Digital Policy Office
The Government of the Hong Kong Special Administrative Region
of the People's Republic of China

香港生成式人工智能研發中心
Hong Kong Generative AI Research & Development Center

Hong Kong
Generative Artificial Intelligence Technical and Application Guideline

Version: 1.1

December 2025

Preamble

The Digital Policy Office (DPO) of the Government of the Hong Kong Special Administrative Region (the HKSAR Government) has commissioned the Hong Kong Generative AI Research and Development Center (HKGAI), which was established under the funding support of InnoHK, to assist in research and formulation of the “Hong Kong Generative Artificial Intelligence Technical and Application Guideline”, so as to provide relevant codes and practices on the technology and the use of generative artificial intelligence (AI) technologies for reference by various sectors.

This Guideline documents the technical background and governance principles of generative AI, and provides a practical guide to Technology Developers, Service Providers and Service Users.

The DPO will continue to review the latest technologies and application development related to generative AI and update the content of the guideline regularly. Through the collaborated efforts of Technology Developers, Service Providers, Service Users and other broad stakeholders, generative AI can better serve Hong Kong’s society, bringing more convenience and welfare to the public.

Note: In the event of any discrepancies between the Chinese and English versions of this document, please refer to the English version and notify the Digital Policy Office.

Table of Contents

Table of Contents

Background

Generative artificial intelligence (AI) is an important branch of modern AI represented by machine learning. By leveraging various machine learning algorithms, generative AI can automatically generate content information such as texts, images, audios, and videos according to complex human intentions and instructions. In recent years, with the breakthroughs in generative AI technology, relevant products and services have been rapidly applied and promoted. Compared to other AI applications and traditional internet applications, their greatest advantage lies in their more personalised and convenient content creation functions, marking an important milestone in the journey towards general-purpose AI.

The wave of technological transformation driven by generative AI is sweeping the world, and it is expected to be widely applied in more fields in the future, exerting a profound impact on economic and social development and the course of human civilisation. However, at the same time, this technology also faces complex challenges such as unpredictable security risks and ethical issues. The governance of AI has become a common issue faced by countries around the world. Only by ensuring that the development and application of generative AI technology always run on a safe track can we promote the steady and sustainable advancement of AI.

The Government of the Hong Kong Special Administrative Region (HKSAR Government) is keenly aware that Hong Kong must keep up with the technological development trends, formulate AI governance plans that are compatible with these trends, and actively promote the development of AI technology to contribute to the economic and social development of Hong Kong. To this end, the HKSAR Government actively draws on international best practices and collaborates with experts in the local industry and the field of innovation and technology, committing itself to enhancing the resilience of the AI industry ecosystem in Hong Kong. The HKSAR Government has commissioned the Hong Kong Generative AI Research and Development Center (HKGAI), which specialises in generative AI technology, to conduct research on appropriate guiding principles regarding credibility, accountability, ethical safety in the development and application of generative AI technology, and to propose recommendations accordingly.

Commissioned by the HKSAR Government, the HKGAI has formulated the “Hong Kong Generative Artificial Intelligence Technical and Application Guideline” (the Guideline), taking into account expert opinions from the industry and international best practices. The Guideline aims to encourage relevant stakeholders within the Hong Kong Special Administrative Region to engage in various generative AI-related business

activities safely and responsibly, strictly adhering to ethical standards, moral guidelines, and legal regulations. The Guideline also provides practical guidance for all stakeholders to effectively address the security issues and social risks arising from generative AI technology.

The Guideline aims to ensure the timely communication of governance developments and the sharing of best practices. It aims to balance innovation with social responsibility in the development of generative AI in Hong Kong, thereby maximising the benefits of technological advancement and minimising associated risks. The establishment of this framework will pave the way for a robust governance structure for generative AI, which is essential for promoting the open application ecosystem and healthy development of generative AI across industries in Hong Kong.

The Guideline is specifically tailored for the following stakeholders:

1. Technology developers and those who commission the development of technology or determine the use of technology (Technology Developers);
2. Service providers, platform providers, and individuals who provide service with additional features and tools based on existing technology (Service Providers); and
3. Service users, content creators, and disseminators of generative content (Service Users).

As generative AI technology services become increasingly prevalent, they bring about a multitude of security and social risks. To this end, this Guideline offers practical guidance and recommendations. For Technology Developers, it is important to focus on technical risks such as data breaches, model biases, and errors. The Guideline provides best practices for secure development and design to ensure the reliability and stability of the technology. For Service Providers, they need to pay attention to the compliance and security of their services, especially when handling data from Service Users and generative content. The Guideline provides suggestions to ensure the secure operation of their services. Service Users need to understand the potential risks associated with using generative AI services, avoid illegal and unethical behaviour during use, and learn how to protect personal privacy and data security. The Guideline offers advice on using technology services, as well as measures for identifying and addressing potential security threats.

1. Brief Introduction to Generative AI

1.1 Overview of Generative AI

Generative AI refers to the use of various machine learning algorithms to enable computer systems to automatically generate content information such as text, image, audio, video, code or other media, based on vast amounts of data, according to complex human intentions and instructions.

Compared with other AI applications and traditional Internet applications, generative AI has the ability of content generation and creation. The information content provided by this service is not the information already existing on the Internet, but the information content newly created and generated by the model algorithm based on the input instructions of the Service Users. Service Users can generate content or solutions that meet their expectations according to their own needs, so as to obtain a more personalised and customised service experience.

Generative AI technology is a branch of the modern AI field represented by machine learning, and it serves as the underlying algorithms, models, and architectures that support generative AI services. Generative AI services refer to the products, solutions, and services developed using generative AI technology.

Throughout the entire lifecycle of generative AI technology, from its research and development to its application, there are three important roles: **Technology Developers, Service Providers, and Service Users.**

1.2 Main Principles of Generative AI Technology

Generative AI technology is based on generative models within deep learning. Through various training strategies such as unsupervised or self-supervised learning, autoregression, and more, the models learn the distributional characteristics of a large amount of training data, thereby facilitating the generation of new content. The methods and strategies for generation vary by model and can involve generating from random noise or initial inputs step by step, or compressing data to form a “Latent Space”, sampling within the “Latent Space”, and then decoding to generate data. Key technologies include latent space learning, probabilistic modelling, and generative strategies, working together to enable the creation of diverse content. Generative AI technology can produce multimodal content, including text, code, images, audio, and video.

Currently, the mainstream models include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Autoregressive Models, Diffusion Models,

Transformer Models, etc.:

- **GAN** trains two neural networks in a competitive manner. The generator network generates data, and the discriminator network determines whether the data is real or not. Through continuous optimisation in adversarial training, the generator is able to generate increasingly realistic data.
- **VAE** uses an encoding-decoding generation method. It maps the input data to the distribution in the latent space, usually a Gaussian distribution, and then samples from this distribution to generate new samples.
- **Autoregressive model** is based on the idea of step-by-step generation. It takes the already generated content as context input and predicts the next element to generate new data, which is suitable for generating ordered data sequences.
- **Diffusion model** uses the generation method of diffusion and denoising. First, noise is gradually added to the original data, and then new data is reconstructed from the noise by learning the inverse process.
- **Transformer model** is constructed based on the self-attention mechanism and feed-forward neural networks, adopting an encoder-decoder architecture. The generation process mainly relies on compressing a large amount of data with the encoder to form a language latent space. Then, the decoder uses the self-attention mechanism to process the context information and samples in an autoregressive manner to gradually generate new content.

Today, the term “Generative AI” primarily refers to large foundation models based on the transformer architecture, including Large Language Models (LLMs) and Vision-Language Models (VLMs). These models use self-attention mechanisms and feed-forward neural networks in an encoder-decoder architecture. The generation process compresses large amounts of data to form a latent space, then uses self-attention to process context and generate new content in an autoregressive manner.

Transformer-based models excel at producing multimodal content including text, code, image, audio, and video. Even diffusion models, which generate images by gradually removing noise from random patterns, typically operate within VLMs to process textual prompts.

While earlier approaches like GANs, VAEs, and pure autoregressive models made important contributions to the field, modern generative AI is predominantly driven by transformer-based architectures that can process and generate across multiple modalities with unprecedented capabilities.

Model pre-training, fine-tuning, and inference are important stages in the lifecycle of generative AI technology, jointly forming a complete process from

construction to deployment.

Pre-training refers to the process of initially training a model on a large scale of unannotated or weakly annotated data, with the aim of allowing the model to learn general features and knowledge from the data. This stage typically employs unsupervised learning frameworks, such as self-supervised or autoregressive methods. Models trained through autoregressive methods can effectively learn the distributional regularities of the data, thereby acquiring strong generalisation capabilities. The knowledge acquired can be easily transferred to specific tasks, thereby reducing the need for annotated data in specific domains.

Fine-tuning is a crucial step to enhance the performance of a model in specific tasks or domains. After the model has been pre-trained, it needs to be further trained on a small-scale labelled dataset that caters to specific requirements. The parameters of some or all layers of the model will be fine-tuned according to the data of the target task, enabling the model to master the knowledge of specific domains. Fine-tuning optimises the model for specific application scenarios. Compared with training from scratch, fine-tuning on the existing pre-trained model can reduce the training time and computational costs, and significantly lower the usage threshold of large-scale parameter models. Common fine-tuning techniques include instruction fine-tuning, Reinforcement Learning from Human Feedback (RLHF), Direct Preference Optimisation, Low-Rank Adaptation (LoRA), etc.

Inference is the final link in the lifecycle of generative AI and also a crucial stage for the practical application of the model. In the inference stage, the already trained model will stop updating its own parameters and instead conduct real-time processing based on the input actual data to generate output results. In order for the trained model to handle the input from Service Users or the data in practical applications, it is necessary to deploy it in the form of a service and launch it into the production environment, so as to meet the needs in different scenarios. The speed, efficiency, and accuracy of inference are key indicators for evaluating the performance of the model. To improve the inference efficiency, technical measures such as model compression and optimisation can be adopted.

