industry, models are often tested on edge cases or rare scenarios to ensure that they perform robustly under all conditions. For example, GenAI models used in market predictions might be tested against historical data from financial crises to assess their resilience.

4.3 Other risks

4.3.1 Dependency risks

GenAI requires specialised infrastructure and expertise which can create dependency risks on models, databases and providers, and such risks can be mitigated with design principles and architecture that allows transferability of model, databases, Cloud and other core components. For example, vector data format portability for data transfers and ensuring dependencies such as libraries can be independently changed or updated. Other considerations include:



Open-source alternatives: Open-source AI models, such as those from Hugging Face or similar platforms, offer more flexibility and control, allowing organisations to tailor solutions according to their specific requirements.



Partner with specialised GenAI startups: Collaborating with smaller AI firms that can provide access to niche technologies and expertise.



Implement multi-cloud strategies: Reducing the dependency on a single vendor while allowing a broader range of services.

4.3.2 Cybersecurity risks

In financial services, security and compliance are non-negotiable. A high-quality GenAI solution should adhere to the battle-tested industry cybersecurity standards including those that are related to the uses of third party open-source codes, and forward-looking human expertise to ensure data protection and system's resilience. Implementation needs to consider LLM-specific types of attacks including data poisoning, prompt injection and adversarial ones that can lead to data leakage, misleading information or generation of harmful outputs.

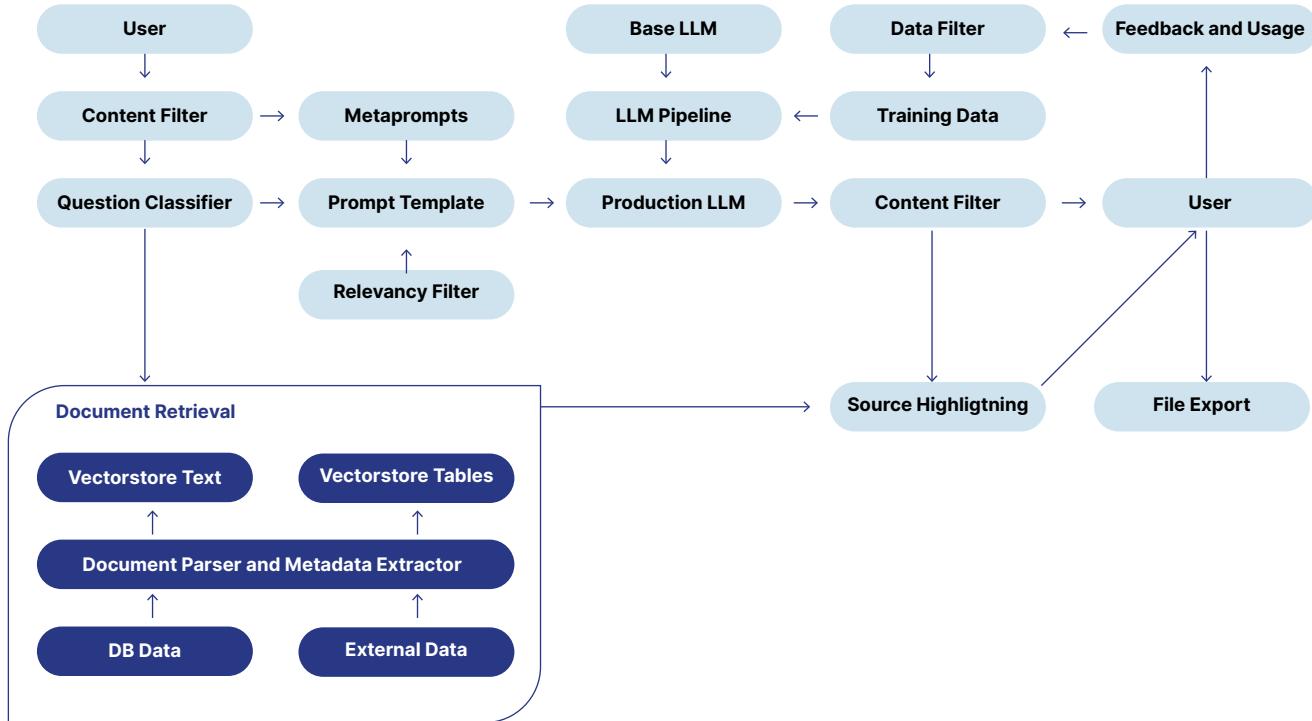
Figure 10: Attacks specific to LLMs

| Attack Type | Description | Probability | Likelihood (same user over time) | Implications |
|-------------|--|--|----------------------------------|--|
| | Data Poisoning Malicious data is injected into the training dataset to alter model behaviour | Low to Medium. Depends on foundation model and fine tuned model data sets | Low | Biassed, misleading or harmful outputs |
| | Prompt Injection Manipulate model behaviour with misleading input prompts | High | High | Biassed, misleading or harmful outputs |
| | Adversarial Inputs crafted to exploit known model weaknesses | Medium | Medium | Biassed, misleading or harmful outputs |

