GENERATIVE ARTIFICIAL INTELLIGENCE GUIDELINES

PUBLIC

Version 1

January 2024

SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Contents

Introduction

Generative Artificial Intelligence (GenAI) has gathered significant attention in the media in recent months. Despite being relatively new, both individuals and businesses have already begun integrating it into their daily lives and activities. While the advantages of GenAI are readily apparent, it is also crucial to acknowledge the substantial risks and challenges associated with potential misuse.

The Saudi Data and Artificial Intelligence Authority (SDAIA) is comprehensively assessing opportunities and related risks in order to promote investment, adoption and responsible use to gain benefits of GenAI technologies while reducing related risks. GenAI tools can improve efficiency and create new products to the benefit of citizens. We are committed to facilitate the wide spreading adoption of the technology in KSA to achieve our national goals.

1. Definitions

For the purposes of this document, the following terms and expressions, wherever they appear in this document, shall have the meanings indicated next to each of them unless the context requires otherwise:

Generative Artificial Intelligence

The generative model is a machine learning model that can create new examples similar to the training dataset. This model is also a sub-part of artificial intelligence that can create new content (including text, images, sounds, icons, videos, etc.) and works by interpreting commands given by users. Generative artificial intelligence can perform tasks that require human cognitive abilities, including responding to and formulating verbal or written commands, “learning,” and “problem-solving.”

Users

Any natural or legal person that consumes or makes use of the goods or services produced by GenAI systems. This is applicable to all stakeholders including companies, NGOs and individuals.

Developers

Any natural or legal person that develops GenAI systems to achieve certain goals. This includes AI developers, data scientists and researchers.

2. The Scope

This document serves as guidance for the public including developers and users of GenAI in the Kingdom of Saudi Arabia.

This guide aims to comprehensively address the responsible and effective development and use of GenAI in KSA. It shall apply to all stakeholders designing, developing, deploying, implementing, using, or being affected by GenAI systems within KSA.

Acknowledging the widespread demand for this technology, SDAIA has formulated this document as a robust guideline, fostering both trust and understanding of GenAI systems and assisting to avoid negative outcomes.

3. Benefits of Generative AI

GenAI is expected to transform various sectors within the KSA in the following ways:

3.1 - Increased Efficiency and Productivity

GenAI can automate a wide range of routine tasks that are currently performed by humans, such as generating summaries of complex documents, transcribing text, or generating images and videos. This frees up people to focus on more complex strategic and critical endeavors.

3.2 - Improved Communication and Collaboration

The deployment of GenAI holds promise in areas like the content creation, digital avatars, optimization of bandwidth for video conferencing, and the formulation of sophisticated virtual collaboration environments. These are critical components that can drive enhanced communication and cooperation.

3.3 - Informed Decision-Making

GenAI can assist in the decision-making process by analyzing and summarizing multiformat data (e.g., text, audio, and video) and helping to simulate scenarios that could impact strategic choices. When applied appropriately, this capability provides an additional layer of information, fostering a data-driven approach to decision-making in diverse areas such as crisis management, healthcare, and education.

3.4 - Enhanced Accessibility and Inclusion

GenAI offers solutions to enhance content accessibility. It can produce materials adapted for individuals with special needs or disability, ensuring that platforms and communications are universally accessible. Additionally, by generating content that aligns with local dialects and traditions, it promotes clear communication between people from diverse backgrounds.

3.5 - Elevated Public Service Quality

GenAI has the potential to elevate citizen experiences and boost overall satisfaction with public services through the customization of service delivery and expanded availability. It can be used effectively to optimize citizens' experiences across various public sectors, including healthcare, education, social welfare, and legal assistance.

4. Guidelines of Generative AI

When dealing with GenAI tools, developers and users should consistently adhere to the AI ethics and GenAI guidelines during all phases of the system's lifecycle to harness their benefits while mitigating risks.

The **AI Ethics Principles**, which were developed by SDAIA, are applicable to stakeholders in KSA. They are designed to apply to the use of all AI systems (not just GenAI) and are listed below:

4.1 - Fairness

Requires stakeholders to take necessary actions to eliminate bias, discrimination, or stigmatization of individuals, communities, or groups in the design, data, development, deployment, and use of GenAI systems. Given that GenAI tools have the capability to generate content that may be discriminatory or lack representativeness. To ensure consistent systems that are based on fairness and inclusiveness, GenAI systems should be trained on data that are cleansed from bias and is representative of affected minority groups.

As such, developers and users should:

- Carefully test GenAI models to insure no bias has been embedded in the codes or algorithms and Ensure that the data used to train the tool are cleansed from bias and is representative of affected minority groups.
- Make efforts to gain a thorough understanding of the data used to train the tool – this understanding should encompass the data's origin, its contents, and how it was selected and prepared.
- Enhance knowledge on bias, diversity, inclusion, anti-racism, and values and ethics – this knowledge will improve their ability to recognize biased or discriminatory content.