1.3 Main Modalities of Content Generation

1.3.1 Text Generation

Text generation refers to the automatic generation of natural language text from input data using natural language processing technology. It usually relies on natural language processing technology, and currently, large language models, which are the

most popular, are also one of the key technologies for text generation. By learning language patterns and rules from a large amount of text data, the model can generate coherent and grammatically correct text according to the given input. There are two basic modes of text generation:

- In unimodal text generation, when generating text, only text is used as both input and output. Based on the given text prompts or context, it predicts and generates the subsequent text content.
- Multimodal generation is capable of processing multiple types of data simultaneously, such as text, images, audios, videos, etc., and establishing connections and interactions among these different modal data, so as to generate text.

1.3.2 Image Generation

Image generation utilises generative technology to automatically generate brand-new image content through the learning and analysis of a large amount of image data. Usually, generative adversarial networks or diffusion models in deep learning are used as the training methods for image generation. The modes of image generation are as follows:

- In the text-to-image mode, Service Users input text descriptions. The model parses the semantic information and converts it into a vector representation, and then generates and decodes it into an image in the latent space based on the learned knowledge.
- In the image-to-image mode, feature extraction and encoding are carried out on the input image, and after operation and transformation in the latent space, it is reconstructed into a new image.

1.3.3 Audio Generation

Audio generation refers to the process of using AI technology to synthesise the corresponding sound waveforms according to the input data, usually relying on deep neural networks to generate the spectrum of audio signals. The modes of audio generation are:

- Text-to-audio converts text input into audio output. By understanding the input text information, it generates audio that conforms to or is related to the text content (such as composing music for the input lyrics).
- Audio-to-audio means generating new audio based on the existing audio. The generative model can learn the patterns and characteristics of the existing audio samples, and generate relevant new audio content based on these patterns and

characteristics (such as composing music by listening to the sound).

1.3.4 Video Generation

Video generation refers to the process of using AI technology to synthesise corresponding video content according to the input data (such as texts, images, video clips, etc.). It can learn the patterns and characteristics of videos from a large amount of data, and generate new video content based on these patterns and characteristics.

Video generation is an extension and expansion of image generation. In video generation, not only does it need to generate high-quality images for each frame, but also deep learning models are required to understand and synthesise the image frames in the time sequence and the corresponding backgrounds, so as to generate a coherent video stream that conforms to the laws of the physical world.

1.4 Main Service Areas of Generative AI

Generative AI services can accept inputs such as texts, images, audios, videos, and codes, and generate new content in various forms. Currently, they have been widely applied in multiple industries, including the following typical service scenarios:

- **Knowledge Q&A:** Based on the knowledge and patterns in its training data, combined with an external knowledge base, generative AI can respond in real-time to the knowledge inquiries of Service Users, providing accurate and detailed information answers. It is widely used in scenarios such as intelligent customer service, internal corporate knowledge bases, academic research, and education.
- **Role-playing:** Generative AI can simulate interactions with specific characters through a dialogue system, such as historical figures, fictional characters, customer service staff, etc., providing Service Users with a virtual reality experience. It can be used in fields such as entertainment, education, and training.
- **Writing Assistance:** Generative AI can help Service Users with various types of writing, such as article creation, story writing, thesis writing, copywriting planning, etc. It can provide functions such as creative inspiration, grammar correction, content polishing, and text translation, improving writing efficiency and quality. It is widely applied in fields such as daily office work, media, advertising, and academia.
- **Programming Assistance:** Generative AI can assist programmers in writing codes, providing code suggestions, conducting code reviews, explaining code

functions, etc., helping to improve programming efficiency. It can be applied in fields such as software development and data analysis.

- **Mathematical Reasoning:** Generative AI is capable of performing complex mathematical calculations and logical reasoning. It can be applied to solving mathematical problems, conducting mathematical proofs, providing ideas and methods for solving mathematical problems, and assisting in mathematical research.
- **AI Agents:** Generative AI has the ability to make judgments and decisions. It can perceive the environment, make decisions, take actions, and interact with other systems at the same time. It can usually be deployed in various complex application environments and is widely used in industries such as industry, healthcare, and finance, for example, in autonomous driving vehicles, intelligent robots, intelligent investment, etc.
- **AI Art:** Generative AI is widely applied in the field of artistic creation. It can generate paintings with diverse styles according to user instructions or reference materials, which are used in fields such as art and design to provide creative inspiration and assistance. It can understand music theory and melody structures, generate music of different styles and types, lower the threshold for music creation, and contribute to music composition and soundtrack production. It can also generate various videos such as animations and short films, and is applied in fields such as film and television, and game development, reducing production costs and thresholds.

Figure 1. Main Service Areas of Generative AI

2. Governance of Generative AI

2.1 Technical Limitations and Service Risks of Generative AI

2.1.1 Technical Limitations

Generative AI models have made significant strides in accurately understanding complex semantics and generating high-quality content, meeting the general standards required for everyday life and production needs. However, generative AI models still have numerous technical limitations, referred to as inherent model issues. These inherent model issues can ultimately lead to the generation of harmful content. Therefore, the Guideline encourages Technology Developers, Service Providers, and Service Users to understand the following technical limitations of the models and their associated risks:

- **Model Hallucination** refers to the phenomenon where the information generated by a model does not align with the actual circumstances of the real world or contradicts the intentions of the Service Users. The root of the problem lies in the model's own generative mechanism. Models generate content based on the distribution and patterns learned from data through statistical learning rules of algorithms. During this process, the limitations of the data, the complexity of the algorithms, and the model's limited semantic understanding are the main causes leading to the occurrence of model hallucinations. According to the current level of technological development, although various methods can be employed to suppress model hallucinations, it has not been possible to completely eliminate them. Taking the large text-to-text generation model as an example, model hallucination can lead to outputs involving fabrication, piecing together, or grafting of information when answering questions or generating content, resulting in responses that deviate from the actual situation. Such issues not only severely affect the reliability and usability of the model's output, but also pose potential risks of misleading downstream decisions and applications.
- **Model Bias** refers to the specific preferences and tendencies that exist when a model generates content or makes decisions. This ethical issue leads to a lack of fairness and impartiality in content. Model bias permeates the entire lifecycle of the model: during the training and development phase, algorithmic design flaws lead to algorithmic bias, incomplete or unbalanced training data generates data bias, and unreasonable evaluation metrics and methods trigger evaluation bias. In the actual usage phase, models are influenced by past data to produce historical bias, affected by cultural differences leading to cultural bias, unfair preconceptions about different groups result in group bias, insufficient

sensitivity to individual characteristics cause individual bias, and influenced by interaction factors leading to interaction bias (bias resulting from the complex interactions between different input variables that the model fails to accurately capture or interpret). In addition, model performance and data timeliness change over time, resulting in temporal bias.

- **The Black Box Problem** means that the internal working mechanism of the generative AI model is complex and opaque, making it difficult for Technology Developers, Service Users, and the audience to understand how the model generates output results. This opacity of the generation mechanism brings about many challenges in terms of technology, ethics, and practical applications. For example, there are technical challenges such as difficult debugging and limited optimisation; ethical and legal issues such as responsibility attribution, fairness, and bias; user trust issues such as lack of transparency and limited decision-making support; and security risks such as adversarial attacks and abuse.
- **Mathematical Reasoning** refers to the capacity to engage with numerical and logical problems, a capability that generative AI has begun to demonstrate significant potential in. However, there is an ongoing debate within the academic community about whether generative AI is merely capable of imitation or has actually developed a certain level of scientific logical reasoning ability. In terms of performance, generative AI tends to perform less well on problems that require the application of mathematical logic, with even simple tasks such as “counting” being prone to errors. Researchers are still exploring the mathematical capabilities of generative AI.
- **Sensitivity to Input Variations** refers to the scenario where generative AI models often demonstrate disproportionate sensitivity to minor input changes, where slight variations in prompts or queries can produce substantially different outputs, undermining consistency and reliability.
- **Data Integrity** is fundamental to trustworthy generative AI systems. Specific risks include data poisoning (deliberate injection of corrupted data), data drift (gradual divergence from training distribution), and unauthorised modifications due to inadequate security controls.

To mitigate these risks, organisations should implement comprehensive data governance practices, including validation techniques, version control, and regular audits. Proper security controls and monitoring systems are essential for maintaining data integrity throughout the AI lifecycle. Organisations should provide appropriate transparency regarding data sources and processing methods while ensuring compliance with relevant regulations. Industries handling sensitive information should

consider additional safeguards such as enhanced auditing capabilities and technologies that provide immutable records where appropriate.

2.1.2 Service Risks

The proposed AI governance framework establishes a four-tiered risk classification system, creating a proportionate governance approach based on potential harm. At the highest level, “Unacceptable Risk” systems that pose existential threats such as uses causing harm or affecting human safety or subliminal manipulation should be prohibited, where developers will need to bear legal liability. “High Risk” applications deployed in critical infrastructure contexts like healthcare diagnostics or autonomous vehicles require conformity assessments, human oversight, and continuous monitoring. Systems with “Limited Risk” that have moderate societal impact, including recruitment tools or educational AI, must fulfil transparency obligations, provide user opt-out mechanisms, and undergo annual compliance audits. Finally, “Low Risk” applications such as spam filters or creative tools face minimal risks, requiring only self-certification. This calibrated approach balances innovation with appropriate safeguards, ensuring governance intensity corresponds to the potential severity of harm across the AI ecosystem.

