The Association of Banks in Singapore

Handbook on Generative AI Guardrails in Banking

THE ASSOCIATION OF BANKS IN SINGAPORE (ABS)'S STANDING COMMITTEE ON DATA MANAGEMENT (SCDM)

May 2025

Foreword by ABS

Artificial Intelligence (AI) is transforming industries worldwide, and the financial sector is no exception. In recent years, Generative AI (Gen AI) has created new opportunities and challenges for banks and financial institutions, improving customer experiences, streamlining operations, and driving innovation. While some early adopters have gained a competitive edge with Gen AI, others are still figuring out how to use it responsibly, assess suitable use cases, and ensure compliance with industry standards like Project MindForge, based on the Veritas initiative and FEAT Principles.

In January 2024, a group of organisations including MAS, ABS, and major banks like ANZ, Citibank, DBS, GXS Bank, HSBC, JP Morgan, OCBC, SCB, SMBC, and UOB, along with Accenture, came together to create the Handbook on Generative AI Guardrails in Banking. Their insights helped identify the risks, challenges, and best practices for adopting Gen AI in banking. We thank all contributors for their valuable input.

This Handbook provides a clear framework for financial institutions to adopt Gen AI in a responsible, secure, and ethical way, focusing on strong governance, technical controls, and human oversight. This ensures that the industry can take advantage of Gen AI's potential while maintaining trust, compliance with data privacy laws, and regulatory requirements.

In addition to the Handbook, the financial sector has a unique chance to collaborate on shared Gen AI tools, best practices, and governance frameworks. ABS will continue to work with its committees and the broader financial community to create solutions that improve risk management and operational resilience. By working together, we can ensure that Gen AI not only fosters innovation but also preserves the integrity of Singapore's financial services industry.

Ong-Ang Ai Boon, Mrs
Director
The Association of Banks in Singapore (ABS)

Foreword by MAS

In the past 2 years, Generative AI (Gen AI) has probably been the most exciting technological innovation that has enthralled the world and this has fired the imagination of many industries, including the financial services industry. There is no doubt that the financial services industry is on the brink of transformation, propelled by Gen AI. This technology promises to revolutionise customer experiences, streamline operations, enhance risk management, and drive innovation across the sector. However, like most emerging technologies, the power of Gen AI comes with significant responsibilities. As we embrace this technology, we must balance innovation with our commitment to trust, compliance, and ethical standards. Potential risks such as hallucination, data privacy concerns, algorithmic bias, toxic outputs, and the challenge of explainable AI necessitate a careful approach to adoption.

The Association of Banks in Singapore's (ABS) Standing Committee on Data Management (SCDM) has developed, as part of the industry self-help effort, a handbook on Generative AI Guardrails in Banking and this is very timely considering the fact that most financial institutions are experimenting Gen AI with different use cases. This handbook provides a comprehensive framework providing clear guidance for financial institutions to harness Gen AI's value while mitigating risks. The handbook's use case-driven approach recognises that different Gen AI applications require varying levels of control, ensuring relevance and practicality.

These guardrails align with the Monetary Authority of Singapore's (MAS) regulatory objectives, including financial stability, institutional soundness, consumer protection, and financial centre development. They provide a foundation for developing trustworthy and compliant Gen AI applications and align with our MindForge initiative.

I encourage all stakeholders in the financial services industry to embrace this handbook as a roadmap for responsible Gen AI innovation. By adhering to these guidelines, the financial services sector can unlock Gen AI's full potential while maintaining the integrity and stability of our financial system. Let us shape the Gen AI-powered future responsibly, ensuring our innovations align with our core values and serve the best interests of our customers and the public. Together, we can build a financial ecosystem that is not only more efficient and innovative but also more inclusive, ethical, and resilient.

Vincent Loy
Assistant Managing Director (Technology)
Monetary Authority of Singapore (MAS)

Foreword

Artificial Intelligence (AI) has the potential to transform industries including the financial sector where trust is paramount. This initiative aims to establish the baseline guardrails to manage key risks associated with common Generative (Gen) AI use cases as part of our effort to promote its responsible use. We hope this will also better facilitate ongoing industry dialogue even as use cases and technologies for Gen AI evolve.

Special thanks are extended to the ABS Standing Committee on Data Management, Accenture, and MAS for their contributions. As the saying goes, fortuna audaces iuvat – fortune favours the bold, but it is also worth calling out that prudent risk-taking builds lasting success!

Andrew Tan Chee Peng
Head, Group Data Management Office, OCBC

In the rapidly evolving world of artificial intelligence, strong guardrails are crucial. This Handbook is a vital resource, guiding organizations in the responsible and principled use of generative AI. By outlining practical best practices and strategic frameworks, we empower stakeholders to harness AI's potential while mitigating risks, ensuring innovation aligns with responsible conduct.

Sameer Gupta
Group Chief Analytics Officer, DBS Bank

Acknowledgement

This Handbook has been developed with valuable contributions from the members of ABS-SCDM working group, representing the following institutions:

- Australia and New Zealand Banking Group Limited
- Citibank NA / Citibank Singapore Limited
- DBS Bank Limited
- GXS Bank Pte. Ltd
- HSBC
- JPMorgan Chase Bank N.A., Singapore
- Oversea-Chinese Banking Corporation Limited
- Standard Chartered Bank (Singapore) Limited
- Sumitomo Mitsui Banking Corporation Singapore Branch
- United Overseas Bank Limited

In addition, The Handbook was completed with the strong support of:

- Accenture

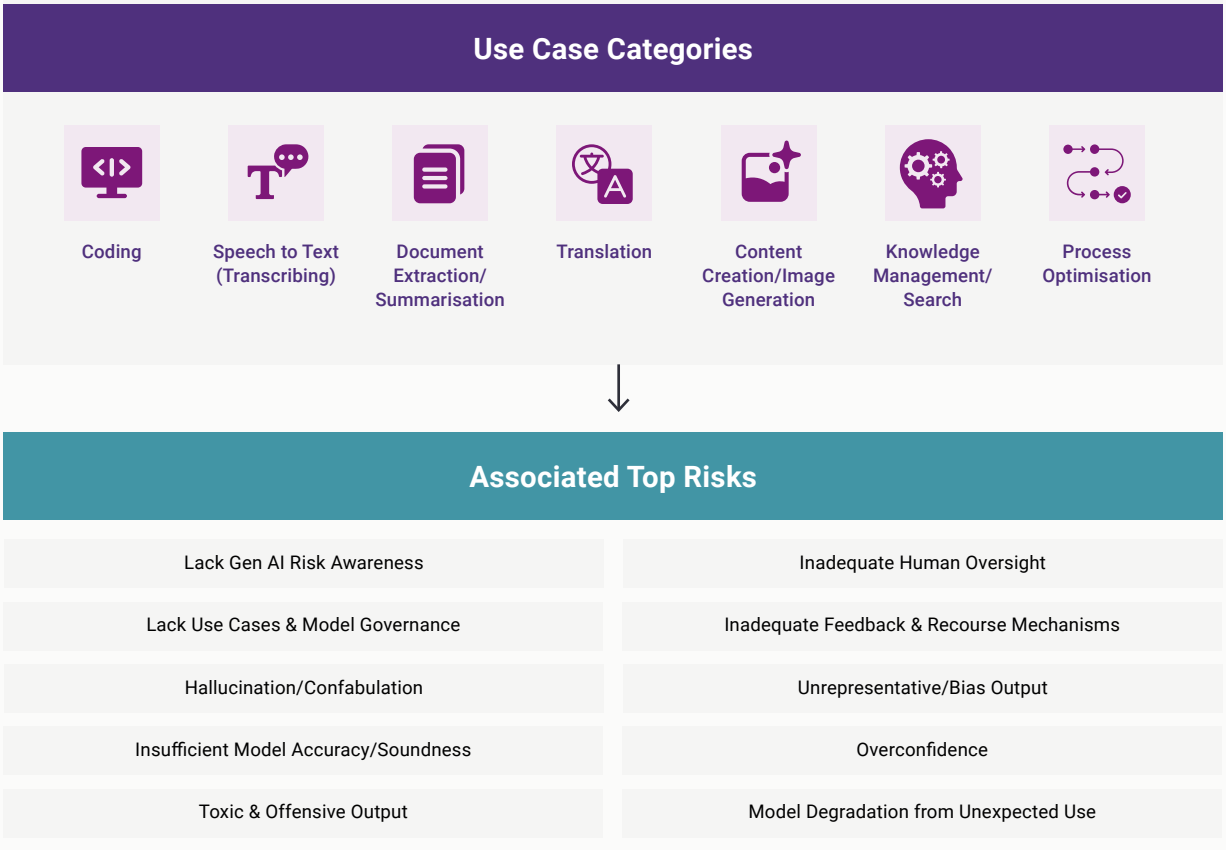
We would also like to extend our gratitude for the strong support received from Monetary Authority of Singapore (MAS) and The Association of Banks in Singapore (ABS).

Executive Summary

Generative AI (Gen AI) has empowered banks to enhance their creative capabilities, make their processes more efficient, and explore innovative solutions across their businesses. To manage the risks and potential challenges of Gen AI use, members of The Association of Banks in Singapore (ABS)'s Standing Committee on Data Management (SCDM) came together to collectively identify **guardrails**[1] to address the specific risks that may arise when rolling out Gen AI.

The approach was based on the experience of ABS SCDM members implementing seven common categories of enterprise Gen AI use across over thirty use cases in their organisations. This was followed by the identification of the most prevalent and applicable Gen AI risks, and corresponding guardrails to mitigate them.

Figure: Overview of the Use Case Categories and the Associated Top Risks



[1]Guardrails are pre-established guidelines and processes that act as a safety net to prevent and mitigate potential risks.

Table 1: Overview of the Guardrails

Guardrail	Description
Enterprise Governance and Training	Governance structures, training and education, and the implementation of risk management practices.
Filtering and Control	Limitations on the inputs or outputs of a Gen AI model, such that undesirable behaviours or unacceptable usage is not permitted.
Customised Model Design	Customisation of the underlying technical characteristics of Gen AI models to significantly alter their behaviour.
Red Teaming	Adversarial testing processes that aim to identify flaws, vulnerabilities, and other undesirable behaviour.
Prompt Design	Standard, bank-defined instructions included in user prompts to Gen AI that promote the use of good practices.
Monitoring and Validation	Activities that aim to determine if and when a Gen AI model is performing as desired.
Human-in-the-Loop Moderation	Human oversight of some or all of a model's outputs to identify undesirable behaviour.
User Feedback and Iterative Improvement	Activities for actively engaging and consulting end users so that Gen AI models can be continuously improved.
User Transparency and Consent	Disclosure of the use of Gen AI technology – and informed agreement to the terms of that use – as well as instructions on responsible user behaviour.

The Handbook describes how banks should select the appropriate guardrails for their situation based on characteristics and level of risk of their use of Gen AI. Alongside the Handbook, the Excel Tool describes these guardrails in detail, articulates where in the Gen AI system lifecycle they should be used, and illustrates a range of specific implementation controls that a bank can implement in practice.