Source: Securities Services, Deutsche Bank

See the figure (Figure 12) below for an example of a GenAI system architecture that is created to minimise cyber risks including risks of inappropriate results.

Figure 11: High level architecture with filters and feedback loop



Source: Kodex AI

4.3.3 Sustainability risks

The energy consumption associated with training large models can contribute to a company's carbon footprint, especially given the high computational requirements for training and deploying such models. One approach to mitigate these impacts is the use of smaller models, quantised models and/or green data centres that leverage renewable energy sources. Another is on efficient model design which optimises the efficiency of model training and inference. Techniques such as Parameter-Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA/qLoRA) allow models to achieve high performance without requiring extensive retraining on new datasets, thereby reducing the computational power needed. These methods can cut down on energy use while maintaining the model's effectiveness.

05

Implementing GenAI & Recommendations

Even after assessing a GenAI as being fit for use and addressing the risks of the systems, there are still significant regulatory, ethical and market considerations that are crucial for successful implementation. This section shares a selection of insights and challenges that can still arise on the road to deployment, and makes some recommendations that private-public sector forums and collaboration can consider to advance the uses of this technology in the financial industry.

5.1 Regulatory considerations

Recognising its power to transform, policymakers and regulators in capital markets globally are also increasingly concerned about the unknown and extensive impacts of GenAI and AI advancements on both markets and individuals.

Currently, policy and regulatory responses can be categorised into several approaches: horizontal AI regulations that apply across all AI types, vertical AI regulations that target specific AI types and their associated risks, ethical guidance, and industry positions asserting that existing regulations are adequate to govern GenAI and AI systems without adding further compliance burdens. Figure 13 provides a non-exhaustive inventory of regulatory topics pertinent to GenAI systems, highlighting the compliance challenges organisations can encounter as they work to implement GenAI.

Figure 12: Regulatory topics relevant to GenAI systems, not in any order of priority**Related relevant regulatory topics (not exhaustive)**

Ethics, Fairness, Explainability and Transparency

Data Quality, Completeness and Bias

Outsourcing and vendor risks management

Model Evaluation and Algorithmic Discrimination

Confidentiality and Privacy

Third Party Risks and Open-Source Code

Accountability and Shared Responsibility

Localisation

End User Agreement

Content which includes harmful content, misinformation, falsehood

Intellectual Property and Copyrights

Technology Risks Management, Cybersecurity and Resilience

Environment, Sustainability and Employment

Cloud and Infrastructure

Market Manipulation, Systemic Risks

Source: Authors' views, not representative of Deutsche Bank or Kodex AI

However, the technology itself is neither inherently good nor bad, even though it may mirror the philosophies of its creators; rather the degree of risks depends on the use case where the technology is applied.

Therefore, implementing GenAI systems will require detailed understanding of the use case, context and applications, and careful articulation of how technology would meet both the business goals and regulatory adherence in a manner that is cost-effective, and not unnecessarily complicated.

Clarity and continued public-private discussions on key topics to streamline governance can be tremendously helpful to the global financial industry. Not new regulations but rather, an equivalent of a mind-map that goes across all the different existing regulations to link sections that are relevant to GenAI can be most helpful for effective and efficient start to compliance and adherence.

5.2 Explainability and transparency

GenAI text-to-text handles imprecise human language and is “creative” by its very nature. However, the probabilistic characteristics of GenAI outputs, along with concerns over accuracy, security, and data privacy, can be significant barriers to adoption which needs trust in the system’s outputs. This is why a sequenced composable approach (Figure 4 refers) to implementing GenAI can be helpful.