Risk Tier	Definition	Governance Strategy
Unacceptable Risk	Systems posing existential threats (e.g., uses causing harm or affecting human safety, subliminal manipulation)	• Full prohibition • Legal liability for development/deployment
High Risk	Critical infrastructure systems (e.g., healthcare diagnostics, autonomous vehicles)	• Conformity assessments • Human-in-the-loop requirements • Real-time monitoring
Limited Risk	Systems with moderate societal impact (e.g., recruitment tools, educational AI)	• Transparency obligations • User opt-out mechanisms • Annual compliance audits
Low Risk	Minimal-risk applications (e.g., spam filters, creative tools)	• Self-certification

Table 1: Risk Classification System

In addition, generative AI models must be packaged as services or products to be brought to market. For example, OpenAI's ChatGPT is essentially a chat service based on large language models. It provides a front-end interactive interface for Service Users, while the back end can employ various models such as GPT-4o, o1, or o1-mini. Even at

the service stage, generative AI can still introduce new safety risks. This Guideline refers to these as service-derived issues, such as:

- **Content Safety** is a critical issue for generative AI services. Such services pose risks of enabling users to create or disseminate harmful content, as well as exposing audiences to such content. For instance, users might exploit generative AI services to produce dangerous content involving pornography, violence, gore, terror, or child abuse. During interaction, generative AI services might also propagate harmful values, disseminating hate speech, discriminatory remarks, or inflammatory statements, thereby making users passive recipients of harmful content. When such harmful content is created, processed further, and widely distributed, it can exert a subtle negative influence on audiences, particularly teenagers who lack discernment, potentially inducing them to engage in illegal activities, criminal behaviour, or self-harm.
- **Fabrication of Rumours** refers to the challenge wherein generative AI services can produce highly convincing text, images, audio, video, and other multimedia content. Due to their low cost, ease of use, and speed, these services can be exploited by users to deliberately create and disseminate rumours at scale, thereby confusing the public and misleading audiences. The general public often struggles to distinguish such fabricated rumours, which can significantly impact personal decision-making. As technology continues to advance, rumours generated using generative AI are expected to become even more realistic, posing severe challenges to the integrity of the societal information environment.
- **Model Jailbreaking** refers to the practice of bypassing the safety mechanisms that developers have put in place to prevent generative AI services from being used for hazardous purposes and abuse. Under normal conditions, these models are designed to identify and reject unsafe requests that fall outside the established safety parameters. However, on the user side, attack methods targeting the safety perimeter continue to emerge, and these are referred to as model jailbreaks. For example, a user might input a carefully crafted sequence of commands as part of an attack, followed by an illicit request. If the model is successfully deceived by these commands, it may process requests that it should have otherwise refused, potentially leading to severe security risks.
- **Data Leaks** refer to the challenge wherein generative AI services, particularly chat-based services, may collect user information in various forms. This includes information voluntarily provided by users, uploaded documents, and personal data accessed through devices. During the transmission and processing of this data, there is a risk of exposing private information belonging

to individuals or enterprise users.

Maintaining equilibrium between mitigating potential risks and preserving freedom of expression presents a persistent challenge to law enforcement operations. While authorities remain committed to upholding speech rights protected by the Basic Law, clear demarcation of legal boundaries becomes necessary when AI technologies are exploited for illicit activities. Such prohibited applications may include: the dissemination of instructions facilitating violent conduct (such as explosive device fabrication or assault methodologies), the generation of obscene material, the production of synthetically manipulated media content or falsified information intended to facilitate fraudulent schemes or public disruption, or the creation of malicious computational code designed to compromise information systems.

2.1.3 Model Lifecycle and Human Oversight

The lifecycle of a generative AI model can be divided into the following stages:

(1) Planning and Data Collection; (2) Model Development; (3) Deployment and Integration; and (4) Usage and Maintenance.

In light of the inherent risks associated with generative AI models, developers, organisations, and individuals should take into account the potential costs and risks while aiming for practicality and effectiveness, and thus adopt a responsible approach in the development and utilisation of generative AI at every stage. Below will elucidate the risks present throughout the lifecycle of generative AI models and indicate the responsibilities that should be assumed by all involved parties.

2.1.3.1 Planning and Data Collection

Generative AI aims to create new content, such as text, images, audio, and video. This requires the system to emulate human creativity and, to some extent, surpass existing methods of data collection and usage. Generative AI often employs self-supervised or unsupervised learning, typically requiring larger datasets than traditional AI, which increases data-related risks, such as:

- **Harmful Data Risk/Data Poisoning:** During data collection, incorrect data, outliers, or specially crafted samples by attackers may distort the model's understanding of data distribution or decision boundaries, leading to degraded performance, inappropriate outcomes or skewed model behaviour.
- **Data Bias Risk:** Variations in data sources, collection methods, and collection times can result in a lack of objectivity and comprehensiveness in the dataset. This leads to biased models that often replicate and amplify biases present in the training data.

- **Personal Data Privacy Risk:** Developers may inadvertently or intentionally handle sensitive personal data during collection, thereby violating individuals' privacy rights.
- **Intellectual Property Risk:** The training of generative AI models and the utilisation of generated content entail potential intellectual property risks, as the inadvertent use of copyrighted materials may lead to infringement claims, while the substantial resemblance of the output to existing protected work may increase the likelihood of legal challenges.

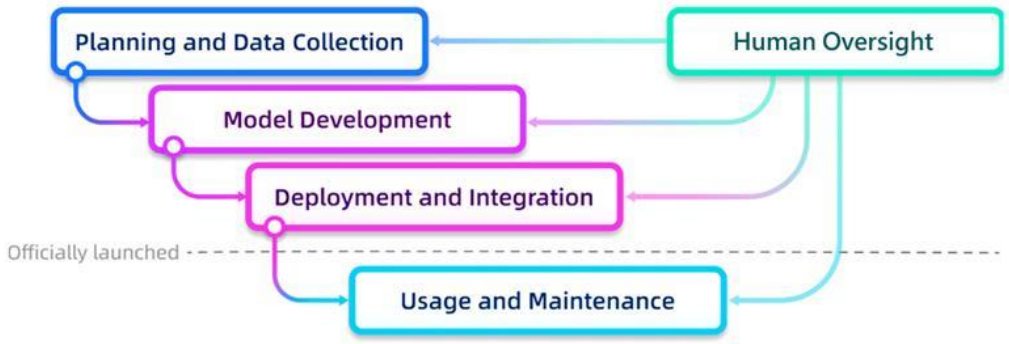

Figure 2. Lifecycle of Generative AI Model

To mitigate these risks, it is crucial to assess the source and quality of data during the planning and collection phase. Measures should be taken to ensure the standardisation of data collection and cleaning. Specifically, data sources should adhere to a principle of diversification; data distribution should be analysed using automated or manual methods, filtering harmful content; and data collection should secure proper authorisation from stakeholders to avoid infringing on personal privacy.

Domain-specific fine-tuning is particularly valuable for high-stakes sectors such as legal, regulatory, government, and banking. Given the nuanced language, legal precision, and confidentiality often required in these domains, generalised models may fall short. Targeted fine-tuning or training on curated, domain-relevant datasets such as legislative archives or regulatory frameworks can enhance accuracy, compliance, and trustworthiness.

Generative AI with multimodal capabilities can significantly enhance the delivery of public services. For instance, integrating generative AI voice assistants into government-operated communication channels can help address manpower challenges by efficiently handling routine inquiries while enabling human agents to focus on complex or sensitive cases. With capabilities such as speech recognition, sentiment analysis, and voice-to-text summarisation, these AI systems can convert large volumes of interaction data into actionable insights that inform real-time responses and support long-term policy development.

2.1.3.2 Model Development

Model development includes the design and selection of the model's architecture, planning of training algorithms, and processes for design, implementation, and evaluation. First and foremost, it is crucial to assess whether generative AI models align with business objectives, with a clear clarification in mind of the basis for decision-making. During the planning phase, ensuring the success of generative AI projects hinges on understanding the datasets required for different models, selecting the appropriate learning algorithms, and clarifying limitations regarding scalability, storage, and development time. The model development process is nonlinear. As new data arrives or business requirements change, models may need to be retrained or updated. During the development of generative AI models, the following risk issues may arise:

- **Technical Risks:** Depending on the model architecture and capabilities, models may fit the training data too closely (overfitting), making the generated content too similar to the training content, thereby losing its generalisability; or the model may not learn the patterns in the training data adequately (underfitting), leading to the model's inability to output content that meets the instructions. Both of these technical risks can impact the overall accuracy and reliability of the model.
- **Intellectual Property Risks:** Model development may involve using third-party architectures or pre-trained weights. Violations of others' intellectual property rights (e.g., patents, copyright) or licensing requirements during this process can lead to intellectual property disputes.

Therefore, during the model development phase, developers should thoroughly understand the accuracy and reliability of the chosen model and design appropriate metrics to fully assess the technical risks associated with the model. If using open-source model architectures or weights, it is also necessary to comply with the requirements of the open-source licences.

2.1.3.3 Deployment and Integration

Deployment and integration represent the final stage before users interact with the AI model. Service Providers must prepare sufficient computational resources to support the model's operation and integrate an intuitive user interface to facilitate interaction. Key risks in this phase include:

- **System Integration Risks:** Combining multiple components into a cohesive system may lead to instability due to compatibility issues, component failures, or security vulnerabilities.

- **Access Control Risks:** Components developed by different developers or third parties may inherit permissions improperly. This can lead to unintended permission propagation, creating issues such as infringement of others' intellectual property rights or data infringement.

To address these risks, Service Providers should build redundant and highly available systems to ensure individual component failures do not compromise overall security; design comprehensive integration testing strategies; and clearly define access permissions for system components and implement strict access control policies to restrict access to critical components and services.

Organisations should evaluate deployment strategies that balance generative AI adoption with security considerations. Since AI prompts may contain sensitive business data or personal information, organisations should carefully assess risks when selecting service models—whether cloud-based solutions, on-premise setups using open-source models, or on-device implementations. A hybrid approach where deployment choices are guided by the criticality of the use case and data protection requirements may be most appropriate for many scenarios.

The adoption of interoperability standards, such as open APIs and secure data-sharing protocols, between government platforms and private-sector AI solutions can facilitate smoother integration of generative AI into e-government services and other public infrastructure. By enabling structured collaboration and technical compatibility, this approach can accelerate AI adoption, reduce duplication of efforts, and foster an open innovation ecosystem.

Regulatory alignment with international frameworks enhances trust and enables cross-boundary AI collaboration. Harmonisation with widely recognised standards, such as those established in other jurisdictions, offers predictability and confidence to global partners and investors. Clear delineation of how legal and regulatory mechanisms will function in the event of AI-related issues—including which authorities have jurisdiction over various AI-related matters—can strengthen institutional transparency and governance effectiveness.

2.1.3.4 Usage and Maintenance

Once generative AI services are officially launched, users interact with the model online or within a local network to receive desired outputs. When providing services to the public, the following risks require close attention:

- **Content Safety Risks:** Generative AI may produce content that violates legal or ethical standards based on improper user inputs.
- **Data Exposure Risks:** When generating responses based on the prompts

provided by users, generative AI might inadvertently include the user's personal data or other confidential information, or it might access data used during the Retrieval-Augmented Generation (RAG) process.