Practical considerations around these guardrails are illustrated through two common use case studies on **Document Extraction / Summarisation** and **Code Generation**. These use case studies describe the common risks and key considerations for implementing the relevant guardrails in a real-world context.

This Handbook will serve as an initial guide based on the experience of ABS SCDM members to-date, and will continue to evolve with new developments in Gen AI and its associated risks. This document eventually will support the development of the AI Governance Handbook as part of Project MindForge, an industry AI initiative supported by the Monetary Authority of Singapore (MAS).

Contents

01 Introduction

1.1 Intent of the Paper

As organisations increasingly scale Gen AI technologies, they unlock unprecedented opportunities for innovation and efficiency, such as enhancing customer interactions, automating routine tasks, and gaining deeper insights from data. The potential benefits of Gen AI are vast, offering significant competitive advantages and driving remarkable progress. However, these advancements come with challenges, particularly in managing data security, privacy, and ethical concerns. For example, Gen AI systems are prone to hallucinations, where the model generates information that is plausible but incorrect. This poses a significant risk, especially when Gen AI is used in critical applications. To mitigate these risks, organisations must use high-quality domain-specific data, structured prompts, and human-in-the-loop reviews, while continuously monitoring performance and implementing robust governance practices.

Equally important is addressing the ethical considerations and biases associated with Gen AI. AI models can inadvertently perpetuate biases from historical data, leading to unfair or discriminatory outcomes, especially in a sensitive area like banking. To counteract these issues, organisations should adopt comprehensive bias detection and mitigation strategies, conduct regular audits of AI models, and establish ethical oversight committees. By integrating these practices, organisations can leverage Gen AI responsibly, ensuring that data is protected, privacy laws are upheld, and AI systems operate ethically and in alignment with organisational values.

This handbook is designed to provide practical guidance on establishing effective guardrails for the responsible use of Gen AI, with a particular focus on the banking sector. It outlines control implementations and strategic approaches to manage and mitigate the risks associated with Gen AI. By integrating these guardrails into their AI governance frameworks, organisations can leverage Gen AI technologies effectively while maintaining trust, ensuring compliance, and upholding ethical standards in their operations.

1.2 Traditional AI vs. Generative AI: A Paradigm Shift in Banking

AI has transformed industries like banking by analysing data to make predictions, automate tasks, and support decision making. Traditional AI relies on pre-defined algorithms to process structured data, optimise operations, and identify patterns, driving efficiency and accuracy in functions like fraud detection, risk assessment, and customer service automation.

Gen AI, on the other hand, represents a revolutionary leap forward by not just analysing data but creating entirely new content, such as text, images, audio, video, and complex data patterns, from the inputs it receives. This fundamental difference allows Gen AI to engage with users in a more creative and interactive manner, producing novel outputs that were previously impossible with traditional AI. Gen AI's capability to generate contextually relevant information is driving its rapid adoption across sectors, including the banking industry, where innovation and efficiency are paramount.

While traditional AI has been instrumental in automating routine tasks, streamlining processes, and enhancing predictive analytics, Gen AI is opening unprecedented opportunities for deeper transformation. For instance:

- **Process Automation and Efficiency:** Traditional AI has helped automate data-driven tasks such as transaction monitoring and risk scoring. Gen AI takes this further by automating or augmenting complex and creative processes like loan underwriting, document generation, and compliance reviews, allowing banks to operate more efficiently and reduce manual workloads.
- **Customer Engagement and Personalization:** Traditional AI-powered chatbots and virtual assistants are commonly used to only provide basic support by answering frequently asked questions. Gen AI enhances this by delivering highly personalised and human-like interactions and responding in a way that feels conversational and intuitive, significantly elevating the customer experience.
- **Advanced Fraud Detection and Risk Management:** Traditional AI has been essential for identifying patterns that signal fraud. Gen AI goes beyond simple pattern recognition, ingesting and parsing a wider set of contextual cues and parameters, including real-time and unstructured data, to identify new fraud patterns or previously unknown tactics, making it a powerful tool for advanced fraud prevention.

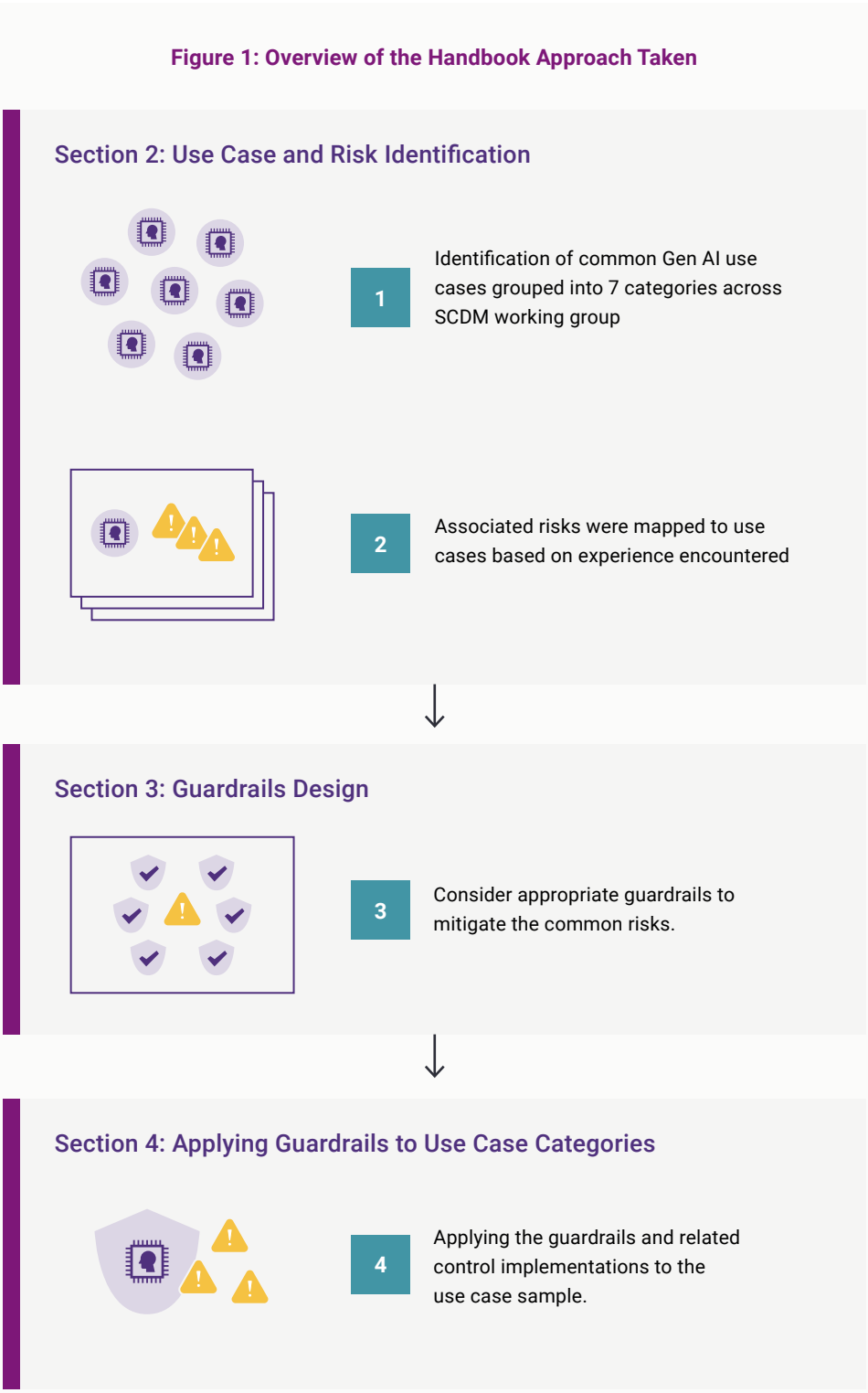
1.3 Handbook Approach

With the growing imperative to balance the immense potential of Gen AI against its associated risks, implementing relevant guardrails and controls throughout the AI lifecycle has become crucial. In response, The Association of Banks in Singapore (ABS), through its Standing Committee on Data Management (SCDM), has initiated industry efforts to identify applicable Gen AI guardrails and corresponding control implementations through the four-step approach shown in Figure 1. This Handbook outlines guardrails for a subset of the Gen AI risks included in the risk taxonomy published by project MindForge[2].

[2]Emerging Risks and Opportunities of Generative AI for Banks - A Singapore Perspective: https://www.mas.gov.sg/-/media/mas-media-library/schemes-and-initiatives/ftig/project-mindforge/emerging-risks-and-opportunities-of-generative-ai-for-banks.pdf

The selection of common Gen AI use cases and the identification of pertinent risks for the purposes of this paper is outlined in Section 2. Section 3 presents this paper's nine guardrail approaches and discusses how they can be implemented. Section 4 illustrates how these guardrail approaches can be applied to practical use cases.

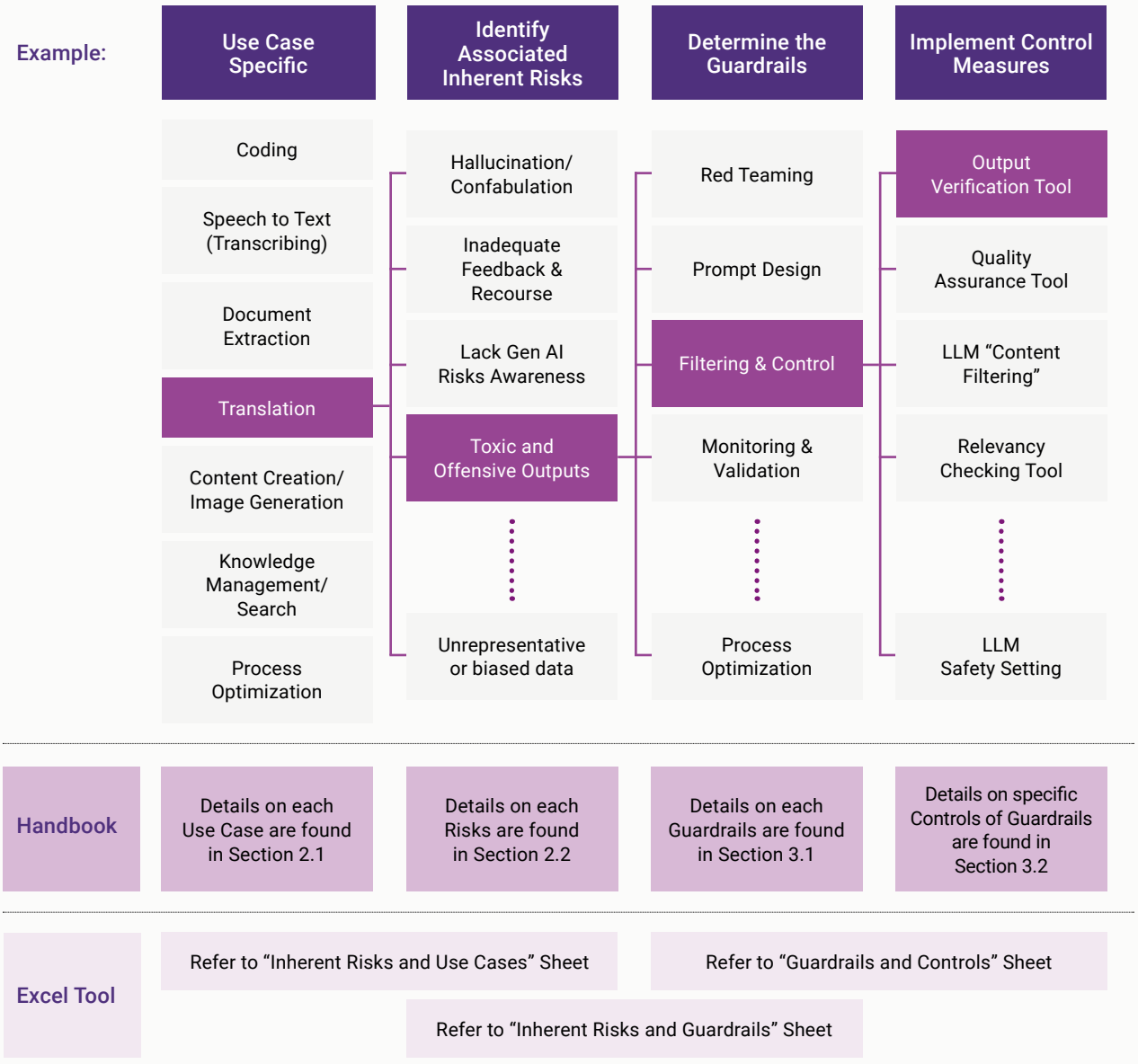
Figure 1: Overview of the Handbook Approach Taken