4.2 - Reliability & Safety

Ensures that GenAI systems adhere to set specifications and that they behave as designers intend. Reliability is a measure of consistency and provides confidence in how robust a system is. On the other hand, safety is a measure of how the AI system does not pose a risk of harm or danger to society and individuals. A reliable and safe working system should have built-in mechanisms to prevent harm.

As such, developers and users should:

- Design and develop GenAI system that can withstand the uncertainty, instability, and volatility that it might encounter.
- Establish a set of standards and protocols for assessing the reliability of a GenAI system to secure the safety of the system's algorithm and data output.
- Predefined triggers/alerts should be in place based on system risks, GenAI systems should trigger human oversight.
- System trigger should be assigned to the appropriate stakeholder. These triggers/alerts can be defined as part of the risk mitigation or disaster recovery procedure and may need human oversight.
- Identify AI-generated content to ensure users are alert to potential reliability issues and can check it against other sources.
- Encourage users to complement AI-generated content with information from reputable sources to ensure accuracy.
- Strive to comprehend the quality and origins of the training data employed by the AI system to enhance content reliability.
- Scrutinize content for factual accuracy and contextual relevance to prevent the dissemination of erroneous information.

4.3 - Transparency & Explainability

Builds and maintains the public's trust in GenAI systems and technologies. In line with this principle, systems must be built to be clearly "explainable" and have features that track how automated decisions are made so that lessons can be learned if these decisions prove to be less than optimal. Transparency is an important consideration when building trust in human-AI interactions in. Notification of AI-generated content eliminates ambiguity, helping users differentiate between automated and human responses.

As such, developers and users should:

- Clearly communicate when GenAI is used in interactions with the public.
- Notify users when messages or content have been generated by AI.
- Offer alternative, non-automated communication channels for users who prefer human interactions.
- Employ tools such as watermarks to assist others in identifying content generated by AI.

4.4 - Accountability & Responsibility

Holds designers, vendors, procurers, developers, owners, and assessors of GenAI systems ethically responsible and liable for the decisions/actions that may negatively affect individuals and/or communities.
The adoption of GenAI systems may entail legal and ethical implications. These warrant thorough consideration and include the risk of infringing intellectual property rights, concerns about data privacy, and the potential for human rights violations.

As such, developers and users should:

- Ensure that data is be properly acquired, classified, processed, and accessible to ease human intervention and control at later stages when needed.
- Ensure data quality checks, cleanse data and validate the integrity of the data in order to get accurate results.
- Build and validate models in a responsible manner to achieve intended results.
- Adhere to relevant legislation, such as Personal Data Protection Law, Intellectual Property Laws to ensure compliance and protect user rights.
- Consult with legal professionals to assess and mitigate risks associated with the deployment of GenAI systems.

4.5 - Privacy & Security

Requires all GenAI systems to be built and operated in ways that protect the privacy of the data they collect. In line with this, GenAI systems should employ best practice data security measures developed by the national authorities to prevent data breaches that could lead to reputational, psychological, financial, professional, or other types of harm.

As such, developers and users should:

- Ensure adequate privacy and security measures in place when using classified data to train GenAI modules.
- Implement rigorous data protection measures and consider the principles such as, the one outlined in the Personal Data Protection Law.
- Assess the risks resulting from the use of the GenAI tool according to the AI ethics principles; little or no risk, limited risk, high risk, unacceptable risk.
- Privacy and security by design should be implemented while building the AI system. The security mechanisms should include the protection of various architectural dimensions of an AI model from malicious attacks. The structure and modules of the AI system should be protected from unauthorized modification or damage to any of its components
- The privacy impact assessment and risk management assessment should be continuously revisited to ensure that societal and ethical considerations are regularly evaluated

4.6 - Humanity

AI systems should be built using an ethical methodology to be just and ethically permissible, based on intrinsic and fundamental human rights and cultural values to generate a beneficial impact on individual stakeholders and communities in both the long and short-term goals and objectives to be used for the good of humanity. Predictive models should not be designed to deceive, manipulate, or condition behavior that is not meant to empower, aid, or augment human skills but should adopt a more human-centric design approach that allows for human choice and determination.

4.7 - Social & Environmental Benefits

This principle embraces the beneficial and positive impact of social and environmental priorities that should benefit individuals and the wider community, focusing on sustainable goals and objectives. GenAI systems should neither cause nor accelerate harm or otherwise adversely affect human beings but rather contribute to empowering and complementing social and environmental progress while addressing associated social and environmental ills. This entails the protection of social good as well as environmental sustainability.