- **Copyright Risks:** Determining the copyright ownership of content created by generative AI services, if any, is not always straightforward. The domestic law governing copyright subsistence and ownership of such content may vary from place to place, necessitating careful attention and consideration to address copyright issues covering copyright ownership of outputs of generative AI services, where applicable, and prevention of copyright infringement by such outputs.
- **Credibility, Ethical, and Social Risks**: Generative AI can create realistic but false content, such as fake news or forged audiovisual material, which may contradict ethical and social values. This poses threats to social trust.

Technology Developers have the responsibility to employ technical measures to filter the input content provided by Service Users to prevent the generation of harmful content due to malicious intent. They should also take relevant measures to label the generated content to distinguish it from real content. Service Providers should provide clear operating instructions and technical documentation to explain to Service Users how to use the service correctly. They must declare the copyright ownership of the generated content to Service Users where practical, and only with the consent of the Service Users, may record their usage (logging) and store it without violating personal data regulations. Service Users should be fully aware of the potential for falsity in content generated by AI and should proactively verify the authenticity of the content.

2.1.3.5 Human Oversight

Appropriate human oversight is critical for ensuring the trust and accountability framework of generative AI systems. The degree of human oversight should be based on the impact of different stages (e.g., data collection, model training, and output generation). The greater the impact, the stronger the need for human oversight. Models can be categorised based on the level of human oversight:

- **Collaborative Generative AI Models:** For models used in less impactful decision-making scenarios, human judgment can complement AI. These systems typically require limited human oversight due to smaller data volumes.
- **Human-Dominated Models:** When collaborative models are insufficient, human-dominated models are used. These rely primarily on human decision-making and operations, with AI serving as an auxiliary tool.

2.2 Five Dimensions of Governance

To promote the effective and beneficial use of generative AI, this Guideline introduces a governance framework for generative AI based on five dimensions. Under this framework, stakeholders should clearly define the scope of their actions and accurately assess potential risks. The five dimensions are:

- Personal Data Privacy
- Intellectual Property
- Crime Prevention
- Reliability and Trustworthiness
- System Security

2.2.1 Personal Data Privacy

The rapid growth of generative AI presents unprecedented challenges to personal data privacy protection. At its core, the complexity of AI systems poses risks of personal data privacy and sensitive information leaks at all stages of development. From data collection and processing to model training, optimisation, and eventual application and service provision, even minor missteps can expose sensitive information to potential threats.

For addressing personal data privacy concerns during the AI training phase, promising approaches like Federated Learning have emerged. This technique allows models to be trained across multiple decentralised devices or servers holding local data samples, without exchanging the data itself. Instead, only model updates are shared, helping to minimise privacy exposure while still enabling effective model development.

Regarding personal data privacy, the purpose and manner of personal data collection, accuracy and duration of personal data retention, use of personal data, security of personal data, openness and transparency in relation to personal data policies and practices, and access to and correction of personal data are the key aspects in the lifecycle of generative AI development and service provision.

Ensuring the security of personal privacy and sensitive information throughout the entire lifecycle of generative AI development and service provision is thus critical. This is not only about protecting individual rights but also about maintaining public trust in generative AI technologies and fostering the healthy, sustainable development of the industry.

2.2.2 Intellectual Property

The intellectual property system of each place provides a dedicated and balanced legal framework for protecting the legitimate rights of creators and inventors, and fostering a culture of creativity and innovation. However, the rapid development of generative AI is posing unprecedented opportunities and challenges to the intellectual property regime, from aspects ranging from data collection and model training to content generation.

In the phases of data collection and model training, the utilisation of materials protected by copyright for AI training has raised widespread concerns about potential infringement and the applicable scope of copyright exceptions. This has driven academia, industry, and legal communities to engage in vigorous discussions and to explore new provisions that explicitly provide for exceptions for model training purposes, in an effort to achieve a balance between fostering innovation and safeguarding the interests of creators.

2.2.3 Crime Prevention

The evolution of generative AI presents both opportunities and challenges for crime prevention and control. While AI integration significantly advances law enforcement capabilities by transforming traditional methods into more sophisticated, data-driven approaches, its governance must extend beyond legal frameworks. A holistic approach must incorporate ethical considerations, social implications, public acceptance, and community response. The implementation of these technologies requires transparency about their capabilities and limitations to maintain public trust and ensure their deployment aligns with societal values and expectations.

Simultaneously, generative AI introduces serious governance challenges as criminals exploit these technologies. Deepfakes powered by generative AI pose significant societal risks by creating incredibly realistic fake audio and visual content that mimics individuals' appearances and voices. These deceptive materials are increasingly used to spread misinformation, manipulate public opinion, and impersonate individuals in communications and electronic transactions for fraudulent purposes, creating multifaceted threats to public safety, privacy, and trust in digital information.

2.2.4 Reliability and Trustworthiness

The credibility of generative AI refers to its ability to demonstrate stable and reliable performance, continuously and accurately outputting results that meet expectations and are trustworthy in various application scenarios. For the system of generative AI, authenticity and credibility are one of the core elements of the trustworthiness and accountability. This accountability involves the task of constructing a scientific, rigorous,

and effective mechanism and framework to ensure that developers, operators, and users of generative AI assume corresponding responsibilities for the logic of model operations, behavioural patterns, and their widespread impacts. Especially when the system generates information that is false, biased, or misleading, it is possible to clearly and accurately allocate responsibilities among the relevant entities based on this mechanism and framework.

Currently, the trustworthiness and accountability of AI-generated content are facing unprecedented and significant challenges. The technical architecture inherent in generative AI is highly complex, with a certain degree of opacity in its internal logic and algorithmic mechanisms, making the tracing and analysis of its decision-making process extremely difficult in practical applications. Issues known to affect the authenticity and credibility of generative AI include defects in algorithm design, biases or noise in training data, and unexpected anomalies encountered during operation. Since it is difficult to precisely determine the source of erroneous information, it is impossible to quickly and accurately identify the responsible entities upon discovering problems, which seriously affects the credibility and application promotion of generative AI.

2.2.5 System Security

Under the governance framework of generative AI, system security is a core element, focusing on ensuring that the system itself and data are not accessed and compromised by unauthorised individuals. However, the unique vulnerabilities that arise during the processing of sensitive information and the risks of reverse attacks faced by models are continuously increasing the security risks associated with generative AI. At the same time, data poisoning attacks are also a prominent issue. Malicious attackers deliberately tamper with training data, interfering with the model's learning process. Once data poisoning is successful, the model may make erroneous or biased decisions, which not only undermines the integrity of AI applications but also significantly diminishes user trust.

To address these challenges, Technology Developers and Service Users need to establish sophisticated monitoring and updating mechanisms and strictly implement security measures to ensure the integrity and credibility of data. In terms of preventing data poisoning, strict data verification processes should be implemented, utilising anomaly detection algorithms to identify and filter suspicious data entries, while ensuring that the training datasets are reliable and of high quality. Additionally, regularly auditing data sources and constructing secure data transmission channels can further reduce the risk of data poisoning, providing a solid guarantee for the safe and stable operation of generative AI.

2.3 Key Principles of Governance

2.3.1 Compliance with Laws and Regulations

Throughout the entire lifecycle of generative AI technology—spanning research and development, service provision, and practical application—all relevant stakeholders must uphold a strong sense of legal compliance and adhere strictly to regulatory standards. In the context of Hong Kong, every stage of generative AI operation must fully align with the specific provisions and core principles of the region's existing legal framework, leaving no room for deviation or unlawful practices. Specifically, Technology Developers must ensure that data collection for model training respects intellectual property rights and protects personal privacy, while generated content must not violate laws, public morality, intellectual property rights, or individual privacy. If the industries, regions, or countries covered by a generative AI service impose stricter requirements, Technology Developers, Service Providers, and Service Users should respect and comply with these higher standards. Additionally, Service Providers and Service Users must recognise their social responsibilities and legal obligations, particularly in avoiding the dissemination of false or harmful information.

2.3.2 Security and Transparency

In the development of generative AI, it is essential to address and resolve both model-inherent and service-derived issues to reduce their insecurity and lack of transparency. At the model level, harmful content that is illegal, violates regulations, or goes against ethical standards should be eliminated through algorithm optimisation and data governance. At the service level, Service Providers must fully disclose risks to users and enhance model security and transparency by employing technologies such as encryption and explainable AI, ensuring the reliable operation of the technology.

Taking an open-source model in the market as an example, its reasoning technology imparts a clear logical context to content generation, vividly presenting the key steps and logic in the content generation process. This makes content generation no longer a “black box” operation to some extent, allowing users to intuitively understand the basis of content generation, thereby greatly enhancing the transparency of generative AI and laying a solid foundation for the safe and reliable application of generative AI technology.

Open-source models typically offer greater transparency, allowing for public scrutiny of training methodologies, data sources, and algorithmic design. This transparency enables broader verification of safety measures and detection of potential biases or vulnerabilities by the wider community. Proprietary models, while often featuring advanced capabilities, present inherent limitations in transparency due to commercial considerations, making independent verification more challenging. Such a

comparison would provide stakeholders with valuable insights into the transparency trade-offs associated with different development approaches and their implications for governance frameworks.