1.4 Navigating the Handbook and Tool

This section explains how the user can navigate the handbook and supplementary Excel Tool to identify risks associated with a Gen AI use case and determine the appropriate guardrails and control measures for implementation.

Figure 2: Approach to Navigating the Handbook and Tool



02 Use Case and Risk Identification

2.1 Identification and Classification of General Gen AI Use Case Categories

This initiative began with member institutions identifying common Gen AI use cases, such as chatbots, data analysis, knowledge search, and content creation, across their organisations. These use cases were then grouped into seven categories:

Coding
Using programming languages to communicate with computers and create software applications or interpret code.

Speech to Text (Transcribing)
Converting spoken words into written text, making it easier to transcribe call conversations, meetings, interviews, or lectures for minutes-taking or further analysis.

Document Extraction / Summarisation
Document extraction is the process of automatically extracting relevant information from documents, such as text, tables, or images, for analysis or storage. Summarisation is the automatic generation of concise and accurate summaries of various types of documents, such as policy documents or reports.

Translation
Converting written or spoken content from one language to another.

Content Creation / Image Generation
Generating ideas and transforming them into compelling content that engages and resonates with the target audience, which are essential processes in marketing, branding, and communication strategies.

Knowledge Management
Using chatbots to automate content creation, summarisation, and personalised delivery, improving information retrieval, organisation, and collaboration, alongside advanced search capabilities, to boost efficiency and informed decision-making within organisations.

Process Optimisation
Analysing workflows/processes, predicting bottlenecks, and suggesting improvements.

2.2 Associated Common Risks

Leveraging the MindForge risk taxonomy published by the Monetary Authority of Singapore (MAS), SCDM members identified the most prevalent Gen AI risks impacting the seven categories of use cases. Of the 14 risks identified, 4 fall outside the data management domain and pertain to areas like third-party risks, data loss protection, and cybersecurity. The working group recommended that these be addressed by relevant ABS forums, such as the Standing Committees on Cybersecurity and Third Party. The remaining top common risks will be covered by the ABS SCDM and are as follows.

Enterprise-wide risks:

- **Lack Gen AI Risk Awareness** – Insufficient education or reskilling resulting in undertrained resources lacking awareness of the unique risks involved with Gen AI.
- **Lack Use Cases & Model Governance** – Failure to implement, enforce and oversee principles, guidelines, protocols and controls to proactively manage risks and ensure traceability and responsibility throughout the lifecycle of the model.
- **Inadequate Human Oversight** – Sufficient human-in-the-loop or oversight is not available, limiting recourse to human correction or intervention in the event of a failure or when generating content whose risk level requires human validation.
- **Inadequate Feedback & Recourse Mechanisms** – No mechanism to provide feedback or seek recourse for those impacted by harmful or biased outputs, and with no consequence for the system's developers and/or owners for any negative outcomes.

Model specific risks:

- **Hallucination / Fabrication / Confabulation** – The models produce outputs that are not grounded on any source content or even convincingly contradict the source content, due to the model's lack of real-world understanding. This can have an adverse impact on social groups or may constitute grounds for libel.
- **Insufficient Model Accuracy / Soundness** – The model outputs are inaccurate and/or do not meet the performance thresholds required to ensure they are fit for purpose.

- **Toxic & Offensive Output** – Outputs contain harmful, offensive, hateful, discriminatory, violent, racist, sexist, or nudity related information.
- **Unrepresentative / Bias Output** – Data, whether intentionally or unintentionally, is biased against, or has uneven representation of, certain individuals or groups of individuals, which can produce biased model outputs.
- **Overconfidence** – The characteristic of Gen AI models to produce convincing outputs that do not properly account for the complexity, uncertainty, or contradictions in their sources – leading to the potential to present false information as factual or uncertain information as clear, and presenting this information in such a way that interferes with the ability of users to review using their judgement.
- **Model Degradation from Unexpected Use** – A wider range of unexpected usage patterns due to the broad capability of generative models creates outcome instability and/or unexpected failure modes.

In addition to the above risks, SCDM members emphasised other Gen AI risks, below, that are important to consider. As these risks are not under the domain expert knowledge of ABS SCDM, they are not included within the scope of this initiative.

- **Data Leakages** – Model outputs reveal sensitive, confidential, or personal data that should have been secured and only privately accessed.
- **Lack of Third-Party Accountability** – Organisation has limited control or oversight over the development, modification, and/or decision-making process for Gen AI models/services from third-party providers.
- **Unavailability of IP Protection** – The outputs of Gen AI built on foundation models are not afforded IP protections, such as copyright or trademarks, due to a lack of legal clarity over IP protection for AI-generated content.
- **Compliance over Location for Model Hosting & Data Processing** – Adherence with foundation model hosting and data processing location requirements.

Figure 3: Mapping of key Gen AI risks to use case categories

		Use Case Categories						
		Coding	Translation	Speech to Text (Transcribing)	Document Extraction/ Summarisation	Content Creation/Image Generation	Knowledge Management/ Search	Process Optimisation
Common Risks	Hallucination/ Fabrication/ Contabulation	✓	✓	✓	✓	✓	✓	✓
	Inadequate feedback and recourse mechanisms	✓	✓	✓	✓	✓	✓	✓
	Lack of Gen AI risks awareness	✓	✓	✓	✓	✓	✓	✓
	Lack of use case and model governance	✓	✓	✓	✓	✓	✓	✓
	Unrepresentative or biased data inputs	✓	✓	✓	✓	✓	✓	
	Inadequate human oversight	✓		✓	✓	✓	✓	
	Insufficient model accuracy / soundness	✓		✓	✓	✓	✓	
	Model degradation from unexpected use	✓		✓	✓	✓	✓	
	Toxic and offensive outputs		✓	✓	✓	✓	✓	
	Overconfidence	✓	✓		✓			✓
	Data leakages	**Out of scope for this initiative**						
	Inability to ensure location compliance for model hosting and data processing							
	Lack of 3rd party accountability							
	Unavailability of IP Protection							

This framework is based on the common risks identified by SCDM members at the time of developing this paper. It is not intended to be exhaustive, prescriptive, or used as the sole reference for evaluating Gen AI risks. Other risks may also apply, including those not mapped in the above table against a given use case category. Organisations need to conduct their own assessment of their use cases to identify relevant risks.

2.3 Selected Use Case Categories

Based on the selected common risks, nine guardrails approaches, and their corresponding control implementation steps, were designed across the Gen AI development lifecycle and are covered in detail in section 3 of this paper. The focus of these guardrails is primarily on managing the usage and outputs of Gen AI, while aspects such as input data used to train models or model explainability are considered beyond the current scope of control for banks. This approach aims to provide banks with a structured framework to address each of the selected Gen AI risks, ensuring that the deployment of these advanced technologies remains responsible and aligned with industry standards.

The practical application of these guardrails is illustrated using two use case categories: document extraction / summarisation and coding.

Document Extraction / Summarisation:

One of the most common applications of Gen AI in the banking sector is document extraction and summarisation, which uses a Gen AI system to ingest various input files, as well as a text prompt providing instructions, and outputs useful information – usually key points or data from the source files or a short summary of their contents. Inputs and outputs can, in theory, be of any modality or can be multimodal, but most systems performing document extraction and summarisation in banks today are exclusively text-to-text.

Document extraction and summarisation can be applied in areas such as enterprise knowledge management, compliance and risk management, agent assist, relationship management copilots, and office assistants.

Coding:

Code generation has become one of the most promising value-adding use cases of Gen AI in the banking industry because of its ability to meaningfully augment human capabilities in one of the most difficult and costly elements of a bank's business: software development. By using a Gen AI model to create customised computer code in the context of an enterprise's software ecosystem – in response just to a user prompt – code generation tools can augment developer capabilities and significantly accelerate coding tasks.