5. Generative AI Risks & Mitigations

While GenAI creates significant opportunities, it is important to make sure that it is used in a safe, ethical, and legal manner. Based on emerging trends in use of GenAI, the following risks and potential mitigation measures should be prioritized.

5.1 - Deepfakes and Misrepresentation

GenAI has transformed the world of digital media through its ability to create realistic text, images, and audio. However, these capabilities can also be used for harmful purposes, like scams, financial fraud, blackmail, and sophisticated identity theft. Often, individuals exploit data from social media and other public platforms to create fake digital representations (i.e., deepfakes). With the technology advancing rapidly, distinguishing between real and fake content has become more challenging.

Mitigation Measures:

- **Watermark Implementation:** All GenAI services should include identifiable watermarks. This measure helps individuals and security services detect and flag deepfakes and AI-synthesized content without affecting legitimate GenAI use.
- **'Know Your Customer' (KYC) Protocols for Cloud Server Providers:** Providers that deliver the intensive computational power needed by GenAI, especially those providing Graphics Processing Units (GPUs) and/or Tensor Processing Units (TPUs), should use verification processes similar to those used by financial institutions. For example, requiring users to provide passport details or other proof-of-identity documents to make it more difficult for malevolent actors to access these powerful tools.
- **Output Verification:** GenAI services should analyze generated content for any inappropriate content (like hate-speech, money transfer request), as well as the use of public figures' faces and voice samples.
- **Enhanced Digital Literacy and Online Safety:** Private companies and public institutions should provide training and have regular awareness campaigns that include the risks of the deepfakes and tools and methods to identify them, as well as information on the importance of limiting personal information exposure online, the dangers of unverified digital communications, and the importance of authenticating information requests through trusted, alternate channels.

5.2 - Safety Threats

While GenAI is a great tool to access information on a broad range of topics, when manipulated with malicious intent, it can compromise public safety and security on an unprecedented scale.

Mitigation Measures:

- **Content Moderation and Filtering Systems:** On platforms where GenAI output is accessible to users, deploying robust content moderation and filtering systems is crucial. Such systems should be able to detect, flag, and prevent the dissemination of harmful or malicious content generated by the AI. Good practice would be to verify both the user prompt and the model's output. Regular updates and refinements to these systems based on feedback and emerging trends will ensure they remain effective against evolving threats.
- **Training Dataset Filtering:** GenAI developers should minimize the amount of potentially dangerous information being used in training datasets for models, keeping in mind that all information used will eventually be available to the general public.
- **Limiting Open Access / Open Source for Scientific GenAI Models:** Scientists working with models that could potentially create safety threats (e.g., protein synthesis) should not publish such models in the public domain without first requiring users to verify who they are and what research they intend to use the model for.

5.3 - Misinformation and “Hallucination”

Due to the nature of GenAI text models, some information they provide might be incorrect. Sometimes such models can even over-confidently generate “facts” that are complete fiction (also known as “AI hallucination”). Without critical review of such outputs, there’s a risk of unintentional misinformation.

Mitigation Measures:

- **Content Verification and Citation:** Publicly available GenAI services should estimate the accuracy of information prior to displaying it to users alongside the content requested. Such services should also be able to cite sources indicating where the information was acquired. If the model cannot produce a response with a high level of accuracy and identify its source/s, the output should be replaced with a pre-defined message warning users that the information generated may not be accurate.
- **Content Labeling:** GenAI services should embed watermarks in outputs based on the type of media used. For photo and video, watermarks can take the form of traditional preprint, or be embedded into encoding or decoding. Audio content can be watermarked through embedded audio snippets or echo modulation. For text, content-specific word and letter combinations can be used, as well as non-standard fonts.
- **User Vigilance and Fact-checking:** It is users’ responsibility to verify the content generated by GenAI. This means users need to scrutinize the content and fact check it against trustworthy sources. Users should also always look for any indication that the content came from a GenAI tool if this was the way it was produced (see Content Labelling above)
- **Raising Awareness:** Users should be regularly reminded that outputs from GenAI can sometimes be inaccurate, outdated, biased, or even deceptive, and be educated about standard fact-checking and content-verification processes.

5.4 - Classified Data Breaches

When interacting with GenAI, users should be aware that they might unintentionally expose sensitive information. This is because GenAI services use the information received through prompts as training data for further development of the model – this can then be exposed to third parties. This risk is complicated by the fact that GenAI models do not store information in traditional document form, meaning it's harder to identify if any sensitive information was in fact leaked or to make the model "unlearn" this information.