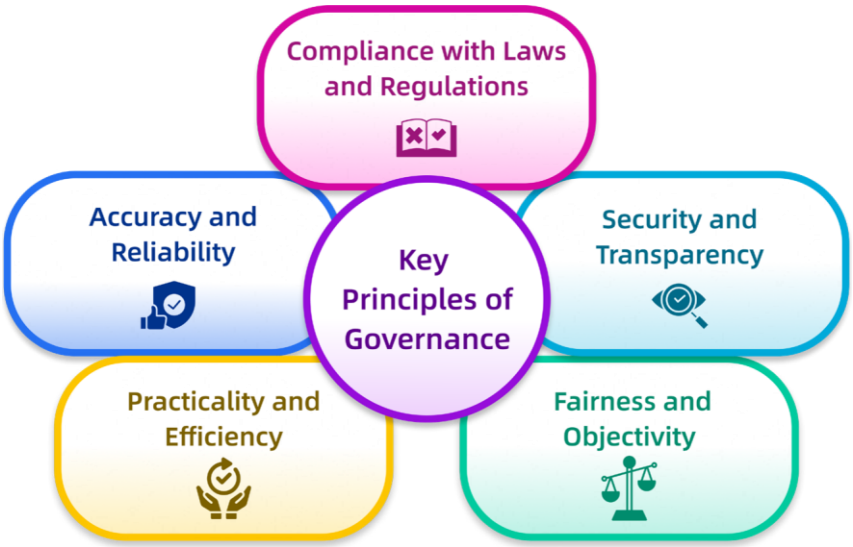

Figure 3. Key Principles of Governance

2.3.3 Accuracy and Reliability

In the research, development, and application of generative AI, careful management is required at both the model development and service stages to effectively address various risks. During model development, developers should leverage advanced technologies and scientific methodologies to minimise issues like model hallucinations that could impact future applications. For example, incorporating RAG technology allows the model to accurately retrieve relevant information from extensive external knowledge sources during content generation and seamlessly integrate this information into the process. This approach enhances the model's understanding and application of real-world knowledge, ensuring the accuracy and reliability of generated content while avoiding factual deviations caused by model hallucinations. Additionally, Service Providers should design and offer user-friendly, efficient fact-checking tools tailored to the specific service context, such as authoritative data retrieval interfaces or intelligent comparison tools, to assist users in manually verifying results. These measures significantly improve the reliability of generated outputs, ensuring the safety, stability, and compliance of generative AI services. This, in turn, supports the steady development of generative AI technologies along a controlled and trustworthy trajectory.

2.3.4 Fairness and Objectivity

Generative AI services must fully embrace the principles of diversity and universality, actively avoiding information imbalances and eliminating the risk of creating “information silos” caused by excessive concentration of certain types of content. From the initial stage of data collection to each critical step of content generation, strict controls must be implemented to eliminate model biases. During data selection, diverse information sources should be included to ensure comprehensive representation of data from varying fields and backgrounds. During model training and content generation, scientific algorithms and rigorous review mechanisms should be employed to prevent the generation of biased or discriminatory content targeting factors such as belief, political opinion, nationality, region, gender, age, ethnicity, skin colour, industry, health status, income level, or lifestyle. By providing fair, objective, and inclusive content, generative AI can help promote information equity and foster societal harmony, allowing the technology to play a positive role in advancing these goals.

2.3.5 Practicality and Efficiency

As an innovative and highly creative technology, generative AI is profoundly transforming operational models across various sectors. Developers and Service Providers bear the responsibility of continuously enhancing the utility of this technology to ensure that generative AI delivers precise, efficient, and high-quality content that aligns with user intentions. On one hand, algorithm and model architecture optimisations should be pursued to improve the accuracy and relevance of generated content, avoiding irrelevant or biased information. On the other hand, the full potential of the technology should be harnessed by exploring its applications across diverse tasks, complex scenarios, and multiple industries. By solving real-world problems and significantly improving work efficiency, generative AI can empower industrial upgrades and contribute to societal progress and the improvement of people's lives. Through innovative applications, generative AI can become a driving force for advancement at both the industry and societal levels.

3. Practical Guidelines for Technology Developers, Service Providers, and Service Users of Generative AI

Based on the above discussion, we offer the following recommendations for generative AI developers, Service Providers, and Service Users, aligned with their respective roles, rights, liabilities, and obligations as outlined in the stakeholder responsibility matrix.

Stakeholder	Definition	Responsibilities
Technology Developers	Organisations and individuals who create, train, and maintain the foundational models and algorithms that power generative AI systems.	• Ethical model development • Technical safeguard implementation • Ongoing oversight and monitoring • Data rights management
Service Providers	Entities that deploy generative AI technologies as customer-facing applications or services, acting as intermediaries between developers and end users.	• Content governance • Privacy protection • Accountability measures • User data processing
End Users	Individuals or organisations that utilise generative AI services for personal or professional purposes.	• Ethical usage • Awareness and control • Community protection • Content verification

Table 2: Role Definition Matrix

3.1 Technology Developers: Build Comprehensive Teams and Adopt Proper Working Practices

A well-structured generative AI development team and sound working practices are essential for advancing technology while ensuring security and compliance. Developers should adhere to principles that emphasise both intellectual property protection and broader considerations such as data integrity and neutrality.

- **Establish a Data Team and Assign a Leader:** A dedicated data team, led by an assigned leader, should ensure compliance with applicable laws and regulations, avoiding any infringement or violations of all other prevailing

applicable laws and regulations. In addition to managing data distribution and maintaining balance to prevent discrimination, the team must align datasets with appropriate values and factual accuracy during fine-tuning. Robust quality control mechanisms should be established to uphold these standards. Furthermore, data acquisition, storage, processing, and transmission must comply with relevant laws and regulations, including those governing personal data privacy and cybersecurity. Stringent security measures are vital to prevent data breaches and misuse. By following these practices, the data team can support responsible and ethical AI development.

- **Establish an Algorithm Engineering Team:** This team should prioritise security and trustworthiness in algorithm design and optimisation, ensuring safety during pre-training, fine-tuning, and inference stages. They should promptly identify and fix vulnerabilities, enhance content explainability through inference techniques, and strengthen system-level guarantees for the reliability of generated outputs. Techniques such as RAG and knowledge bases should be employed to ensure the timeliness and accuracy of the generated content.
- **Establish a Quality Control Team:** Comprehensive testing is a key method to ensure that AI applications are safely deployed and function properly before deployment. Establishing formal testing requirements to verify the model's vulnerability to security threats, such as its ability to resist adversarial attacks and sensitivity to data breaches, is crucial. Involving Service Users in the testing process can help identify security issues from the perspective of the users. Standardised testing criteria facilitate effective comparison between candidate models. To this end, the system must have concrete performance metrics. Developers should devise clear, measurable implementation paths based on performance metrics (such as precision, recall, and accuracy). Control measures or means for manual supervision and intervention in the model's operation should be established. Additionally, implementing an anomaly-based reporting system to highlight performance deviations and regular assessments to maintain the model's accuracy and effectiveness are essential. Lastly, Technology Developers must regularly review test results to further ensure the accuracy of the test results, guaranteeing the stability and reliability of the model and product, and ensuring they can operate normally in various application scenarios, outputting reliable results.
- **Establish Operational Principles:** Transparency and explainability are crucial, requiring developers to disclose training data sources (where feasible), model architectures, and evaluation metrics. Organisations should establish policies on when to accept AI-generated content, such as requiring users to double-check AI-generated materials, verify references, and ensure correctness before

usage. Data quality must be prioritised, with training data that is high-quality, diverse, representative, and regularly updated. AI systems should provide transparent explanations for outputs and incorporate mechanisms for source verification, fact-checking, and validation. Continuous monitoring, including regular audits, is essential to identify and address errors, biases, or inaccuracies. User feedback should be actively encouraged to improve system reliability, while domain expertise is critical for developers and content creators. Accountability mechanisms must be established to address dissemination of misinformation, along with industry-wide standards for consistency and accuracy. Collaboration with academia and research institutions, both locally and internationally, is vital to stay updated on advancements and leverage expertise. Finally, industry leaders and regulatory bodies should work together to develop common guidelines covering data quality, transparency, accountability, and bias mitigation.

- **Establish a Compliance Team:** Regular compliance reviews and assessments of the models and products developed are essential to prevent negative impacts on social order, public safety, and ethical standards. Establishing a comprehensive documentation system encourages Technology Developers to adhere to the principle of transparency by disclosing the technical principles and usage rules, thereby allowing Service Users and regulatory agencies to understand and supervise the application of the technology, building trust, and reducing the risk of misuse.
- **Technology Developers and Service Providers should follow higher standards.** High-risk AI-generated content, such as deepfakes, ID document images, and financial materials, should include irremovable watermarks or embedded codes to ensure traceability and accountability. Transparency is crucial—AI models must be trained on verified, reliable sources, with Service Providers integrating trusted references and disclosing information sources. AI systems should alert users when outputs cannot be verified, preventing the spread of unverified or false content. Responsibility for accuracy must rest with developers and providers, not users.
- **Independent evaluation mechanisms** should apply not only to Service Providers but also to Technology Developers from the development stage. Developers should undergo audits covering risks like content accuracy, bias, harmful outputs, and privacy compliance. As major stakeholders, developers must bear greater responsibility for AI safety. Implementing these measures will enhance accountability, transparency, and the overall effectiveness of the Guideline in preventing AI misuse.

3.2 Service Providers: Build Responsible Service Frameworks and Service Development Processes

3.2.1 Establish a Responsible Generative AI Service Framework

Generative AI Service Providers must identify specific business or opportunities that can bring significant value, and determine the priority of service provision based on service feasibility, alignment with strategic objectives, and potential impact, establishing a responsible framework for generative AI services. We propose below aspects to consider when establishing such a framework:

- **Ensure Service Compliance:** Service Providers should select and use base models that comply with the relevant laws and regulations and social ethical standards of Hong Kong. Service Providers are responsible for ensuring that their service systems do not output illegal, non-compliant, or inappropriate content. They should establish mechanisms to enhance the traceability and auditability of the system, effectively reducing the risk of inputting malicious data; label content such as generated images and videos; and strengthen risk notification for content that is unsuitable for output or contains biases, as well as for services used by special groups such as minors.

- **Ensure Data Security:** Service Providers must comply with relevant data protection regulations such as the Personal Data (Privacy) Ordinance (PDPO) (Cap. 486) when collecting, processing, using, storing, retaining and deleting of user data including personal data. They must fully protect the privacy rights of Service Users, avoiding excessive collection, misuse, or disclosure of user data. At the same time, they should strengthen the encryption and desensitisation of sensitive data to ensure the security and privacy of data during its transfer. When necessary, they should work closely with Technology Developers and conduct data security surveys among Service Users to promptly identify and fix security vulnerabilities.