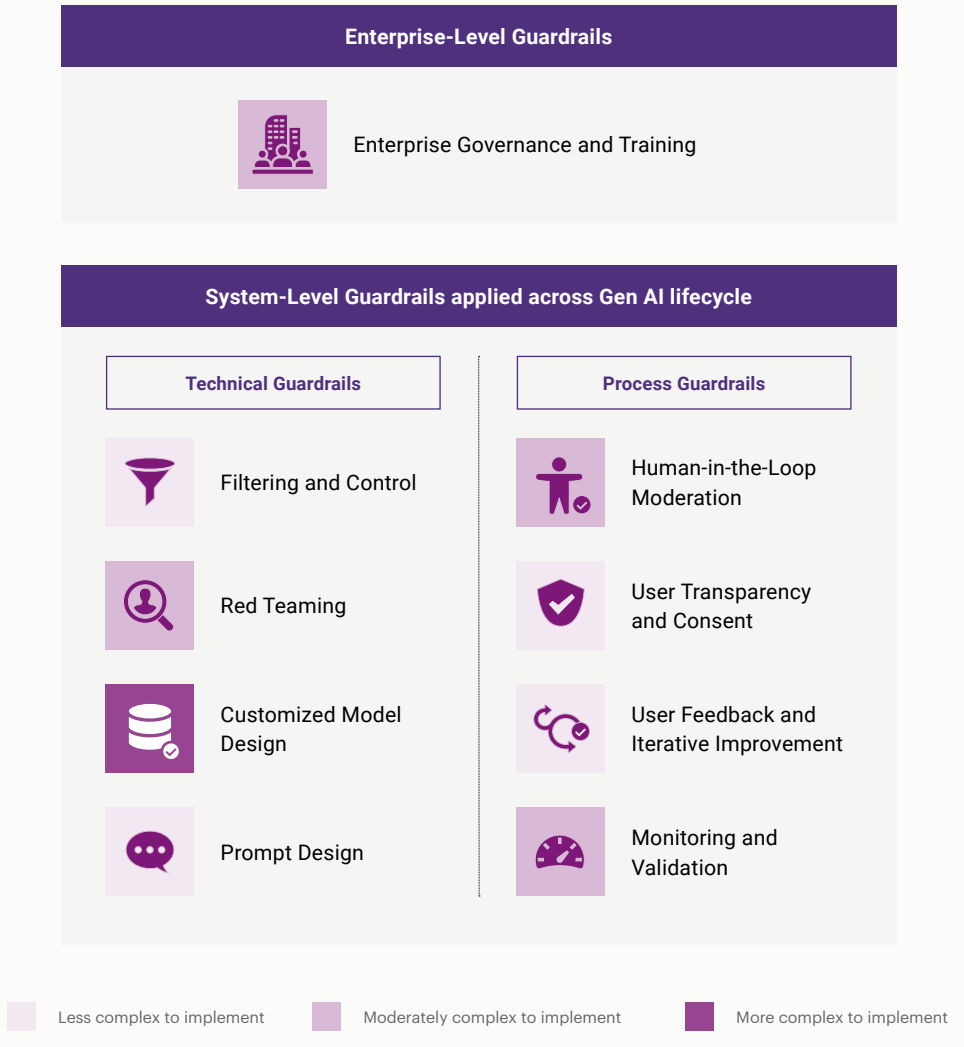
In a bank, code generation has several practical uses such as software development assistance, automatic testing and debugging, accelerated or automated translation between coding languages, documentation generation based on code, and prototype or proof-of-concept development based on natural language prompts.

03 Guardrails Design

3.1 Proposed Framework for Designing Guardrails and Relevant Controls at Enterprise- and System-Level

Guardrails for Gen AI are measures that are implemented across Gen AI use cases to address specific risks. Guardrails are designed at two levels: Enterprise-level guardrails and System-level guardrails.

Figure 4: Select Gen AI Guardrails



Enterprise-Level Guardrails: These guardrails focus on overarching governance measures that apply across the entire organisation and for all Gen AI use cases. They include establishing institutions, committees, high-level policies, procedures, standards, trainings, Standard Operating Procedures (SOPs) and risk management frameworks.

System-Level Guardrails: System-level guardrails are more granular and are applied across the Gen AI development lifecycle. These guardrails are divided into two subcategories:

- **Technical Guardrails** – These guardrails involve technology-based activities that are implemented at the system level. Some of the technical guardrails include filtering and control, customised model design, red teaming, and prompt design
- **Process Guardrails** – These are procedural elements of Gen AI deployment such as human-in-the-loop moderation, user transparency and consent, user feedback and iterative improvement, and monitoring and validation.

Table 1: Description of the Common Guardrail Approaches

Guardrail	Description
Enterprise Governance and Training	Governance structures, training and education, and the implementation of risk management practices.
Filtering and Control	Limitations on the inputs or outputs of a Gen AI model, such that undesirable behaviours or unacceptable usage is not permitted.
Customised Model Design	Customisation of the underlying technical characteristics of Gen AI models to significantly alter their behaviour.
Red Teaming	Adversarial testing processes that aim to identify flaws, vulnerabilities, and other undesirable behaviour.
Prompt Design	Standard, bank-defined instructions included in user prompts to Gen AI that promote the use of good practices.
Monitoring and Validation	Activities that aim to determine if and when a Gen AI model is performing as desired.
Human-in-the-Loop Moderation	Human oversight of some or all of a model's outputs to identify undesirable behaviour.
User Feedback and Iterative Improvement	Activities for actively engaging and consulting end users so that Gen AI models can be continuously improved.
User Transparency and Consent	Disclosure of the use of Gen AI technology – and informed agreement to the terms of that use – as well as instructions on responsible user behaviour.

Figure 5: Application of Common Guardrail Approaches to Common Gen AI Risks

Common Risks	Guardrail Approaches: Enterprise Governance and Training	Filtering and Control	Customized Model Design	Red Teaming	Prompt Design	Monitoring and Validation	Human-in-the-Loop Moderation	User Feedback and Iterative Improvement	User Transparency and Consent
Hallucination/ Fabrication/ Contabulation	✓	✓	✓		✓	✓	✓	✓	✓
Insufficient model accuracy/ soundness	✓	✓	✓			✓	✓	✓	✓
Model degradation from unexpected use	✓		✓	✓		✓	✓	✓	
Toxic and offensive outputs	✓	✓	✓	✓	✓	✓	✓	✓	
Overconfidence	✓	✓	✓		✓	✓	✓	✓	✓
Lack of Gen AI risks awareness	✓								✓
Lack of use case and model governance	✓								
Inadequate feedback and recourse mechanisms	✓							✓	✓
Unrepresentative or biased data inputs	✓	✓	✓		✓		✓	✓	✓
Inadequate human oversight	✓						✓	✓	

The above table shows the mapping of common Gen AI risks to their corresponding guardrail approaches. These are the possible guardrails identified by the SCDM members at the time of developing this paper. Not all Gen AI use cases will necessarily require the implementation of all relevant guardrails. The choice of which guardrails to use requires a judgement based on the degree of risk of that particular use case, the extent to which it involves an AI system interacting with clients or external stakeholders, and the technology complexity and predictability of the underlying AI tool. Use cases that are simple, internal, and limited in risk can often be governed with a light touch, using a subset of relevant guardrails and focusing on those that are less costly to implement. Conversely, a use case that is client-facing, complex, or which poses significant business risks may need the full range of available guardrails – up to and including the most complex and expensive ones, like model customisation and reinforcement learning. This determination depends on the enterprise risk appetite and the use case's unique context.

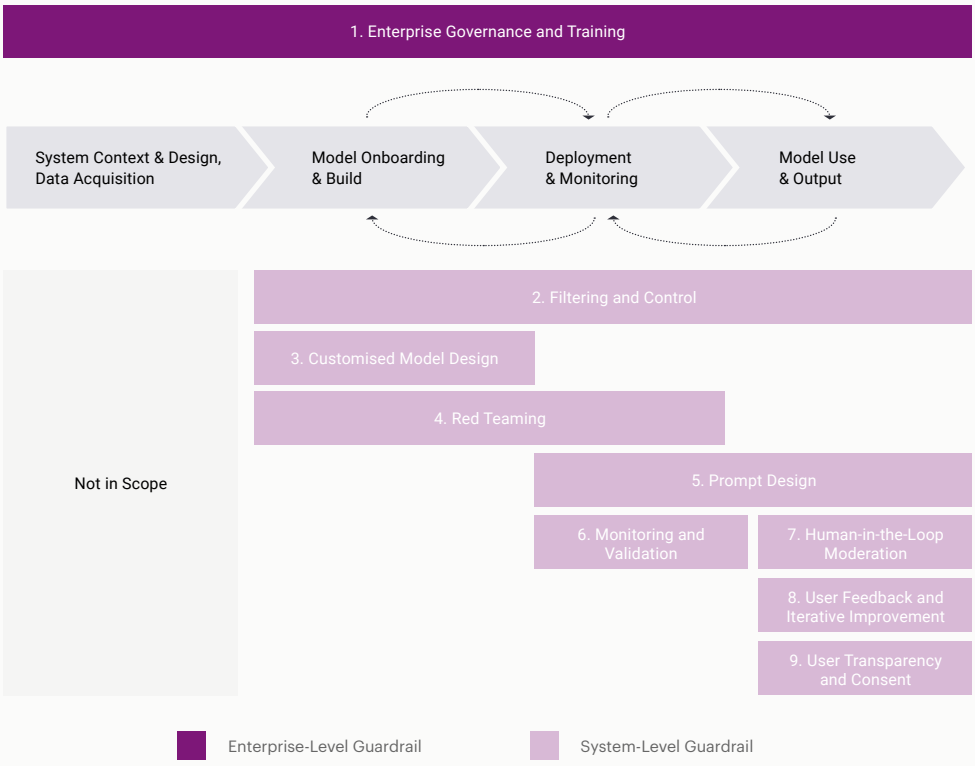
The first step for an enterprise to take in completing the governance of a Gen AI system is to complete a risk tiering and assessment, a key element of Enterprise Governance and Training. The extent to which those risks are mitigated will depend on several factors, chiefly including the maturity and effectiveness of the guardrail implementation, the business context, and the details of the use case. This risk tiering and assessment facilitates a value assessment of the proposed Gen AI use case. In general, riskier use cases for a given capability will require more expensive guardrails (such as model customisation and reinforcement learning), requiring that their value be sufficient to offset these lifecycle costs. Gen AI use cases that are not economically viable to implement responsibly should not be implemented at all.

The risks and guardrails proposed in this paper are part of a rapidly evolving landscape. While the paper provides a baseline framework, organisations will need to monitor changes and update their Gen AI risk and control strategies accordingly for each of their Gen AI use cases.

3.2 Guardrail Approaches and Control Implementations

Guardrails for managing Gen AI risks are applied at various stages of the Gen AI development lifecycle. Figure 5 illustrates the application of each of the nine guardrail approaches throughout the system lifecycle.

Figure 6: Potential Application of Proposed Guardrail Approaches Across the Gen AI System Lifecycle



The Gen AI development lifecycle consists of the following stages: System Context & Design, Data Acquisition, Model Onboarding & Build, Deployment & Monitoring, and Model Use & Output. This paper will primarily focus on the last three stages – Model Onboarding & Build, Deployment & Monitoring, and Model Use & Output – since banks, as Gen AI deployers, have control over the activities implemented in these stages. The first two stages – System Context & Design and Data Acquisition – are mainly managed by the developers of foundational models, with banks having limited influence.

ENTERPRISE-LEVEL GUARDRAILS

3.2.1 Enterprise Governance and Training

Description: Enterprise governance and training represents three distinct activities: the establishment of governance structures and guidelines, the completion of training and awareness activities, and the establishment of robust risk management. Together, these represent a unique *enterprise-wide guardrail* that applies transversally across all Gen AI use cases. These are foundational mitigation activities for every inherent risk of Gen AI that underpin the application of all other guardrails.

This guardrail includes three sub-approaches:

Establishing governance structures and guidelines: Creating the institutions, processes, policies, and organisational responsibilities that can facilitate the universal and effective application of AI governance.

Completing training and awareness activities: Ensuring that employees throughout the organisation are equipped with the knowledge and skills to apply AI guardrails and controls in their work, as well as to take steps to create a responsible AI culture throughout the organisation.

Establishing robust risk management: Systematically implementing risk management procedures around the use of Gen AI – and wherever feasible, integrating these risk management procedures with existing enterprise functions like Model Risk Management (MRM).

Indicative complexity to implement: Moderate

Control implementation: Enterprise governance and training should be implemented through the following controls:

Establishing governance structures and guidelines

Control 1.1: Define and promulgate a consistent set of organisational AI Principles applicable to all uses of AI and, above all, congruent with the enterprise's mission, values, and business context. Organisational principles and policies governing the use of AI may include characteristics of trustworthy AI, such as being valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhancing, and fair, with harmful biases identified and mitigated.