Related to the topic of trust is the issue of understandable explainability and transparency in GenAI model’s output. It is recognised that LLMs are complex and therefore, its explainability and transparency can be complex but not understandable by lay persons, or risks being too simplified. Together with the public sector, the financial industry can benefit from clarity on how these values can be satisfactorily achieved. Currently, techniques followed to ensure explainability and transparency include[‡]

Model and process documentation: Documenting the training data, model architecture, and decision pathways.

Source attribution: Highlighting clearly the source of any generated content.

Audit trails: Retaining the prompts and recording the model’s decision-making process to enable comprehensive post-hoc analysis.

Human oversight: Integrating human-in-the-loop (HITL) systems to review and validate critical outputs.

Navigating ethical guidelines require a robust governance framework that incorporates data quality and accountability. Organisations should consider AI-focused forums and oversight teams to guide GenAI deployments, ensuring that these systems meet both internal standards and external regulatory expectations.

5.3 Data and Hallucination

The effectiveness of GenAI training and the accuracy of its outputs hinge on the quality and comprehensiveness of the data used. High-quality data significantly reduces the risk of “hallucinations,” which we define here informally as errors, falsehoods, or outdated information. GenAI models depend on training data to be finely tuned for specific applications, ensuring that the content generated aligns with the requirements. Incomplete or low-quality training data can result in flawed, misleading, or biased outputs.

For example, models trained on unrepresentative data would only be effective within a narrow context. There can also be data distribution mismatch – that is where the data set that trained the model does not match the actual live ways it needs to respond. When systems extend beyond such scope, the resulting

outputs can be misleading, out-of-date, with falsehood to have harmful implications depending on the application.

It's not just the data that matters. A lack of suitable infrastructure to fit business requirements, from either insufficient expertise or budget, can also lead to inaccurate or hallucinated outcomes. Key factors such as text tokenisation strategies, Retrieval Augmentation Generation (RAG) methods, dynamic content filters, and the involvement of human domain experts throughout the fine-tuning and training stages play crucial roles in the precision of a GenAI's output.

Therefore, while hallucination is a risk related to GenAI, it can be effectively managed via data pre-post processing, architecture that includes input-output content filters, training, retraining and reinforcement learning by domain experts as well as hardware considerations including RAG and memory size. Users setting the creativity level of the system, for example through the "Temperature level", Top K and/or Top P, can also influence the degree of creativity-hallucination.

Concerns of this risk are legitimate but should be grounded by these factors and mitigants.

5.4 Synthetic data

Synthetic data can augment incomplete data sets and also offers the potential for privacy preservations.

Utilising synthetic data or data augmentation techniques like Self Play Fine Tuning (SPIN) to tackle incomplete datasets presents both opportunities and challenges. On the positive side, synthetic data can effectively mitigate privacy concerns and address issues with limited or biased datasets by offering more diverse examples. However, it also necessitates scrutiny to ensure that the synthetic data remains representative and accurate without perpetuating the inherent weaknesses of the original training dataset. There is also a risk of synthetic data being flawed and failing to capture real-world complexities, potentially leading to inaccurate outcomes.

To help data quality and completeness, the financial industry can establish centralised repositories of standardised, anonymised, and high-fidelity datasets that are purpose-built to test and fine tune financial applications as well as applicable benchmarks. Awareness of synthetic data and how it can be generated should be fostered, supported by clear documentations and assumptions used in its creations. Indeed, industry-wide collaboration is essential on synthetic data, and to create industry-specific training datasets to accelerate progress and reduce implementation barriers.

5.5 IP and copyrights

Without synthetic data, the risks of flawed and incomplete data can become more pronounced from the rise of intellectual property and copyright issues that would constrain the availability of public data for use in GenAI applications. The copyright paradox is both ironic and challenging. Intellectual property rights and its protection are vital to ensure creators of content are protected for their work and their generosity in sharing. But on the other hand, these same rights and protection can prevent access to the comprehensive-ness of data that GenAI needs to be effectively trained. Limited access to good quality data can ultimately lead to lower quality inaccurate outcomes.

For example, announcements by regulators about market changes, which serve the public good, can be subject to terms and conditions prohibiting commercial use. However, the definition of 'commercial use' for public information has become increasingly ambiguous. For instance, if a library of public market change news is integrated into a GenAI system for comprehensive textual analysis for clients, enabled by GenAI systems, is it considered commercial use even if no fees are charged for using the system?

This would bring us to fair data use that the next point touches on.