  Service Providers must take specific measures to effectively manage personal data privacy. First, they should standardise cross-industry data protection agreements to ensure data integrity and enable international data flow cooperation. They may make reference to various guidelines, such as the “Guidance on Personal Data Protection in Cross-border Data Transfer”, published by the Office of the Privacy Commissioner for Personal Data (PCPD). Second, they must implement robust technical safeguards, such as advanced anonymisation techniques and enhanced encryption technologies, to protect data and prevent re-identification (i.e., removing or encrypting personally

identifiable information from data sets). Third, compliance with local and international data protection laws (such as the PDPO and European Union General Data Protection Regulation (GDPR)), as appropriate, is essential, along with obtaining explicit consent from Service Users and safeguarding their rights and interests. Finally, Service Providers should conduct continuous monitoring and evaluation to adapt to evolving data security threats, ensuring that their measures remain effective and up-to-date.

Service Providers should introduce opt-out mechanism that empowers users to avoid sharing their data for future model training. Besides, Service Providers should also establish a mechanism to accept and review feedback and complaints from Service Users about the use of personal data and take immediate actions to correct or remove the data concerned.

- **Ensuring System Security:** Service Providers must continuously monitor and assess the security of their systems, develop preventive measures and emergency response plans for potential system and data attacks; when there are significant changes in the system's functions or operation, a reassessment should be conducted to identify and mitigate new risks.
- **Ensuring System Credibility:** Service Providers need to adhere to the principle of transparency by disclosing the technical principles behind their services, which allows Service Users and regulatory bodies to understand and supervise the services provided. They should offer detailed service descriptions and usage guides so that Service Users can correctly understand and utilise the services. Service Providers should establish a clear framework for trustworthiness and regular review cycles, engage in monitoring with stakeholders to foster a culture of compliance, and enhance the reliability of the services. Additionally, they should hire independent auditors to regularly audit the quality, safety, and compliance of the services; and engage ethics policy experts to strengthen the alignment of service standards with ethical norms and policies.

3.2.2 Responsible Service Development Processes

- **Service Procurement:** When Service Users procure generative AI services, economic and security factors are often key considerations. On the one hand, the price needs to be within the budget and offer reasonable cost-effectiveness; on the other hand, security indicators such as data security, privacy protection, and system stability directly impact the quality and user experience of the services. Therefore, Service Providers should establish clear financial agreements and service security agreements to ensure the orderly conduct of services.

Additionally, an independent evaluation mechanism and proper documentation should be established. Independent evaluations can objectively and neutrally review services and uncover potential issues, while comprehensive documentation records every aspect of the services in detail, enhancing transparency and providing strong support for the construction of a trustworthiness and accountability framework.

- **Risk Assessment:** Service Providers of generative AI services should conduct comprehensive service risk assessments at various stages of service development. The assessment includes: at the data level, reviewing the quality of training data, identifying issues such as missing data, incorrect annotations, and duplication, analysing data biases, ensuring data security, and preventing privacy breaches, tampering, and misuse. At the algorithm and model level, checking for logical and security vulnerabilities in algorithms, evaluating the explainability of model decision-making processes, and assessing stability under different input conditions. At the application scenario level, ensuring compliance with relevant laws and regulations as well as industry standards for each scenario, analysing service deficiencies such as misinformation and negative impacts on user experience, and considering social risks. At the operation and management level, assessing the infrastructure and operational capabilities of the service, focusing on personnel operation and ethical risks, and evaluating the reliability and stability of third-party suppliers. At the same time, ensuring the independence of the assessment and establishing a proper documentation mechanism are crucial. Independent assessments can objectively and neutrally review services and uncover potential issues, while comprehensive documentation ensures transparency throughout the entire service process, providing strong support for the construction of a trustworthiness and accountability framework.

- **Pilot Projects:** Before rolling out services on a large scale, Service Providers should carry out small-scale pilot projects to verify the feasibility of the services in specific business scenarios under clear objectives and scope, consider the impact on business efficiency and costs, and define business processes, user groups, and data boundaries. Such pilots select appropriate models based on objectives and data characteristics, assess performance, scalability, and costs, and fine-tune models.

- **Service Maintenance:** Service maintenance encompasses the continuous improvement of service quality, security, and user satisfaction. Service Providers must prioritise transparent communication with Technology Developers and Service Users to build trust and ensure compliance. This includes disclosing the rules, purposes, benefits, and limitations of generative

AI services, while respecting the rights of data subjects and users, and informing individuals about how their personal data is used. Service Providers should strive to enhance the transparency of generative AI services, provide insights into the decision-making process, and establish feedback channels to promote an inclusive and responsible generative AI service process using clear and understandable language, thereby improving the user experience.

3.3 Service Users: Proactive Stewards of Benevolent AI

Service Users of generative AI assume a critical role and bear significant responsibilities when utilising these services. It is essential to heighten awareness of the technological and service security risks to consciously contribute to a compliant and secure AI ecosystem. Here are our recommendations for Service Users of generative AI services:

- **Legal and Regulated Use:** Service Users should use generative AI services in a manner consistent with the guidance and requirements provided by the employer, Service Providers and regulatory authorities (such as PCPD), refraining from using the services for any illegal, non-compliant, or inappropriate purposes. When generating content through generative AI services, users should take into account the laws and regulations of Hong Kong, including but not limited to laws relating to copyright, privacy, and anti-discrimination; should the generated content contain potential or obvious violations or inaccuracies, Service Users should promptly report to Service Providers to remove the corresponding and prevent the dissemination of illegal, non-compliant, or inappropriate content, which is a fundamental aspect of proper generative AI usage. Service Users should not use AI services to damage the reputation, legitimate rights and interests of others. When utilising generative AI services at work, employees should also adhere to organisation policies regarding permitted devices, tools, use, information input to the generative AI and proper reporting procedures, if applicable.
- **Maintain Independent Discretion:** Generative AI serves as our tool and assistant, not a replacement for us. While the content generated can inform our work and daily life, it should not be adopted without human verification and judgment. Service Users should be aware that AI-generated content may include misleading, false, or inaccurate information. Responsible users should possess a multifaceted awareness and knowledge capability in law, ethics, and risk management, so as to validate and review generated content and make independent information judgments. If necessary, submit requests to Service Providers to rectify or remove any inaccurate data in the generated content.

- **Establish Organisation Policies and Guidelines:** Organisations adopting generative AI services should developing their internal policies or guidelines, which cover[1]:
  - the permitted tools (include publicly available and internally developed generative AI tools or applications);
  - the permissible use (for example, drafting, summarising information and/or creating textual, audio and/or visual content);
  - the policy applicability;
  - the permissible types and amounts of input information;
  - the permissible use of output information;
  - the permissible storage of output information;
  - the compliance with other relevant internal policies to ensure that the policy on the use of generative AI is aligned with organisation's other relevant internal policies;
  - specify that employees shall not use generative AI tools for unlawful or harmful activities;
  - emphasise the importance of employees acting as human reviewers to ensure that generated output aligns with the ethical values and standards of the organisation;
  - the permitted devices;
  - the permitted users;
  - robust user credentials (for example, using strong, unique passwords and multi-factor authentication);
  - security settings (for example, disabling chat history or data sharing functions);
  - response to AI incident and data breach incident (for example, reporting data breaches, unauthorised data input, or unlawful output results); and
  - specify the possible consequences of employees' violations of policies or guidelines.
- **Understand Responsibilities and Obligations:** Before engaging with any generative AI services, Service Users should thoroughly read and familiarise themselves with the terms of use of the relevant platforms or software to understand their responsibilities and obligations. These terms often encompass themes such as privacy, security, ethical standards, and legal compliance. For instance, it is explicitly stipulated that Service Users may not instruct AI to

[1] https://www.pcpd.org.hk/english/resources_centre/publications/files/guidelines_ai_employees.pdf

generate content that contains hate speech, bias, discrimination, defamation, or other immoral and illegal content, as well as protections for users' rights, such as restrictions on the sharing of personal data to prevent Service Providers from recording and disseminating users' personal information beyond the scope of consent. Personal data should be anonymised or cleansed before being inputted into generative AI services.

- **Citation and Attribution:** To ensure transparency, accountability, and safety, Service Users should explicitly indicate by watermarked, labelled, metadata, digital signatures, etc., whether generative AI has been involved in content generation or decision-making, and must bear the responsibility for its ethical and legal implications.
- **Privacy Protection:** Service Users should familiarise themselves with the privacy policies of generative AI services to understand their specific practices regarding data collection, use and sharing. It is recommended to choose services that do not use shared data for training generative AI and to avoid transmitting sensitive personal information, while adopting pseudonymisation or anonymisation methods to protect privacy. Regularly reviewing and deleting data within generative AI services, and submitting correction or deletion requests if inaccuracies are found in the generated content, is also important.
- **Prudent Dissemination:** Any content generated by generative AI systems and further disseminated, which may impact society, economy, or culture, ultimately holds the content disseminator responsible. This means that Service Users need to assess and take responsibility for the potential misdirection or negative consequences of the content. Service Users should proactively verify the authenticity, legality, and appropriateness of generated content and seek professional advice or conduct secondary reviews when necessary to reduce potential risks. Additionally, if Service Users intend to publish AI-generated content, they should disclose its source when making it public. Disclosure of the source is especially important when the content involves commercial use or mass dissemination. This level of transparency not only helps to increase public trust in generative AI but also ensures the legality of one's own content creation and dissemination.
- **Respect for Intellectual Property Rights:** To maintain the generative AI ecosystem and encourage legal creation by Service Users, it is essential to consciously respect intellectual property rights. This includes avoid using generated content that constitutes the whole or substantial copying of copyright works so as to prevent copyright disputes. Specific practices may include searching and assessing whether the generated content constitutes

infringement of copyright, trade mark or patent rights. If the generated content is found to have infringed others' intellectual property rights, it should be deleted or modified in a timely manner according to the administrative and/or legal measures (e.g., terms of use, take-down notice and/or court order, where applicable).

Acknowledgement

Institution:

The Hong Kong University of Science and Technology

Hong Kong Generative AI Research and Development Center

Lead Authors:

Sirui Han, Yike Guo, Hongying Huang

Professor Sirui Han is an Assistant Professor at The Hong Kong University of Science and Technology (HKUST) and an RGC-Fulbright Research Scholar, as well as the Head of Large Language Model Division at the Hong Kong Generative AI Research and Development Center (HKGAI).

Professor Yike Guo is the Provost and Chair Professor at HKUST, as well as the Director of the HKGAI.