Control 1.2: Define an enterprise AI governance operating model that specifies how the organisation's AI activities should be organised and what roles and responsibilities are required. This operating model should specify what the organisation's AI oversight measures are and how they are to be applied.

Control 1.3: Define appropriate organisational policies on AI, taking the form of a dedicated AI policy and/or updated policies on related subjects that include, but are not limited to, data (including preparation, processing, and storage), risk management, information security, and procurement. Implementing these policies requires teams to be educated on how AI is involved in the policies and on each team's role in implementing them.

Control 1.4: Establish a consistent enterprise-wide AI inventory that tracks each use case and its key characteristics, including data sources and intended uses. Ensure that this inventory is integrated into ways of working. De-conflict all AI inventory activities with existing MRM controls, integrating the two where feasible. Implement audit mechanisms that utilise third-party or external tools to verify and track the compliance of each Gen AI use case within the inventory.

Control 1.5: Develop standard operating procedures (SOPs) that equip employees with the skills for running and governing Gen AI, including around understanding Gen AI limitations and understanding and responding to Gen AI risks, disclaimers, and failures.

Completing training and awareness activities

Control 1.6: Implement role-specific Gen AI training, addressing in particular the roles of leadership, developers, system owners, and employees at large. Ensure that training is an ongoing and evolving practice.

Establishing robust risk management

Control 1.7: Understand key categories of AI risks and articulate risk levels that correspond to the organisation's needs. Integrate this risk categorisation into the organisation's AI governance model, attaching controls to specific levels of risk where relevant. De-conflict all AI risk assessment activities with existing MRM controls, integrating the two where feasible.

Control 1.8: Assess the enterprise's existing MRM function for potential overlaps with AI governance activities. Consider the ways in which existing MRM frameworks can be extended to AI, empowering MRM teams with the tools and expertise wherever practical to complement or conduct AI governance activities, especially AI system inventory and risk assessment.

SYSTEM-LEVEL – TECHNICAL GUARDRAILS

3.2.2 Filtering and Control

Description: Filtering and control represents a family of Large Language Models (LLM) techniques that broadly seek to achieve the same objective: imposing limits on the kinds of behaviour that the system is capable of exhibiting, either through input filtering, output filtering, or the introduction of other system-level controls. These approaches shape the outputs of LLM, including masking Personally Identifiable Information (PII) for enhanced data security. These methods help mitigate several risks. Hallucination or fabrication is mitigated by directing models toward more reliable outputs. Overconfidence is mitigated by detecting and modifying overly certain responses. Biased inputs are mitigated by filtering them out or catching biased outputs before returning them to users. Toxic or offensive content, meanwhile, is blocked at the output level to ensure safer interactions.

These filters and controls may be:

- Rules-based, which search for specific phrases or content and respond to it in a deterministic way.
- Model-based, which detect targeted content in user prompts or in LLM outputs through a scoring approach or using forms of natural language processing.
- Agentic, in which case an AI model – often a second LLM – checks inputs and outputs according to set criteria and generates an appropriate response.
    - Agentic models should have sufficient autonomy, proactivity, and carefully specified goal orientation to effectively take actions within a system architecture.

The controls such as predefined structured instructions and hyperparameters are relevant for "build" or "boost" deployment models, but are less applicable to "buy" deployment models, where hyperparameters may be unavailable and as such different controls may be used.

Indicative complexity to implement: Lower

Control implementation: Filtering and control should be implemented through the following controls:

Control 2.1: Predefine structured instructions to guide the LLM towards desired behaviours, such as passing relevant documents or context to the LLM, citing sources, excluding undesirable information, expressing uncertainty appropriately, or declining harmful prompts. Evaluate meta-prompt language to ensure it addresses desired responses, testing alternative phrasings.

Control 2.2: Adjust LLM hyperparameters and safety settings to achieve an appropriate level of reliability for the desired application. Test the LLM under various hyperparameter conditions and assess the extent to which these impact performance on core tasks and the incidence of errors.

Control 2.3: Deploy an output verification tool – whose role can be filled by the principal LLM in question or a secondary model, depending on the cost and observed performance of those models – for applications needing enhanced reliability. Configure the system so that this verification tool screens LLM outputs before reaching the user, using tailored instructions for each undesirable behaviour. Based on verification results, trigger actions such as rejecting the prompt, regenerating the response, or adding qualifiers to the output.

Control 2.4: Deploy a relevancy checking tool for applications requiring additional accuracy. Configure the system architecture such that LLM outputs are tested for their similarity to source data, and segments or whole responses that contain material irrelevant to the source are flagged. Trigger appropriate actions depending on the outputs of the relevancy checking tool, including refusing the user prompt, re-generating the response, or adding qualifiers to the system output.

Control 2.5: Deploy an input quality assurance tool to screen user prompts before they are processed by the primary LLM. Based on the input quality assurance tool's findings, take actions such as rejecting the prompt, censoring or replacing undesirable terms, or using another LLM to rephrase the prompt.

Control 2.6: Explore the configuration of LLM "Content Filtering" to achieve appropriate content output levels. Test various configurations to evaluate their performance and ensure it meets the desired standards.

3.2.3 Customised Model Design

Description: Customised model design refers to a broad family of techniques that, together, all involve modifying the structure or functioning of an LLM itself. Customised model design can take the form of adjustments to a model's weights, which is possible for open-source models, internally developed models, and certain proprietary models with APIs for weight adjustment. This is a more transformative modification of the model's internal characteristics and can significantly change its behaviour. Because of its cost and complexity, this guardrail is best suited to the most risky or complex use cases.

A key technique for adjusting model weights is fine-tuning, which is an approach for modifying a model's internal parameters using a limited dataset. Fine-tuning builds on an existing model, and as such has lower computational and data requirements than training one from scratch. The effectiveness of fine-tuning is strongly linked to the quantity and quality of the data that is applied to these additional training rounds. Common fine-tuning datasets include:

- Fact-checking datasets, which consist of pairs of prompts and correct responses contextualised to the LLM's usage context and are often written by a human.
- Negative examples, which consist of examples of outputs that the LLM should avoid.
- Domain data, which consists of general subject-relevant data that may be useful for the LLM to respond more accurately and pertinently to prompts in a particular subject or context.

It can also take the form of architectural interventions or design decisions. This can include selecting a Small Language Model (SLM) instead of an LLM for a certain task that requires high levels of accuracy and predictability, and where complex, creative responses are not required.

This guardrail is suited for high-risk or complex use cases, addressing risks such as hallucination and fabrication by retraining the model with domain-relevant data, insufficient accuracy by fine-tuning with relevant data, and model degradation by periodic updates in changing contexts. It can also reduce overconfidence through calibration methods that synthesise more robust responses. Additionally, biased inputs are mitigated by training the model with unbiased data, while toxic or offensive outputs are minimised by fine-tuning with safe, appropriate examples for the use case.

Indicative complexity to implement: Higher

Control implementation: Customised model design should be implemented through the following controls:

Control 3.1: Conduct fine-tuning on the LLM using a limited dataset. Goals for fine-tuning may include reducing the incidence of bias, confabulation, inaccuracy, toxicity, or overconfidence. Conduct fine-tuning and benchmark re-trained model versions against older versions before accepting the fine-tuned version.

Control 3.2: Implement Retrieval-Augmented Generation (RAG) techniques to enhance the accuracy and reliability of Gen AI models with relevant information retrieved from trusted sources. This grounds outputs in factual data, reducing the risk of hallucinations.

Control 3.3: Based on use case objectives and context, during model selection, evaluate whether an LLM or a more constrained model like an SLM may be suitable. Assess the performance of multiple potential models on a use case, considering their cost, performance on core objectives, and performance on metrics of undesirable behaviour, such as irrelevant outputs, toxicity, and bias.

Control 3.4: Periodically collect refreshed domain-specific data to fine-tune the LLM over the course of its lifespan to ensure that its outputs remain well-contextualised and pertinent. Conduct fine-tuning and benchmark re-trained model versions against older versions before accepting the fine-tuned version.

Control 3.5: Introduce model calibration techniques to align the predicted probabilities of a model with the probabilities of those events to improve the model's performance on tasks where overconfidence or inaccuracy are potential risks. Consider benchmarking results against the original system before accepting a calibrated approach.

Control 3.6: Introduce additional agents, tools, or models into the LLM's system architecture. Introduce a tool to divide prompts into actionable steps or components, passing each to the LLM or to another tool that may handle them better. Test different architectural configurations for performance and the production of undesirable responses.

Control 3.7: Conduct reinforcement learning on domain-specific knowledge and subject matter. Determine which types of outputs are most relevant to assure using reinforcement learning, such as accuracy or bias. Determine the most appropriate technique and monitor progress for results and the introduction of bias.

Control 3.8: Re-conduct reinforcement learning periodically, in an ongoing fashion, to address potential concerns around data or model drift. Track the rewards or penalties assigned during reinforcement learning to benchmark model performance over time.

3.2.4 Red Teaming

Description: Red teaming is an interactive and structured testing approach designed to identify flaws and vulnerabilities in AI systems, such as harmful behaviour, leaks of sensitive data, and the generation of toxic, biased, or factually inaccurate content. It involves deliberately provoking a Gen AI system to produce outputs it was specifically trained not to, or to reveal biases and vulnerabilities that were previously unknown to its developers. This process involves extensive robustness testing during the development phase by simulating unexpected scenarios and edge cases, including adversarial testing.

The adversarial prompts are usually crafted according to prompt engineering techniques including: (1) Prompt Injection, (2) Prompt Probing, (3) Gray Box Attacks, (4) Jailbreaking, (5) Text Completion Exploitation, and (6) Biased Prompt Attacks. This is often completed by teams other than the one that developed the model, providing an outside perspective on the system's functioning. Issues identified through red teaming can later be mitigated.

Indicative complexity to implement: Moderate

Control implementation: Red teaming exercises should be carried out during the development process, before deployment in the pre-production environment, and periodically after deployment in sandbox environment.. It helps to address risks such as model degradation from unexpected use and toxic and offensive outputs. It is implemented through the following controls:

Control 4.1: Establish and monitor clear degradation objectives by identifying the most concerning potential harms that an AI system might cause. Key degradation objectives could include topics such as: (i) Copyright and Intellectual Property violations, (ii) Privacy breaches involving PII, (iii) Bias and fairness issues, (iv) Generation of hateful content, and (v) Misleading or hallucinated information, among others.

Control 4.2: Construct a sufficiently large red teaming dataset of adversarial prompts to simulate a range of possible adversarial attacks. This is best accomplished by first constructing an initial set of adversarial red teaming prompts, before progressively evolving it to increase the scope and complexity of attacks. A robust red teaming dataset should unveil a wide variety of harmful or unintended responses from an LLM.

Control 4.3: Systematically apply unexpected or out-of-distribution data to the model during development, validation, and post-deployment testing stages to understand model behaviour under these conditions. This approach helps identify potential weaknesses and vulnerabilities before deployment, ensuring that the model can handle real-world scenarios safely and effectively.

