“PETs Sandbox Expansion – Use of PETs for Generative AI Use cases”

Introduction

As a recognition of the potential that PETs hold, The PETs sandbox was introduced in 2022 to provide a testing ground for businesses to pilot their PET use cases, with technology, financial, and regulatory support from IMDA, across three archetypes: (a) identifying common customers in multiple datasets; (b) deriving additional features of common customers from multiple datasets; and (c) making more data available for AI. The sandbox has already seen active industry participation, with some projects targeting the data protection challenges inherent in traditional AI.

The IMDA is keen to extend similar support and exploration into the realms of generative AI. We acknowledge that the use of PETs in generative AI is nascent today. Nonetheless, our conversations with industry partners indicate that businesses recognize both the potential of generative AI and the personal data protection risks that it poses throughout its lifecycle and are interested in the use of PETs to address such risks.

The use of PETs can drive growth in generative AI use cases by unlocking more data and addressing data protection risks

Since November 2022, generative AI has surged in popularity, exemplified by the success of ChatGPT and the subsequent proliferation of similar applications. This technology has the potential to deliver significant value to the global economy over the next few years by driving innovation, improving efficiency, and creating new opportunities.

Central to this potential is the critical role of data. Generative AI models require large amounts of data for training to generate accurate and contextually relevant output. In contrast to traditional AI, training data for generative AI is more voluminous, diverse, and complex. This poses a greater challenge in curating and removing personal data, resulting in a higher risk that confidential data (i.e. personal or commercially sensitive data) may be included in the training dataset.

Further, when users interact with generative AI, both the inputs (prompts) and corresponding outputs can include anything, including confidential data. Confidential data entered as input prompts may be stored and may be inadvertently used to further train the generative AI models. Consequently, generative AI may output such confidential data in interactions with users.

The use of PETs could help to unlock more data for generative AI use cases, and address confidential data related risks largely across the following areas:

1. [Input] To “make available” more data:
    a. E.g. The use of Synthetic Data (SD) techniques to create statistically similar data for training or testing purposes.
    b. E.g. The use of Homomorphic Encryption (HE) to encrypt and enable compute on sensitive datasets, so that confidential data can also be used in a privacy preserving manner.
2. [Output] To obfuscate or remove confidential data:
    a. E.g. The use of Differential Privacy (DP) to add noise to output (reports or analysis) to lower the likelihood of re-identification.

The use of PETs has been steadily growing in traditional AI use cases, but it is still nascent for Generative AI

PETs have been successfully used to protect the collection, sharing and use of confidential data in traditional AI[1] use cases. Examples include:

PET Used	Example in Traditional AI
Differential Privacy (DP)	**Apple uses DP** when collecting user data to train machine learning models to power keyboard predictive text features.[2]
Federated Learning (FL)	Using FL, **Moorfields Eye hospital NHS Foundation Trust deployed machine learning models for the diagnosis and treatment of common eye diseases**. [3]
Fully Homomorphic Encryption (FHE)	**Banco. Bradesco used FHE on financial data for machine learning**, and proved similar levels of accuracy and privacy could be achieved.[4]
Synthetic Data (SD)	**American Express used synthetic data to train its AI models** to improve detection of rare and uncommon frauds.[5]
Trusted Execution Enclaves (TEEs)	**The Weather Company (an IBM business)** used AWS Clean Rooms to enable advertisers to analyse their data together with weather data and **used predictive machine learning to identify engaged audiences at scale**.[6]

However, examples of PETs in generative AI, while growing, are still explored primarily by larger technology companies:

- Microsoft[7], Meta[8], and Anthropic[9] have used **synthetic data** to train their respective generative AI models.
- Apple recently launched Private Cloud Compute[10], which uses **secure enclaves to process user inferences for generative AI services.**

As the use of PETs for generative AI is still nascent, to facilitate a more comprehensive understanding of PETs and their potential to enable privacy-preserving generative AI use cases, it is essential to expand real-

[1] Traditional AI refers to AI models that make predictions by leveraging insights derived from historical data. Typical traditional AI models include logistic regression, decision trees and conditional random fields. Other terms used to describe this include "discriminative AI".
[2] Apple Machine Learning Research, Learning with Privacy at Scale, Dec 2017
[3] Open Data Institute, Federated Learning: An Introduction, Jan 2023
[4] IBM Research, Top Brazilian Bank Pilots Privacy Encryption Quantum Computers Can't Break, Jan 2020
[5] Fortune, American Express is trying technology that makes deepfake videos look real, Sep 2020
[6] AWS, AWS Clean Rooms ML
[7] Microsoft, Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone
[8] Meta Responsible AI, Our responsible approach to Meta AI and Meta Llama 3
[9] Anthropic, The Claude 3 Model Family: Opus, Sonnet, Haiku
[10] Apple Security Research, Privacy Cloud Compute: A new frontier for AI privacy in the cloud, Jun 2024

world implementations across various industries (e.g. finance, healthcare, supply chain, etc) and different types of organisations (e.g. MNCs, governments, SMEs, etc).

Encouraging experimentation and feedback through a new Archetype for Generative AI

At IMDA, we adopt a "use case centric" approach through the PETs Sandbox, emphasizing close partnerships with industry to learn from the practical implementation of solutions. It is crucial to ensure that feedback and learnings are factored in, so that any policy or guidance[11] developed is meaningful, grounded, and reflective of industry's requirements.

In working with industry on use cases, the Sandbox has validated the potential of PETs to support the use and sharing of data in a trusted and accountable manner across industries such as finance, healthcare and AdTech.

With greater interest in Gen AI applications from industry, and the need to better understand and address data protection risks in such applications, the PETs Sandbox will be expanded to include a new archetype – **"Data Use for Gen AI"**, that would focus on model and application 1) development and 2) use (See Annex A for more information)

While not traditionally considered PETs, Gen AI can be used to identify and flag personal data, **which can then be removed or obfuscated**. The use of such technologies would also be relevant, and we encourage use cases employing such solutions to also come onboard our Sandbox.

We invite industry to propose use cases for collaboration under this new archetype.

[11] For example, the PDPC has published the "Proposed Guide on Synthetic Data Generation" that would be finetuned iteratively with new inputs from industry based on PETs Sandbox use cases

Annex A

Potential use cases would focus on 2 key areas:

Areas across Model / App	**Scope**	**Potential examples of PETs use cases**
Development	Making more data available for Gen AI model and app development, e.g. when data needs to be shared for pre-training, fine-tuning or RAG	• Gen AI model training in a TEE • Generating synthetic unstructured data for training gen AI models
Use	Addressing the potential for confidential data to be included in the input and output when interacting with Gen AI apps	• Encrypted Inferences using SMPC • Conducting Inferences in a TEE • Gen AI PETs solutions to identify and anonymise CD in input and output