Dr Hongying Huang is a Special Advisor at HKUST and the Chief Operating Officer of the HKGAI.

Appendix

With the advancement of generative AI technology, countries and industries have begun to enhance governance and oversight to ensure the legality and compliance of the technology. In addition to providing specific action guidelines for Technology Developers, Service Providers, and Service Users, the Guideline compiles the relevant requirements for the governance of generative AI from various countries and industries, offering relevant policy information to help them conduct generative AI activities more safely and compliantly on a global scale. Furthermore, the governance of the AI industry in Hong Kong has distinct characteristics. This Guideline summarises the principles of trustworthy generative AI industry applications in Hong Kong to provide a reference for relevant parties, which can be found in the appendix.

1 Governance Requirements for Generative AI in Specific Countries and Regions

1.1 The Mainland

The Mainland has made significant strides in the field of generative AI, emerging as a pivotal player in the global AI development landscape. Recognising AI governance as a matter of global significance that affects the destiny of all humanity, the Mainland has issued the “Global AI Governance Initiative”. This initiative calls on nations to uphold a collective, comprehensive, cooperative, and sustainable security perspective. It emphasises the importance of balancing development with security, fostering consensus through dialogue and cooperation, and establishing an open, fair, and effective governance mechanism. The aim is to harness AI technology for the betterment of humanity and to advance the building of a community with a shared future for mankind.

The Mainland has been consistently advancing the governance and regulation of AI. It has previously set targeted management requirements for technologies such as recommendation algorithms and deep synthesis services. In the realm of generative AI, the Cyberspace Administration of China, in collaboration with relevant departments, has promulgated the “Interim Measures for the Management of Generative AI Services” (“the Measures”) under the legal framework of the “Cybersecurity Law”, “Data Security Law” and “Personal Information Protection Law”. The Measures, which officially came into effect in August 2023, serve as the core regulatory document for the governance of generative AI in the Mainland at present. The Measures underscore the nation's commitment to a principle that prioritises development and security, encourages innovation, and integrates law-based governance. Effective measures are taken to

stimulate the innovative development of generative AI and to implement inclusive, prudent, and classified supervision over generative AI services.

1.2 Other Major Countries and Regions

Currently, major countries around the world are actively developing generative AI, but there are differences in understanding of AI safety, governance concepts, and legal and regulatory requirements. Service Users or Service Providers planning to expand into overseas markets should consider understanding the local governance principles and specific requirements for AI, such as:

- **United States:** The United States has been actively using means such as export controls to solidify its leading position in this transformative technology field. In terms of AI governance, the United States adopts a policy guidance and corporate self-regulation approach, with key AI companies committing to self-regulation of AI systems. The United States exhibits a decentralised and localised characteristic in AI governance through legislative means; there is currently no unified federal law for the AI field, nor is there a unified regulatory agency. Some industries such as finance and healthcare, along with relevant regulatory departments, have introduced their own self-regulation standards and regulations. State-level AI legislation also shows significant differences, with states like California and New York, which are more active in AI research and application, having passed their own AI bills with different focuses. To comply with United States AI laws and regulations, it is essential to first clarify specific industry and state background information and further understand the requirements in those areas.
- **European Union:** The European Union (EU) has one of the most significant AI markets globally and primarily ensures the regulated application and development of AI technology services through legislation, ethical frameworks, and technical standards. The EU has issued a series of normative documents, including the “Ethics Guidelines for Trustworthy AI”, the “General Data Protection Regulation” (GDPR), and the “Artificial Intelligence Act” (“the Act”). The Act, which directly targets AI systems, officially came into effect in August 2024. The Act aims to improve the functioning of the internal market, promote human-centric, trustworthy AI applications, and prevent AI systems from posing risks to health, safety, fundamental rights, including democracy, the rule of law, and environmental protection. The Act mainly adopts a risk-based and role-based regulatory approach, classifying AI systems into unacceptable risk, high risk, limited risk, and low risk. It requires the prohibition of AI systems with unacceptable risks, strict regulation of AI systems with high risks, appropriate relaxation for AI systems with limited risks, and no regulation for AI systems with

low risks. The Act also defines six AI system participant roles, including providers, deployers, distributors, importers, authorised representatives, and product manufacturers, and specifies the obligations each must undertake. It is important to note that the Act has a broad scope of application; in addition to applying to businesses and individuals within the EU, it also applies to AI technology service providers outside the EU that provide services to users within the EU.

- **United Kingdom:** The United Kingdom (UK) closely monitors the development and safety of AI and has established the world's first AI safety research institute. In terms of AI governance, the UK has adopted a more flexible approach, with AI regulation not introducing new specialised legislation but mainly relying on existing regulatory bodies to issue guidance and application standards within their functional scope. To enhance businesses' and the public's confidence in using AI, in March 2023, the UK Department for Business, Energy & Industrial Strategy published the white paper “A Pro-Innovation Approach for AI Regulation”, proposing its AI governance approach based on five principles of safety, security, robustness, appropriate transparency, and explainability, providing clearer and more consistent regulatory guidelines for the industry.
- **Singapore:** Singapore believes that AI has significant transformative potential but also comes with risks. In terms of AI governance, Singapore mainly adopts the method of incorporating the regulation of AI into various industry regulatory departments, implementing governance by issuing non-binding guidelines and recommendations. Among them, the Infocomm Media Development Authority (IMDA), responsible for regulating the communications industry, and the Personal Data Protection Commission (PDPC), responsible for personal data protection, are the two most active departments in AI governance, releasing the first and second versions of the “AI Governance Framework” in 2019 and 2020, respectively. To address the challenges brought by the rapid development of generative AI, IMDA released the “Generative AI Governance Framework”, aiming to propose a systematic and balanced approach to solve generative AI issues while promoting innovation, considering nine aspects including responsibility, data, trustworthy development and deployment, incident reporting, testing and assurance, security, content source, safety-aligned research and development, and AI for public good, to comprehensively consider building a trustworthy ecosystem.

2 Key Governance Areas of Generative AI in Hong Kong

After reviewing global AI governance frameworks, it has been observed that many jurisdictions employ non-binding frameworks to guide the development and use of AI systems. By issuing practical guidelines under such frameworks, governments can promote the responsible use of AI without imposing strict regulations that might hinder innovation.

- For instance, Singapore's “Model AI Governance Framework” provides detailed guidelines on how to ensure the ethical use of AI, focusing on internal governance, risk management, and best operational practices.
- Japan's “Social Principles of Human-Centric AI” prioritise human rights, data protection, and social acceptance, encouraging stakeholder collaboration to mitigate risks while harnessing the potential of AI.

These global frameworks demonstrate that non-binding guidelines can effectively guide the ethical development and deployment of AI technology, while allowing for flexibility and innovation.

As for the case of Hong Kong, the HKSAR Government always recognises the importance of addressing ethical considerations in the implementation of AI projects and services. The HKSAR Government formulated the “Ethical Artificial Intelligence Framework” (“the Framework”) in 2021 to provide a set of practical guidance when implementing projects that involve the use of AI technology. Since then, the HKSAR Government has been keeping on updating the Framework with reference to the latest AI development. Moreover, in response to the needs of various industries, the HKSAR Government has formulated corresponding policy statements. For instance, in October 2024, the Financial Services and the Treasury Bureau (FSTB) issued the “Policy Statement on Responsible Application of Artificial Intelligence in the Financial Market” to set out the HKSAR Government’s policy stance and approach towards the responsible application of AI in the financial market.

Hong Kong places significant emphasis on safeguarding personal data privacy in the application of AI technology. PCPD published in 2021 the “Guidance on the Ethical Development and Use of Artificial Intelligence” and “Artificial Intelligence: Model Personal Data Protection Framework” in June 2024, with an aim to help organisations understand and comply with the relevant personal data privacy protection requirements under the Personal Data (Privacy) Ordinance (PDPO) (Cap. 486) when developing, customising and using AI. PCPD published in March 2025 the “Checklist on Guidelines for the Use of Generative AI by Employees” to assist organisations in developing internal policies or guidelines on the use of generative AI by employees at work while complying with the requirements of the PDPO.

In addition, to enhance the intellectual property regime, the HKSAR Government launched a public consultation on 8 July 2024 on the enhancement of the Copyright Ordinance (Cap. 528) regarding the protection for AI technology development.

A number of regulatory authorities and public bodies also published several sector-specific guidelines. These regulatory authorities and public bodies oversee various industries, applying specific regulations, guidelines, and codes of conduct to regulate and govern various sectors. Several principles and guidelines by these regulatory authorities and public bodies have been established to facilitate AI governance in specific areas, balancing innovation with risk management and ethical considerations. For example, the Hong Kong Monetary Authority (HKMA) has published several key documents, including “High-level Principles on Artificial Intelligence”, “Consumer Protection in respect of Use of Generative Artificial Intelligence” and “Use of Artificial Intelligence for Monitoring of Suspicious Activities”. The HKMA and the Hong Kong Cyberport Management Company Limited collaborated to launch the GenAI Sandbox. This initiative provides banks with a risk-controlled environment to develop and test AI solutions tailored to real-world banking scenarios, driving innovation in the banking sector. The Hong Kong Judiciary also issued the “Guidelines on the Use of Generative Artificial Intelligence for Judges and Judicial Officers and Support Staff of the Hong Kong Judiciary”.

2.1 Hong Kong Governance Framework for Generative AI: Hong Kong Features

When formulating or adjusting regulatory measures for generative AI, Hong Kong should adhere to a pragmatic balance strategy. It is essential to address the risks and concerns associated with technology in areas such as personal data, security, and fairness, while also avoiding excessive hindrance to innovation. Hong Kong has long had a rigorous and comprehensive legal system, which lays a solid foundation for the governance of emerging technologies, including generative AI. Under the existing framework for AI governance, some core regulatory issues are already covered by current laws, such as the Personal Data (Privacy) Ordinance (Cap. 486), which is sufficient to address privacy disputes that may arise from the use of personal data in AI systems. Further to the consultation regarding the enhancement of the Copyright Ordinance (Cap. 528) for the protection for AI technology development, it proposed specific text and data mining (TDM) exception to allow reasonable use of copyright works for computational data analysis and processing.