Control 4.4: Consider quantifying the model's ability to maintain performance when exposed to perturbations or unexpected inputs using metrics such as a robustness score.

Control 4.5: Automate adversarial prompt generation through an LLM or using third-party tools which can create customised adversarial prompts at scale.

3.2.5 Prompt Design

Description: Prompt design is a crucial aspect of optimising Gen AI applications to ensure they produce accurate and reliable outputs. This approach overlaps with the "Filtering and Control" approach which focuses on a zero-shot approach (where the model performs tasks without being provided with specific examples). In contrast, prompt design typically utilises a few-shot approach (where the model is given a few examples to guide its response). *Chain of Thought* prompting is one such technique that involves breaking down complex problems into smaller, sequential steps. This method improves reasoning by allowing the model to address each step of a problem individually, resulting in more accurate and coherent responses.

Another technique, *Constraint-Based Design,* involves incorporating task-specific constraints into prompts. These constraints help the model focus on relevant information, support logical reasoning, and minimise inaccuracies by reducing the risk of hallucinations. By setting clear boundaries and guidelines, this approach ensures that the Gen AI model adheres to defined criteria, improving both the relevance and reliability of its outputs.

Prompt design is essential for mitigating Gen AI risks such as hallucinations, biased inputs, and toxic outputs. Hallucinations can be reduced by breaking queries into smaller parts and combining results. Biased inputs can be addressed by incorporating guidelines in the prompts to detect and prevent biased intent. Toxic or offensive outputs can be mitigated by including sample responses that guide appropriate behaviour in response to unsafe prompts.

Indicative complexity to implement: Lower

Control implementation:

Control 5.1: Implement a prompt design strategy that utilises Chain of Thought techniques to decompose complex queries into smaller, more manageable queries. Provide relevant user training to design prompts using this strategy.

Control 5.2: Incorporate specific constraints within prompts to guide the model's responses. These constraints should be tailored to the domain and specific task, enhancing the model's focus and minimising the risk of hallucination.

Control 5.3: Include samples of safe responses in model prompts, mapped to potentially harmful inputs that the model may encounter in the course of operation.

Control 5.4: Implement structured and organised prompts through a templatised approach (e.g., task, instructions, guidance, examples) to ensure clarity and consistency.

Control 5.5: Create and maintain a centralised prompt library to minimise errors, maintain knowledge, and promote learning between users. Regularly iterate and refine prompts, incorporating reference inputs and outputs to enhance accuracy and safety.

SYSTEM-LEVEL – PROCESS GUARDRAILS

3.2.6 Monitoring and Validation

Description: Monitoring and validation are two closely related activities that aim to determine if a model is performing as desired at the time of procurement, during deployment, and over the course of the system's lifecycle.

Monitoring refers to the measurement and collection of information related to a model's performance in an ongoing fashion. Monitoring typically is either conducted continuously, with the model's outputs being captured in real time, or periodically, with the model being tested occasionally and its results compared to a benchmark.

Monitoring approaches should include contingency plans that are triggered by deviations from key metrics. These contingency plans can range from switching to manual processes, switching to backup models, to switching to older model versions, and typically include a plan for deploying guardrails to mitigate the issues detected.

Validation refers broadly to techniques that can determine whether a model meets a threshold of performance, such as on accuracy, bias, or stability. Validation is particularly useful when selecting between different LLMs, such as those offered by several vendors, as it can quantify the performance of each candidate LLM on metrics relevant to the use case.

Validation approaches include:

- Search- or knowledge-based validation, which measure the extent to which the model's outputs match a knowledge base; this typically is used to measure the accuracy of the model's outputs. This knowledge base can consist of internal domain data or of public data, such as search results on the open Internet.
- Statistical or linguistic validation, which calculates a standard industry benchmark on the outputs generated by the model.
- Validation datasets, such as question and answer pairs, against which model responses can be graded for correctness. Some validation datasets, such as those offered by Project Moonshot, are industry-standard and widely available.

Monitoring and validation ensure effective governance of AI systems in production, addressing various risks. Hallucination or fabrication is mitigated by accuracy metrics and comparisons to ground truth, while insufficient accuracy is managed through domain knowledge checks and consistency metrics. Model degradation is handled by tracking performance changes over time, and toxic/offensive outputs are reduced by measuring the model against bias and toxicity standards.

Indicative complexity to implement: Moderate

Control implementation: Monitoring and validation should be implemented through the following controls:

Control 6.1: Designate an appropriate source of ground truth that is relevant to the model's usage context and progressively expanding the golden dataset and ground truth to ensure that monitoring is robust to variety of scenarios. This is achieved through search or reference to an enterprise knowledge base and the use of quantifiable metrics for the measurement of the model's adherence to that source of ground truth. This score can be used to validate the model. To the extent that it is feasible, integrate the measurement of this metric into the model's ongoing operation, either in real time or periodically.

Control 6.2: Identify appropriate metrics for accuracy, bias, toxicity, and other relevant undesirable behaviours. Use evaluation or acceptance testing to validate the model, either for the purposes of procurement, production testing, or acceptance testing. To the extent that it is feasible, integrate the measurement of this metric into the model's ongoing operation, either in real time or periodically.

Control 6.3: Revise organisational procedures to include model validation in standard practices for the procurement and acceptance testing of Gen AI systems. Set common enterprise standards for tests that need to be passed. Set out common enterprise standards for ongoing model monitoring.

3.2.7 Human-in-the-Loop Moderation

Description: Human-in-the-Loop (HITL) moderation is a critical approach for managing the safety, accuracy, and appropriateness of outputs generated by Gen AI systems, especially in sensitive or high-risk applications. This process involves incorporating various levels of human oversight, including expert evaluations or dedicated roles for moderating and approving Gen AI output. Depending on the use case, human reviewers, including Subject Matter Experts (SMEs), may assess generated content to identify inaccuracy, harmful content, and specific cultural or organisational sensitivities that the model may struggle to fully capture.

HITL moderation involves systematic checks like output sampling, targeted human fact-checking in response to system-generated requests, and co-piloting models with human reviewers to provide continuous oversight. In some high-risk applications, HITL might involve a human checking every output, either in real time or after the fact, to ensure that responsibility for the final output remains with a human.

To effectively implement HITL, a risk-based approach should be adopted to determine the level of human supervision required for each model. This classification can be based on factors such as materiality, impact, subject matter, data sources, and complexity. While minor tasks (e.g., directing users to forms) may only require periodic human oversight, more critical interactions (e.g., reimbursement approvals) may necessitate comprehensive HITL monitoring. Identifying the necessary process and layers of human oversight – whether on a sample basis or for every system output – is crucial. Additionally, ensuring that human experts have adequate training and tools is essential for effective and efficient review.

Indicative complexity to implement: Moderate

Control implementation: HITL moderation is an important guardrail to address risks and can be operationalised using these controls:

Control 7.1: Designate an employee to review Gen AI-generated outputs, using random sampling or full scans, depending on the risk level of the use case to ensure they meet quality and safety standards before deployment.

Control 7.2: Identify high-risk models and establish a workflow for human experts to validate their outputs. Define risk levels and the degree of control attached to each, set evaluation criteria, and ensure experts are trained and equipped to perform their roles effectively. Establish a process for review or approval, and document the actions and outcomes of human reviewers to ensure accountability and continuous improvement.

Control 7.3: Implement an escalation process for AI-generated outputs that require further investigation. This process should include clear procedures for escalating issues to higher levels of review and investigation, ensuring that critical or problematic outputs receive appropriate attention and resolution, including outreach to end users to update or correct Gen AI-generated outputs or decisions, as applicable.

3.2.8 User Feedback and Iterative Improvement

Description: This approach focuses on actively engaging users throughout the post-deployment lifecycle to gather insights on the quality and relevance of Gen AI-generated outputs. Feedback mechanisms enable users to report discrepancies, biases, or inaccuracies, which are then used to refine the model's performance. This iterative process involves adjusting model confidence levels and fine-tuning the model to produce more relevant and accurate answers, ultimately reducing the likelihood of generating biased or harmful content. Furthermore, user feedback helps in continuously updating and improving the model based on real-world interactions. By integrating this feedback into the training and development process, AI systems can evolve to better meet user needs and adhere to safety standards.

User feedback is crucial for identifying issues such as overconfidence, hallucinations, or misinterpretations, allowing for systematic improvements through fine-tuning and updates. Continuous monitoring through user input enables swift corrective actions for problematic outputs, ensuring that the AI system remains aligned with safety standards. Clear feedback channels also promote shared accountability between the organisation and third-party providers in the AI's development and deployment.

Indicative complexity to implement: Lower

Control implementation: HITL moderation is an important guardrail to address risks and can be operationalised using these controls:

Control 8.1: Implement a robust feedback mechanism that allows users, customers, and stakeholders to rate the accuracy and relevance of AI-generated outputs. Inform and educate users about the purpose, scope, and limitations of the system, guiding them on how to engage with these feedback tools effectively. Analyse feedback to identify and prioritise potential problems, inaccuracies, limitations, or biases in the system, and monitor them over time to track the effect of mitigations. Consider further information-gathering approaches, such as user interviews, where needed to supplement this feedback.

Control 8.2: Provide relevant context links alongside Gen AI-generated responses, allowing users to review the source material from which answers are drawn. Encourage users to review links and provide feedback on output accuracy.

Control 8.3: Actively engage potential end users in requirements-gathering workshops to understand their needs and develop features for current and possible future use cases. Consider ways to involve users in testing to ensure that the model meets their needs. Establish feedback mechanisms and consider how to prioritise and address key issues effectively. Capture lessons learned throughout this process to enhance future projects, ensuring alignment between Gen AI solutions and user requirements.

3.2.9 User Transparency and Consent

Description: This approach includes providing disclaimers and disclosures to the end user when they access Gen AI applications. Rather than focusing solely on whether Gen AI is being used, disclosure obligations should be based on the level of risk posed by the AI application. Higher-risk AI applications, particularly those with potential to harm individuals, require clear and transparent communication. This may include disclosing the role of AI in content generation or decision-making processes when it has significant implications for users. These measures collectively ensure that users are well-informed and can make conscious decisions about their interactions with Gen AI content, and will never mistake an AI interaction for a human one.

Indicative complexity to implement: Lower

Control implementation:

Control 9.1: Depending on the risk level of the use case, consider incorporating clear and prominent disclosures within the model interface to indicate when Gen AI technology has been used in producing content. The language used should be straightforward and accessible, so users can easily understand which content that they are interacting with was generated or influenced by AI.

Control 9.2: Implement effective disclaimer messages within the user interface that remind users to exercise caution when utilising AI-generated information. Additionally, where doing so would not be overly burdensome or repetitive, include reminders to fact-check outputs and provide reference links, citations, or contact points to help users verify information and address any accuracy concerns. Provide a level of disclaimers that is appropriate to the level of risk of the use case, and where users are internal, ensure that they are trained to understand and act on those disclaimers effectively.