Mitigation Measures:

- **GenAI Usage Protocols:** Organizations should implement policies for GenAI use that prohibits users from entering classified information into third-party tools. These should detail acceptable content generation practices, actions to prevent sensitive data leaks, and steps to support ethical use of the technology. Oversight and regular reviews are essential to confirm adherence to policies, and rapid, organization-wide communication is key to their successful implementation.
- **Employee Awareness and Training:** Organizations must emphasize to their workforce that existing and upcoming legal requirements remain binding when using GenAI tools. This includes adherence to company policies, as well as broader legal mandates encompassing data governance, cybersecurity, government data classification, personal data protection, intellectual property (IP) rights, and other pertinent legal or policy areas. Targeted training sessions can provide guidance on how to reduce potential legal risks associated with the deployment of GenAI.
- **User Data Control:** GenAI service providers should provide users with options to give – or refuse – consent to use their data for AI model training purposes, they should also provide an option to remove all prompt history on request in accordance with relevant laws and regulations.

5.5 - Certification Fraud

Human certification processes, such as exams and professional evaluations, stand as essential benchmarks of individual competence and institutional credibility. However, with the advent of GenAI, these processes are facing novel threats. GenAI, with its ability to craft 'human-like' specialized content, can be misused to produce answers, essays, or even detailed research, effectively undermining traditional educational and professional standards. The sophistication of this AI-generated content poses a distinct challenge: it is original and not directly copied from known sources, making it harder to detect with standard anti-plagiarism tools. Thus, the unchecked misuse of GenAI could jeopardize the integrity of foundational educational and professional assessments.

Mitigation Measures:

- **Assessment Enhancement:** Review all forms of assessment and evaluation to enhance their resilience against AI-aided fraud. Consider adjustments such as transitioning to in-person assessments, revising question formats, or adopting innovative evaluation methodologies. Moreover, organizations should conduct a thorough evaluation of critical certification exams, particularly those essential for positions like national critical infrastructure access or medical practice licenses, where certifications determine eligibility. In cases where GenAI could potentially be exploited to pass such certifications, organizations must proactively revise the certification processes to reduce the likelihood of GenAI use. This ensures the assessments maintain their integrity and security, especially in vital areas.
- **Education and Training:** Conduct comprehensive training to raise awareness and educate both educators and students about responsible GenAI use. This includes recognizing misuse, adhering to ethical practices, and identifying suspicious behaviors that may indicate fraud.
- **Guidelines and Policies:** Collaboratively develop and implement clear guidelines in conjunction with students and educators to govern GenAI usage in academic settings. These guidelines should align with learning objectives and specify when and how GenAI can be applied.
- **Query Proficiency:** Provide specialized training to enhance the competency of educators, researchers, and students in crafting precise and effective prompts for GenAI systems. Emphasize the significance of formulating inputs that yield optimal AI outputs.

5.6 - Intellectual Property Infringement – and Protection

The rise of GenAI has brought to the forefront the issue of unauthorized use or replication of copyrighted material, potentially leading to legal liabilities of IP. This risk extends to both the unauthorized use of existing copyrighted material and the creation of novel content by AI that may inadvertently infringe upon IP rights.

Mitigation Measures:

- **IP Licensing and Due Diligence:** GenAI developers should obtain licenses for any IP included in training data, preventing indiscriminate content scraping. Customers should perform their own due diligence to confirm whether AI models were trained with protected content before using them.
- **Creator Permission and Compensation:** GenAI services should obtain original content IP holder permission before generating content based on their IP. GenAI services should establish compensation mechanisms (e.g., contributor funds) to fairly remunerate creators if their IP has been used in training sets. For instance, Shutterstock offers creators the option to opt out of having their work used in AI training sets and has set up a contributor fund to compensate them when their work is used.

5.7 - Variability of Outputs

GenAI services operate differently from traditional “programmed” services where output generation adheres to pre-defined algorithms in 100 percent of cases. Furthermore, developers of such services may update their offerings without informing end users. Consequently, users should be very vigilant when relying on outputs generated by GenAI.

Mitigation Measures:

- **Clear Annotation for AI-Generated Code:** Code developers who leverage GenAI coding tools, such as ChatGPT and GitHub Copilot, should add clear annotations within the source code to indicate that it was generated by AI. This annotation not only promotes transparency but also ensures accountability for the code’s origin and quality.
- **Regularly Verify and Validate:** Users should adopt a practice of regularly verifying and validating AI–generated code and other content, especially in critical applications where accuracy is paramount. Implementing thorough testing and validation processes can help identify and rectify any inconsistencies or errors in the outputs.
- **Stay Informed About Updates:** To mitigate the risk of unannounced updates to GenAI services, users should stay informed about any changes or enhancements made by the service providers. Subscribing to notifications and updates from the developers can help users adapt to modifications in the service’s behavior or output quality.
- **Build Expertise:** Users, particularly those in software development and coding roles, should develop expertise in understanding GenAI. This proficiency will help when evaluating and refining AI-generated outputs to meet specific requirements.