However, some AI applications, although posing potential risks, can practically be assessed and tested through industry self-regulation or phased regulatory sandboxes; as long as necessary privacy protection and security standards are observed, excessive regulation is not necessarily required. This gradual institutional arrangement can more

effectively maintain Hong Kong's attractiveness to international investment and innovation and consolidate its position as a regional technology hub.

2.2 Hong Kong's Generative AI Governance Policy Framework

The rise of Generative AI has brought unprecedented synergistic effects to Hong Kong's innovation and technology ecosystem, spawning various new opportunities in finance, healthcare, education, smart city domains, etc. However, as AI systems become increasingly prevalent in social and economic activities, concerns over personal data privacy, intellectual property rights, cybersecurity, and AI biases and ethical risks are also escalating. To maintain a consistent emphasis on personal data, security, and fairness while encouraging innovative development, Hong Kong must maintain a flexible and adaptive policy framework, fostering an environment that supports innovation and compliance in a balanced manner.

The Guideline references widely recognised international AI governance models and incorporates Hong Kong's legal and industrial characteristics, continuously optimising regulatory strategies with the principles of “application-oriented” and “risk-based” approaches. This practice corresponds with non-binding guidelines in jurisdictions like Singapore and Japan, emphasising that governments and industries should retain sufficient flexibility and provide necessary guidance when facing rapidly evolving technologies. This framework focuses on:

2.2.1 Application-Oriented Principles of Trustworthy AI in Hong Kong

To implement principles related to transparency and explainability in trustworthy AI, clear guidelines for AI system documentation should be established, enabling Technology Developers, Service Users, and regulatory authorities — including relevant Hong Kong Government agencies, international regulators, and industry associations — to fully understand system design, decision-making processes, data sources, and intended uses.

Additionally, promoting the application of explainable AI technologies can effectively assist the citizens of Hong Kong and stakeholders without a technical background in better understanding the basis and logic behind AI outputs. This includes selecting or developing technical tools that can explain AI decisions in a user-friendly manner, aiming to make the general public more comfortable in accepting and using these emerging technologies.

At the same time, in the face of the increasingly severe threats posed by deepfakes and the misuse of AI-generated content, Hong Kong is actively formulating comprehensive governance strategies. These strategies aim to promote responsible AI innovation while protecting individual rights and public trust. Specific measures include:

- Promoting the ethical development of AI, emphasising transparency and accountability, and mandating clear labelling and traceability of AI-generated content to further support this objective.
- Raising public awareness by educating the public on identifying and understanding risks including deepfakes, combating false information to maintain social safety, and participating in international cooperation to address the cross-boundary nature of such challenges.

By implementing the aforementioned measures, Hong Kong can continue to promote AI innovation on a foundation of trustworthiness, reliability, and compliance, maintaining a competitive edge in the global market. By raising awareness of threats such as data poisoning and implementing preventive measures, Hong Kong aims to create a secure and reliable AI environment that supports innovation while mitigating potential risks.

2.2.2 Industry-Oriented Trustworthy AI Principles in Hong Kong

In the process of promoting the development and application of AI, various industries in Hong Kong must balance ethical norms with industry-specific characteristics to achieve the goal of fostering innovation while ensuring effective governance. Some ethical principles are non-negotiable, such as ensuring personal safety, protecting personal data, and maintaining system reliability. However, other principles can adopt an application-driven approach, allowing for flexible management based on industry-specific needs. For example, the financial services sector should emphasise system transparency to maintain fairness and user trust; the healthcare sector should prioritise patient privacy protection; autonomous vehicles need to ensure the establishment of trust and model safety; and the education technology (EdTech) sector should ensure equitable access, preventing learning outcomes from being affected by algorithmic bias.

Hong Kong has a well-established and mature industry regulatory framework and public institutions, with specific laws, guidelines, and codes of conduct for different industries. As AI adoption becomes more prevalent, industries must establish additional requirements and guidelines based on their specific services and risk points. In particular, when applying generative AI, strict compliance with the Personal Data (Privacy) Ordinance (PDPO) (Cap. 486) and relevant data security and protection measures must be ensured, while integrating industry-specific needs to facilitate stable development. PCPD has published an information pamphlet on “10 Tips for Users of AI Chatbots” in September 2023 to promote safe and responsible use of AI chatbots, which applies across all industries and sectors. Below are key recommendations for various industries when using generative AI:

- **Finance:** The financial sector should strengthen fairness in the use of generative AI. When providing recommendation or decision-support services, all potential candidates should have an equal opportunity to be recommended, and necessary mechanisms should be in place to prevent human manipulation. Where possible, the financial sector should consider restricting interference with recommendation weights through manual settings, model training interventions, or other means. Where applicable, financial institutions like banks may need to customise models to align with specific user requirements. The financial sector should consider disclosure of information as much as practical and user choice should be provided to help users understand the working mechanisms, effects, and potential negative impacts of generative AI. If possible, users should be able to opt into using generative AI knowingly, and they should have the ability to terminate related services at anytime. Instead of opt-in, the HKMA circular "Consumer Protection in respect of Use of Generative Artificial Intelligence" dated 19 August 2024 states that customers should be provided with the option to opt out of using GenAI and request human intervention on GenAI-generated decision at their discretion as far as practicable, during the early stage of deploying customer-facing GenAI applications. Where an opt-out option cannot be provided for some reasons, banks should provide channels for customers to request for review of the GenAI-generated decisions.
- **Healthcare:** When using generative AI to assist in diagnosis, extreme caution should be exercised. Users must be explicitly informed that generated content may contain errors or fabrications, and such content should not be used directly as diagnostic reports but should be reviewed by licensed professionals as a reference. Personal data protection is crucial, and generative AI should adhere to the principle of data minimisation when collecting sensitive information such as identity details, biometric data, medical conditions, and patient histories. The purpose, usage, and processing methods of collected data must be clearly communicated to individuals, and explicit consent must be obtained before collection. Measures should be implemented to prevent data leaks during collection, transmission, processing, and storage. Additionally, data collected for healthcare purposes should not be repurposed for insurance, job recommendations, or other industries.
- **Legal:** In the legal industry, the accuracy and reliability of generative AI outputs must be ensured, and generated content should include citations that trace back to the original legal texts. AI-generated content should not be used directly as legal documents but should serve as a reference after review by legal professionals. To protect personal data, sensitive legal cases involving trade secrets or private information should not be processed using public AI services

that lack security and confidentiality guarantees.

- **Education:** The education sector should regulate the use and scope of generative AI rather than outright banning students from using it. However, students should obtain teacher approval before using AI in their coursework, and AI-generated content should be clearly identifiable to prevent misuse that violates academic integrity. When teachers use generative AI in teaching, they must ensure that the generated content is truthful, accurate, and consistent in both textual and visual representation. If AI is used for grading assignments and exams, final results should always be reviewed by human educators.
- **Journalism:** When using generative AI for news gathering, the principles of truthfulness, objectivity, and impartiality must be upheld. A diverse range of information sources should be used as input, and both technical and procedural safeguards should be implemented to minimise factual inaccuracies, misleading content, or distortions caused by model hallucinations. It is recommended that AI-generated news content include source attributions to facilitate manual verification. Generated content must undergo fact-checking and full editorial review before being published. Journalistic ethics must be strictly adhered to, and generative AI should not be used to create fabricated text, images, or audio-visual content that misrepresents facts or infiltrates news reports in any form.
- **Tourism:** When using generative AI to process customer personal data or preferences, service providers must clearly inform users of the data usage purpose and obtain their consent. In applications such as travel recommendations, hotel bookings, or AI-powered customer service, generative AI must ensure fair and non-discriminatory treatment of different customer segments. Regular reviews should be conducted to identify any false or misleading information within generated content. When using AI for customer service enhancement or marketing strategies, a balance between privacy protection and targeting accuracy must be maintained to prevent excessive data collection and misuse.
- **Retail:** Retail businesses using generative AI for product recommendations, dynamic pricing, or customer service must ensure that algorithmic outcomes remain fair and transparent across different customer segments and geographic locations, maintaining market fairness. When collecting customer preferences and purchasing behaviour, businesses must comply with data protection and privacy regulations while conducting personalised marketing and promotions. To address customer concerns or confusion regarding generative AI, human support and real-time response mechanisms should be available to safeguard

consumer rights.

- **Logistics:** In transportation scheduling and intelligent route planning, generative AI should rely on reliable and up-to-date traffic and geographic data to minimise bias risks. In logistics processes such as shipping, warehousing, and delivery, any handling of personal addresses or consumer habits must incorporate appropriate encryption and access control measures to prevent data breaches. If industrial robots, such as automated sorting arms, are used in conjunction with generative AI, regular security and stability assessments should be conducted to prevent collisions or accidents.
- **Industry:** In industrial process monitoring and production line optimisation, AI models should be trained with high-quality, rigorously validated datasets to ensure accurate system judgments and predictions. If predictive maintenance or automated fault diagnosis functions are introduced, AI-generated results must be reviewed by supervisory engineers or quality control personnel. Strong security measures must be in place, particularly when handling confidential formulas/know-hows, or other trade secrets to prevent technology/confidential information leaks.

These industry-specific guidelines aim to ensure the ethical and responsible use of generative AI across different sectors in Hong Kong, balancing innovation with governance to support sustainable AI development.

3 Reference

- Ethical AI Framework
- Checklist on Guidelines for the Use of Generative AI by Employees by Office of the Privacy Commissioner for Personal Data (PCPD)
- 生成式人工智能服務管理暫行辦法

  Interim Measures for the Management of Generative AI Services
- 人工智能生成合成內容標識辦法

  Measures for Labeling of AI-Generated Synthetic Content
- GB/T 45652-2025

  網絡安全技術 生成式人工智能預訓練和優化訓練數據安全規範

  Cybersecurity technology—Security specification for generative artificial intelligence pre-training and fine-tuning data

- GB/T 45674-2025

  網絡安全技術 生成式人工智能數據標注安全規範

  Cybersecurity technology—Generative artificial intelligence data annotation security specification

- GB/T 45654-2025

  網絡安全技術 生成式人工智能服務安全基本要求

  Cybersecurity technology—Basic security requirements for generative artificial intelligence service