Control 9.3: Apply watermarking to Gen AI-generated image and video content to clearly indicate its origin, especially when that content is used for promotional or marketing activities that could be viewed externally. Ensure that internal users are informed when interacting with generated content, such as in training modules, so they can be watchful for potential Gen AI issues. Provide guidance to internal users on the potential pitfalls and uses of AI-generated content and ensure they can provide traceability for any actions or decisions made based on this output.

04 Applying Guardrails to Use Case Categories

4.1 Methodology for Use Case Application

Organisations can leverage the below methodology to help them to identify relevant guardrails and controls to address Gen AI-related risks.

Identify Business Case: Start by identifying the specific business case to which Gen AI would be applied. This involves understanding the business problem or opportunity, the goals the organisation aims to achieve, and the potential benefits of using Gen AI in this context.

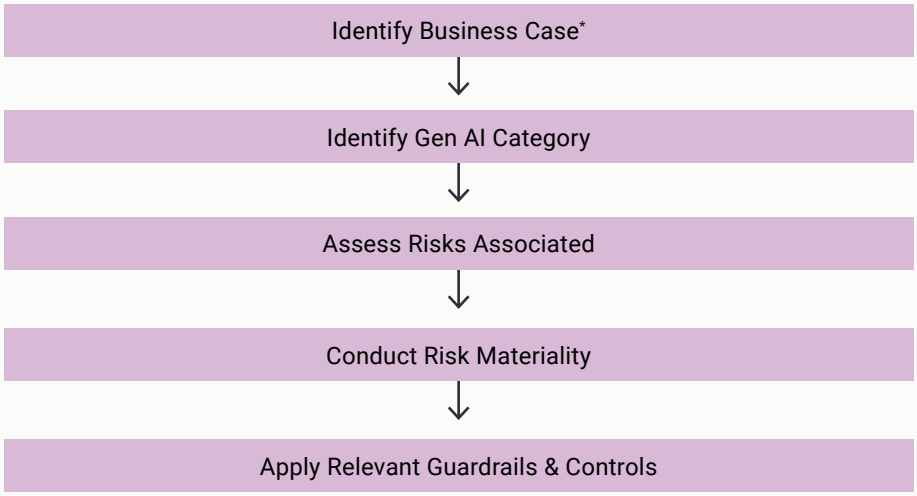
Identify Gen AI Category: Next, determine the category of Gen AI capability that will be used. This could include, but is not limited to, capabilities such as knowledge management, document summarisation, or translation.

Assess Risks Associated with the Use Case: Conduct a thorough risk assessment to identify potential risks associated with the use case. This includes both enterprise-wide risks (e.g., lack of Gen AI risk awareness, inadequate human oversight, etc.,) and system-specific risks (e.g., hallucination, insufficient model accuracy, toxic outputs, etc.,).

Conduct Risk Materiality: Evaluate the materiality of the identified risks. This involves assessing the likelihood and impact of each risk, and determining which risks are most significant and require immediate attention.

Apply Relevant Guardrails and Controls: Depending on the materiality of the risk identified, implement appropriate guardrails and controls covered in section 3.2 to mitigate the identified risks. The extent and complexity of guardrails employed should be proportional to the use case's risk materiality. This includes both technical measures (e.g., filtering and control, customised model design) and process measures (e.g., human-in-the-loop moderation, user feedback and iterative improvement). Ensure that these controls are applied such that risks are mitigated throughout the Gen AI development lifecycle.

Figure 7: Methodology for selection of relevant guardrails and controls for use case application



*Applicability and relevancy of respective guardrails have to be taken in context to the materiality of Gen AI usage in relation to the business case.

4.2 Use Case Study – Document Extraction / Summarisation

4.2.1 Description of the Use Case Category

In the banking industry, Gen AI is utilised for document summarisation and extraction to streamline the processing of vast amounts of financial documents, such as loan applications, compliance reports, and transaction records. In summarisation, AI models distil lengthy reports and financial statements into concise summaries, highlighting key metrics and insights, which expedites decision-making and risk assessment. In document extraction, AI models identify crucial data points like transaction details, customer information, and compliance-related data from unstructured text, converting them into structured formats for easy analysis and reporting. This not only enhances operational efficiency and accuracy, but also helps banks in meeting regulatory requirements and improving customer service by rapidly transforming relevant information into a usable, structured format. Document extraction and summarisation use similar technologies, and can often be applied together in a single solution to extract key information from documents and provide concise summaries.

Document extraction and summarisation as a technology has several practical use cases in a bank:

- **Enterprise knowledge management.** Over the past several decades banks have accumulated large knowledge bases of extensive internal documentation, especially on their technology systems. Gen AI-powered document extraction and summarisation can make large or unwieldy knowledge bases easier to interact with by allowing users to pose detailed queries and receive customised responses without the need to manually review numerous documents.
- **Compliance and risk management.** A consortium of banks in Singapore, as part of Project MindForge[3], experimented a utility powered by Gen AI for extracting useful information from documents related to compliance and risk management, such as supervisory notices and news stories, and summarising them for bank employees to rapidly action.
- **Agent assist.** Call-centre agents in banks are being equipped with Gen AI "assistants" that, among other functionality, can call up help and support documents and summarise notes or transcripts from previous calls with the same customer. This an relieve pressure on agents and reduce the time they spend searching for information and can improve the customer experience.
- **Relationship management copilot.** Relationship managers, especially in private and commercial banking, can use Gen AI "assistants" to summarise customer files, notes from meetings, and other key information. This can free up relationship managers to spend more time working with their clients, furthering a key trend in commercial and private banking towards more hands-on human interaction.
- **Know Your Customer (KYC) advisor.** The large volume of unstructured information, especially text, available for Anti Money Laundering (AML)/KYC can be automatically reviewed using Gen AI agents. These agents can rapidly ingest documentation related to a specific party or transaction and direct the attention of human reviewers towards key documents or issues, facilitating more accurate compliance activities at reduced cost.
- **Office assistant.** A Gen AI-powered tool with general document summarisation capabilities can support office workers on day-to-day productivity tasks, such as summarising emails or memos.

[3]Read more about Project MindForge at: https://www.mas.gov.sg/schemes-and-initiatives/project-mindforge.

4.2.2 Common Risks Involved

Document Extraction / Summarisation, in a banking context, presents several of the risks outlined in Section 2.3. Refer to Figure 2 for a full mapping of the risks pertinent to each Gen AI use case category.

Insufficient model accuracy and hallucination, fabrication, or confabulation can each result in inaccurate or misleading summaries, potentially harming users and undermining the intended use case. This is especially challenging in banking applications where accuracy is critical – such as where document summaries are employed as part of KYC obligations – but can also have business impacts in common applications like relationship management. Additionally, these systems may exhibit overconfidence, leading them to firmly assert conclusions based on tenuous or contradictory sources and cause users to overlook important details or differing perspectives. This is especially threatening for emerging banking use cases that deploy Gen AI tools for risk identification and horizon scanning from media reports. The absence of adequate human oversight further exacerbates this issue, where the neglect of complex validation tasks can allow errors to persist unchecked. This poses a significant risk when tools are used for AML/KYC obligations.

These systems may also generate toxic or offensive outputs when drawing from unfiltered sources, which can damage user trust and the organisation's reputation when present in a customer-facing application, such as a customer service bot that summarises data from web pages that contain user-generated content. Finally, a lack of awareness among users regarding the risks associated with Gen AI outputs could lead to uncritical usage and potentially harmful consequences, such as when document summarisation tools are used in employment decisions, used to summarise sources that contain personal data, or used in regulated applications like AML/KYC without appropriate risk management.

4.2.3 Application of Guardrails for Risk Mitigation

Each of the risks potentially posed by this use case can be mitigated through the application of a risk-informed selection of the guardrails discussed in the earlier section of this paper. See Figure 4 for a detailed mapping of each guardrail approach to the risks it primarily impacts.

Document summarisation for internal knowledge management on day-to-day productivity tasks may present a very low level of risk. Conversely, document summarisation that handles confidential customer data, and especially document extraction / summarisation that is used to make critical managerial decisions or decisions with compliance implications like underwriting or KYC assessments, would have a high level of risk and would require a more rigorous and extensive guardrail application.

Each of the nine guardrails could be applied to a document extraction / summarisation use case. It is not always necessary to do so; the selection of guardrails among and beyond these options should be based on the specific characteristics, context, and riskiness of the use in question. The guardrail descriptions below are therefore neither prescriptive nor exhaustive for this use case.

Where relevant and risk-appropriate, this paper's guardrails could be applied to a document extraction / summarisation use case in the following ways:

- **Enterprise Governance and Training:** As a transversal, enterprise-wide guardrail, enterprise governance and training does not differ by use case.
- **Filtering and Control:** In addition to screening for toxicity and discriminatory language at the input and output levels, set the temperature to 0 and equip the LLM with meta-prompts instructing it to summarise documents accurately, cite sources, and express reasonable levels of confidence. If risk-appropriate, deploy a screening tool at the output level which measures relevancy Key Performance Indicators (KPIs) like semantic similarity between outputs and source documents and re-generates irrelevant content. If risk-appropriate, add an input-filtering tool that will automatically reject prompts that are not relevant to the document extraction / summarisation task at hand.
- **Customised Model Design:** For higher-risk use cases, consider fine-tuning the LLM with a series of accurate and non-sensitive document/summary pairs – with summaries written by a human – that demonstrate the desired style and which deal with use case-relevant data. Periodically review whether the types and formats of documents being summarised by users match the intended use cases.
- **Red Teaming:** Red team the system before deployment by attempting to prompt it to produce toxic text or to inaccurately summarise documents / extract irrelevant information. Test in particular for the system's suggestibility and ensure that document summaries are accurate despite variations in user suggestions or prompt wordings.
- **Prompt Design:** Implement a prompt design strategy that uses chain of thought techniques to ensure that the system addresses all sources and all aspects of the prompt. Give the system several examples of accurate, well-written summaries.
- **Monitoring and Validation:** Measure and validate the system against the semantic similarity between document summaries and source files. Also measure and require the system to adhere to low thresholds for industry-standard measures for toxicity. Set a low threshold for system perplexity and test it under a variety of circumstances, requiring it to produce summaries with a high degree of certainty or to reject the prompt.

- **Human-in-the-Loop Moderation:** Periodically sample document summaries and refer them to the system owner or a relevant expert for human verification. For very high-risk extraction or summarisation use cases, consider implementing a mandatory human review procedure that requires the system owner or the user to check all extracted data and confirm its veracity.
- **User Feedback and Iterative Improvement:** Give end users an option within the tool itself to flag and share inaccurate summaries.
- **User Transparency and Consent:** Clearly indicate at the interface level where document summaries are generated using AI and indicate that AI summaries can be inaccurate. Prompt users to fact-check extracted or summarised information; in the case of customer users, they can be directed to public web pages or information in their account, while employees can be periodically retrained on the limitations and use of Gen AI in their workflows, especially when impacting critical applications like KYC. Employee consent to take responsibility for their own use of Gen AI can be obtained as part of onboarding training activities.