5.6 Open standards and fair data practice

To address the broader implications of lack of suitable training data, industry and regulatory bodies should consider policies that promote fair data practices – policies that facilitate data sharing while protecting IP and copyrights. This will have certain “Butterfly Effects” to benefit level playing fields for smaller AI firms to develop competitive AI solutions, mitigating concentration and dependency risks, as well as addressing data-related risks like hallucination.

Data is the glue and catalyst that allows any GenAI model to become a domain relevant application, and in doing so, ensures system resilience and mitigating dependency risks.

5.7 Cross-border scalability

The diversity, volume and scope of regulatory requirements within and across jurisdictions complicate GenAI implementations. Considerations include data protection laws, cross-border data flow restrictions, data sovereignty, transfer, data usage rights, guidelines on algorithmic accountability, and specific rules on model validation and auditability.

For instance, some jurisdictions may require that personal or any data be stored locally or impose restrictions on cloud-based solutions. As a result, firms looking to implement GenAI across regions need legal and compliance capabilities to handle these nuances, together with informed technologists and AI engineers, to discuss topics like hybrid cloud strategy or leveraging federated learning which allows training to occur locally without moving sensitive data across borders. That is to say, combining technology and compliance views as a solution to address regulatory concerns.

Scalability discussions also extend to maintaining consistent quality and adherence across diverse regulatory environments. Ensuring that a GenAI system can generate consistent, high-quality outputs while respecting local laws can require training data customisation and will need ongoing monitoring.

In their use cases, organisations would also need to consider adapting models to dialects, local legal contexts and regulatory nuances to ensure that the system's output can remain relevant and trusted. This leads us to the next point on non-English benchmarks that would play a central role here.

5.8 Non-English benchmarks

An earlier chapter has highlighted the importance of benchmarks as quality assessors and indicators. In markets where non-English languages dominate, deploying GenAI systems can be challenging if it is without sufficient linguistic datasets for training and benchmarks. Such a situation can lead to inaccurate translations, misunderstandings, and cultural insensitivity, and to heightened concerns about legal liabilities in commercial applications.

From a practical perspective, there's a trade-off to consider. Adopting a conservative approach by delaying GenAI deployment until the necessary benchmarks are available could be a solution. However, depending on how swiftly these benchmarks become accessible, firms risk falling behind and widening the technology gap. On the other hand, rushing to implement GenAI systems on a large scale might expose them to various risks, including reputational damage.

Therefore, non-English markets can benefit from its own industry developed, published local LLM benchmarks with an emphasis on translation and industry specialised nomenclature. This would allow simple but powerful functions like English queries directly into local language materials for effective communication by that market to the world at large.

5.9 Expertise availability, jobs and reliability

The successful integration of GenAI within regulated entities is not merely a technical challenge but also a human one. The complexity of these systems requires not only deep technical expertise but also a nuanced understanding of regulatory requirements, ethical considerations, and business domain dynamics. As such, GenAI talent pipeline is important, and one that an organisation can already build through reskilling, practical experiences and vocational training that is also age inclusive.

A strategic plan that includes upskill/reskilling can also assuage a level of fears of job loss due to AI-driven automation and internal focus on cost streamlining, which could lead to resistance from employees in various ways. A resistance can come in the form of not accepting anything less than 100% perfection from a GenAI system. Addressing this requires transparent communication around how GenAI will augment rather than replace human roles and thoughtful new procedures that allow AI systems with probabilistic results to fit.

Reskilling and vocational skills initiatives can focus on model training, writing new operational procedures, supervising, managing, and improving AI systems with reinforcement learning. This can foster a culture of innovation and growth, and to accept AI systems as a positive driver of change and career opportunities in an organisational fabric.

5.10 Quantifying ROI and productivity gains

For GenAI to be accepted for implementation within regulated entities, it has to demonstrate tangible business value. A challenge lies in being able to quantify the return on investment (ROI) from GenAI implementations that generates productivity gains as its main benefit. Traditional ROI metrics such as cost savings and efficiency gains are unlikely to fully capture the benefits of enhanced decision-making capabilities, better compliance, and faster informed processes. To objectively assess GenAI's value, firms can consider productivity metrics that include:

Reduction in compliance review times: How much time GenAI saves teams in reviewing and analysing large text datasets.

Accuracy in risk assessments: Comparing pre- and post-GenAI deployment risk management effectiveness.

Enhanced customer experience: Measuring the impact of GenAI on customer satisfaction and engagement scores.

Scalability and flexibility: Evaluating the ability to scale regulatory processes with minimal additional cost or resource strain.

Industry ROI and risk assessment frameworks that are accepted by participants and regulators can create a consistent minimum business case standard and enable smoother adoption of AI systems in organisations.

06

Conclusion Can GenAI thrive in the regulated financial industry?

Implementing GenAI in highly regulated sectors like financial services presents a unique set of opportunities and challenges. While the potential benefits—ranging from greater operational efficiency, augmented people capacity and enhanced customer service to risk management and compliance automation—are substantial, these advantages come with a backdrop of complex regulatory landscapes, ethical considerations, and technological expertise availability.

These and other challenges underscore that shared responsibility can become dispersed among various parties with unclear boundaries, leading to increased distrust when issues arise. Perception of novel risks to the industry is itself a risk that can result in excessive regulations on GenAI and AI systems, potentially stifling their capacity for positive impacts. However, the financial industry is a highly regulated one that has been using AI in different forms and there is a well of experience to pivot and address challenges in GenAI. For example, responsibility can be shared based on the level of control that stakeholders have in the development and deployment of a Gen AI system, referencing and leveraging established standards in Cloud environments.

To harness such experiences, accessible and regular public-industry engagements are important to raise awareness in forward looking specific topics and to agree on pragmatic approaches towards new but yet old topics; like determining level of control for responsibility assignments, data and privacy enhancing technologies, IP and copyrights for access to quality data, local benchmarks, public trusted training data sets, agreed good practices for explainability and transparency; and other topics.

Such discussions are also important to find balance, understand trade-offs and accept those that can be accepted at this time, establish clear governance frameworks, and foster good practices across different stakeholder segments that AI sandboxes can be useful.

The focus is not all about risks, but growth and the relevant risk management that should be applied for it to be sustainable. An environment for local GenAI ecosystem development – including education, infrastructure, research and homegrown AI – to ensure that GenAI's transformative potential can benefit the broader economy and the participants therein.

GenAI represents a significant transformative leap forward in human-computer interface to unlock new possibilities for organisation and people alike. It democratises user access to powerful tools that enhance creativity, efficiency, responsiveness and decision making. As we advance forward, pragmatic approaches to ensure that it drives business outcomes and inclusivity would benefit a diverse empowered workforce to create and thrive in an AI-enhanced future of the financial industry.

References:

Jan 2024, Enabling an Efficient Regulatory Environment for AI – Practical Considerations for Generative AI, Asia Securities Industry and Financial Market Association, 2024-asifma-gen-ai-paper-final-updated-18012024.pdf

March 2024, Grobelnik, Perset, Russell. What is AI? Can you make a clear distinction between AI and non-AI systems?, OECD.AI Policy Observatory, What is AI? Can you make a clear distinction between AI and non-AI systems? - OECD.AI

May 2024, Key Considerations for Artificial Intelligence in Capital Markets, Global Financial Markets Association, GFMA Letter on Key Considerations for AI in Capital Markets | GFMA | Global Financial Markets Association

May 2024, Model AI Governance Framework for Generative AI, AI Verify Foundation, Infocomm Media Development Authority (Singapore), Model-AI-Governance-Framework-for-Generative-AI-19-June-2024.pdf (aiverifyfoundation.sg)

June 2024, On AI Agentic Workflows and their potential for driving AI progress, Luminary Talk at Snowflake Summit 2024, Andrew Ng's Luminary Talk: A Look At AI Agentic Workflows (landing.ai)

September 2024, Nunez. OpenAI tackles global language divide with massive multilingual AI dataset release, VentureBeat, OpenAI tackles global language divide with massive multilingual AI dataset release | VentureBeat



This document is intended for informational purposes only and is designed to share general insights about the solutions and research of Kodex AI. It does not constitute professional advice, nor does it imply any commitment or guarantee regarding the future performance or outcomes of any solutions mentioned herein. Any figures, statements, or references made within this document are provided without representation or warranty, express or implied, and are subject to change without notice.

Kodex AI makes no guarantees as to the completeness or accuracy of the information presented, and it is not responsible for any errors or omissions, or for any losses or damages that may arise as a result of reliance on this material. The strategies, views, and opinions expressed may change as new information becomes available.

All content within this document is proprietary and the intellectual property of Kodex AI. Unauthorized use, distribution, or reproduction of this material without express written permission is prohibited.
Copyright © 2024 Kodex AI. All rights reserved.