4.3 Use Case Study – Code Generation

4.3.1 Description of the Use Case Category

Generative AI for coding leverages recent advancements in LLMs and Natural Language Processing (NLP) to enhance programming efficiency. By employing deep learning algorithms trained on extensive datasets of publicly available code, these tools enable programmers to input plain text descriptions of desired functionality. The AI then suggests code snippets or complete functions, automating or streamlining repetitive and simple tasks such as code translation and modernisation, like converting Common Business-Oriented Language (COBOL) to Java. Despite the increasing accuracy of AI-generated code, human review remains crucial as the output may still have errors, though some tools assist with this by automatically generating unit tests.

In a bank, code generation has several practical uses:

- **Software development assistant.** Developers can use a code generation system to assist them in creating new software, either at a tactical and incremental scale – helping developers fill in boilerplate code or complete mundane tasks more quickly through code snippets – or at scale, to generate large segments of code based on the functionality specified in the prompt. Software development assistants receive prompts from a developer or from an intermediate system that specify, in natural language, the desired functionality. This can improve the speed of software development and its adherence to common enterprise rules, making it particularly useful for in the complex and highly regulated software environments in banking.

- **Automatic testing and debugging.** Code generation systems can rapidly identify areas of potential issues by predicting code that is likely to create errors. It can also generate unit tests and other debugging scripts that are customised to a section of code, accelerating the testing and debugging process.
- **Accelerated or automated translation between coding languages.** One of the most promising and powerful use cases of Gen AI in coding is its ability to translate code from one language to another – usually deployed in the context of updating old or legacy code to modern standards. This is especially relevant for legacy banks with large codebases in older languages.
- **Documentation generation based on code.** Using code-to-text generation tools, Gen AI can create documentation in the format and style typically used by the enterprise to document its software. This saves significant developer time, which can more usefully be spent on technical tasks, and can be enhanced when developers provide key bullet points or an outline. This can facilitate regulatory compliance or risk management activities by improving the availability of code documentation to non-technical staff.
- **Prototype or proof-of-concept development based on natural language prompts.** While it is not currently feasible to put AI-generated software directly into production without verification, code generation tools can be used to rapidly generate features or tools to prototype or demonstrate functionality for design purposes. The ability to do so based on natural language prompts allows even non-technical executives to do so. This code can be re-developed using standard software development processes if it is desired for production.

4.3.2 Common Risks Involved

In a banking context, code generation systems face several risks that can undermine their effectiveness. Refer to Figure 2 for a full mapping of the risks in scope to each Gen AI capability.

Hallucination, fabrication, or confabulation can lead to the creation of code with inaccurate variables or functionalities. Insufficient model accuracy may result in non-functional code due to logic errors, endless loops, or syntax issues, and even functional code might not meet security standards. These models are also susceptible to model degradation from unexpected use. Additionally, they may display overconfidence, generating subpar or non-functional code when uncertain about a request. A lack of awareness among users regarding potential errors and risks can lead to improper deployment without adequate testing or mitigation. Furthermore, the absence of use case and model governance and inadequate feedback mechanisms could allow flaws to go uncorrected.

Unrepresentative or biased data inputs may result in low-quality code, while insufficient human oversight can allow optimisation, security, and functionality errors to persist undetected. This can be particularly challenging when generating documentation about code in cases where that documentation's quality has implications for technology risk or regulatory compliance.

These are common risks that applications of code generators across industries would share. Banking is primarily differentiated in the potential impact of these risks. Developing software that touches on core banking functions raises the risk that functionality or security flaws could have far-reaching impacts on the business or customers, or could result in regulatory scrutiny. Even software that does not impact core banking functions can have significant impacts on a bank's customers, reputation, or regulatory exposure. Banks as such face a higher level of risk than peer institutions in other industries.

4.3.3 Application of Guardrails for Risk Mitigation

Each of the risks potentially posed by this use case could be mitigated through the application of a risk-informed selection of the guardrails discussed above. See Figure 4 for a detailed mapping of each guardrail approach to the risks it primarily impacts.

The risk posed by code generation implementations is linked to the use of that code and the extent to which the model is involved in producing it. A small code generation utility that checks errors or generates documentation to assist developers may qualify as low risk, given that the harm created by potential errors is small, and that redundancies have the potential to catch errors in advance. High-risk applications may involve automated code production or translation at scale, where large volumes of code are automatically produced with minimal human involvement, or where generated code controls high-risk systems like critical infrastructure.

Most code generation in banks will be conducted by Software-As-A-Service (SaaS) coding assistants licensed from major cloud service providers. Major SaaS coding assistants are highly capable and often general-purpose, having the ability to be used in a variety of use cases ranging from providing minor suggestions to generating code wholesale in the development of critical features; it is important to balance governance of the tool from governance of each use case. Banks may consider balancing the application of guardrails on the tool transversally – for all users no matter what they are using the tool for – with guardrails or restrictions that are applied only when the tool is used in a high-risk use case. User education is especially important when governing the use of general-purpose tools.

Each of the nine guardrails could be applied to a code generation use case. It is not always necessary to do so; the selection of guardrails among and beyond these options should be based on the specific characteristics, context, and riskiness of the use in question. The guardrail descriptions below are therefore neither prescriptive nor exhaustive for this use case.

Where relevant and risk-appropriate, this paper's guardrails could be applied to a code generation use case in the following ways:

- **Enterprise Governance and Training:** As a transversal, enterprise-wide guardrail, enterprise governance and training does not differ by use case and is applicable to all Gen AI use cases.
- **Filtering and Control:** Set the temperature to 0 and equip the LLM with meta-prompts specifying reference datasets and referencing good coding practices; together, these can improve its ability to generate code that meets enterprise requirements and avoids hallucinations. If risk-appropriate, deploy a screening tool at the output level which can test generated code for known security vulnerabilities, efficiency, or simple functionality. If risk-appropriate, deploy a user prompt screen using an agentic model to determine if requests for code go beyond the system's intended use, and inform the user accordingly.
- **Customised Model Design:** For higher-risk use cases, consider fine-tuning the model on internal code that is well-constructed, accurate, and which covers a variety of predictable use cases. Include a range of types of coding problems, languages, and enterprise systems to ensure that the model is well-trained on a sample that is representative of the types of problems it will encounter in usage.
- **Red Teaming:** Red team the system before deployment by attempting to prompt it to produce code or documentation that is beyond the intended scope of the model, code that contains impossible requirements, or code that contains security vulnerabilities. Code generation red teaming is particularly well-suited to red teaming because the functioning of generated code can be objectively and quantitatively verified.
- **Prompt Design:** Implement a prompt design strategy that uses chain of thought techniques (for e.g., separately prompt the model to generate code for use login, data encryption, push notifications, and in-app purchases in a mobile app).
- **Monitoring and Validation:** Track standard metrics for code performance, such as that code's performance, efficiency, and the detection rate of errors, both through automated verification tools and by human reviewers.

- **Human-in-the-Loop Moderation:** Require human review of all code generated by the model before deploying to a production environment to verify the presence of security issues, its efficiency, and its ability to function. Ensure that AI-generated code and documentation are subjected to appropriate testing before being placed in production, and that all AI-generated code is marked as such. Escalate any issues to the model development team for future improvement, and periodically review code samples to assess and improve the model's functioning.
- **User Feedback and Iterative Improvement:** Give end users an option within the tool itself to flag and share code that contains issues or that fails to respond to their prompt.
- **User Transparency and Consent:** Clearly inform end users of the limitations of the software, and mark AI-generated code with comments wherever appropriate. Integrate reminders to check AI-generated code into the tool's user interface. Watermark AI-generated coding documentation and highlight the potential for errors.

05 Next Steps

This Handbook outlines guardrails for a subset of Gen AI risks included in the risk taxonomy published by Project MindForge. These guardrails are intended as a foundational reference for the banking industry as it begins to explore and implement various Gen AI use cases. This paper will serve as an initial guide, providing a structured approach to managing and mitigating specific Gen AI risks.

Gen AI risks and guardrails extend beyond those included in this paper. The insights of this Handbook will serve as a foundation for ABS' engagement in the development of the MindForge AI Governance Handbook, which is being collaboratively developed by a consortium of financial institutions supported by the Monetary Authority of Singapore.

The handbook will build upon this foundational work to enhance and broaden the scope of guardrails, addressing emerging risks and incorporating new learning as the Gen AI landscape evolves. Through this iterative process, the industry will continuously improve its understanding and management of Gen AI risks, ensuring that the guardrails in practice remain relevant and effective in addressing the challenges faced by the industry both today and in the future.

The guardrails outlined in this Handbook also present an opportunity to contribute to Project Moonshot, which focuses on developing LLM evaluation toolkits designed to test the robustness of LLM models and applications. The proposed guardrails will be integrated into the toolkit to enhance its banking-specific evaluation capabilities.

This Handbook, by articulating risks and practical guardrails for Gen AI based on the experience to date of ABS members, will enable the next phase of innovation in Singapore's AI ecosystem. While this Handbook's scope has been limited to a selection of Gen AI risks and guardrails, future work will expand on it to address emerging risks such as – but not limited to – Intellectual Property (IP), privacy, monitoring, robustness, cybersecurity, and data security. Taken together, the Handbook and this future work will enable a more systematic approach to managing the diverse risks associated with Gen AI in the financial sector.

Appendix

This Handbook is supplemented by an Excel Tool, which is intended to help practitioners understand Gen AI risks and the possible guardrails that can help them address those risks in a practical format. While this Handbook is intended for leaders and business users, the Excel Tool explains the key findings of this paper with a focus on developers, deployers, model reviewers and validators. Practitioners in the AI field are encouraged to use this file pragmatically and often when designing Gen AI systems and considering how to govern those systems through the application of effective guardrails.

The choice of a spreadsheet structure for this supplementary tool was deliberate and reflects the emerging nature of this field; banks that use the Excel Tool are encouraged to add to its controls and guardrails based on their own experience and needs.

The Excel Tool, like this paper, is meant to be a guideline, and banks retain responsibility for their implementation of Gen AI. Users must perform due diligence, applying a robust risk materiality-based approach when selecting guardrails and implementing the recommendations in the file.

The Excel Tool consists of five sheets:

0 How to Use

This introductory sheet describes the purpose of the file and its structure. It also contains key disclaimers on its use.

1 Inherent Risks and Use Cases

This sheet maps the inherent risks (as identified through the MindForge risk taxonomy) across the seven primary use case categories. It helps users identify specific risks tied to various Gen AI applications, providing a foundation for understanding the risks involved in different use cases.

2 Inherent Risks and Guardrails

This sheet includes the mapping of top inherent risks to the appropriate guardrail approaches. Each risk is mapped to potential guardrails that can mitigate or manage those risks.

3 Guardrails and Controls

This sheet outlines several guardrails and how their related controls can be implemented to address the identified inherent risks. This sheet serves as a practical guide for guardrail application.

4 Definitions

This sheet offers clear definitions of the use case categories, inherent risks, and guardrail approaches covered in the document. The terms it defines are:

- The use cases described in this Handbook
- The risks described in this Handbook
- The guardrails described in this Handbook