www.astri.org | TECH FOR IMPACT

人工智能與私隱保障

發展與安全並重 - 開發及應用AI的隱私保護方案

鄧羽眞博士 | 總監

香港應用科技研究院有限公司
Hong Kong Applied Science and Technology Research Institute Company Limited

應用科技研究院的科技發展

香港最具規模的應用科技研究機構

研究人員中有
20% 博士　56% 碩士

780+ 名
僱員分佈在香港和內地

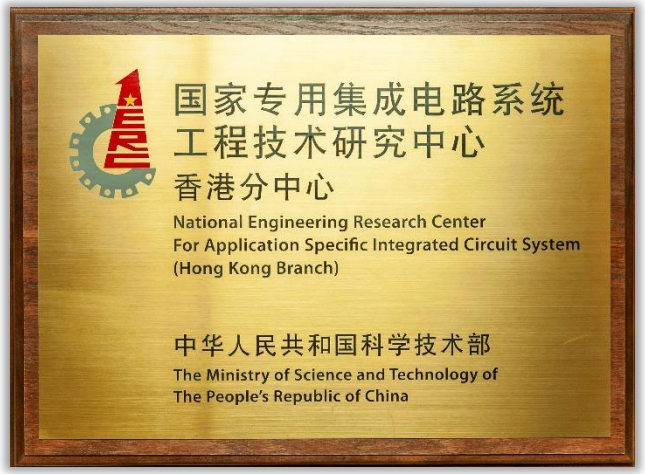
完成
1,100+ 個
研究項目

轉移
1,500+ 項
技術至業界

1,100+ 項
專利授權

數據截至2024年3月31日

中华人民共和国科学技术部
Ministry of Science and Technology of the People's Republic of China



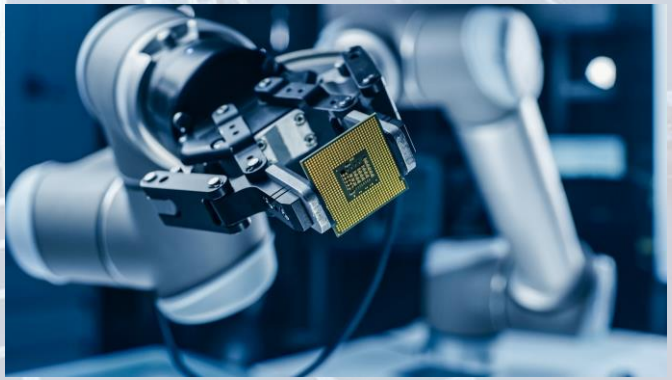
六項重點範疇

智慧城市

金融科技

數碼健康科技

新型工業化及智能製造



專用集成電路

元宇宙

ASTRI

金融科技生態系統多元化

渣打銀行、平安壹賬通銀行、OpenRice、 Lalamove

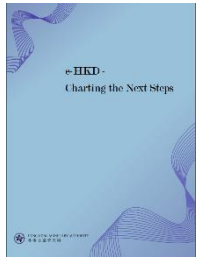
聯盟式學習
協助中小微企業獲得融資信貸評級

- *人工智能（AI）模型*
- *全面的信貸分析*
- *加強私隱保護及資料安全性*

香港金融管理局

央行數碼貨幣 (CBDC)
「數碼港元 e-HKD」先導計劃啟動

- *應用零售層面央行數碼貨幣技術*
- *與業界聯手探索創新用例*



ASTRI

參考資訊

香港金融科技生態系統

替代信用評估
中小企業智慧信貸

- *利用替代資料進行信用評分*
- *確保資料安全和共用專業見解*

生成式人工智能 — 助力金融科技與永續發展

私有化部署 — 專屬ChatGPT
具有可追蹤連結和高精準度生成答案

- *設特定領域和向量資料庫*
- *專屬內部使用，確保資料機密*

工銀亞洲、周大福、中國移動 (香港)

AI 聊天機器人、光學字元辨識 (OCR)
提升產品體驗和服務效率

- *特殊的語言環境：香港廣東話*
- *語音情緒辨識*
- *手寫中文字元識別*

大灣區碳中和協會

人工智能 ESG 報告分析
實現海量資料自動提取

- *發佈「碳中和100強榜單」*
- *自動收集、提取並分析關鍵資訊*
- *減少人工檢查流程及合規成本*

ASTRI

創新科技屢獲殊榮

Inventions Geneva

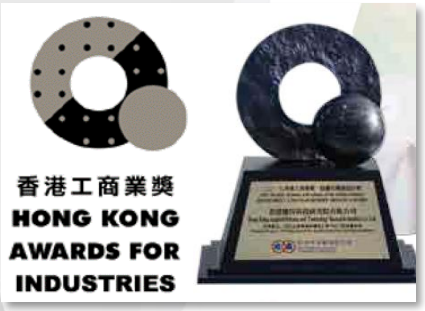
2018 – 2024
歷屆榮獲
超過120個獎項

- 特別創意大獎 1 項
- 評審團嘉許金獎 10 項
- 金獎 36 項
- 銀獎 59 項
- 銅獎 21 項

ICT 中國創新獎 2021

最佳「技術創新應用」

香港工商業獎



- 科技成就獎 (2022)
- 設備及機器設計獎 (2022)
- 設備及機器設計優異證書 (2022)

香港資訊及通訊科技獎

- 智慧生活 (2021- 2018)
- 智慧市民 (2021)
- 智能出行 (2020)
- 智慧商業 (2018, 2022, 2023)

科創中國 2021全球百佳技術轉移案例

8 項技術榮獲最佳跨境創新技術轉移案例

科 创 中 国
INNOVATION CHINA

2020年國家科學技術進步獎

一等獎

第三屆亞洲創新發明展覽會

榮獲 10 個獎項

- 評審團嘉許金獎
- 日內瓦發明一等獎
- 2項金獎及6項銀獎

獲獎


2024
BUSINESS
GOVIRTUAL
TECH AWARDS
EXCELLENCE AWARD
TECH COMPANY OF THE YEAR
(LARGE CORPORATE)
CIENCE AND
H INSTITUTE
(ASTRI)


27th 周年紀念 ANNIVERSARY
創新科技署
Innovation and Technology Commission
慶祝中華人民共和國成立75周年暨香港回歸祖國27周年活動
2024年創新科技獎項嘉許禮
香港創科成就 揚威世界舞台
HONG KONG'S I&T ACHIEVEMENTS SHINE IN INTERNATIONAL ARENA


EXCELLENCE AWARD
TECH COMPANY OF THE YEAR -
FINTECH
2023
BUSINESS
GOVIRTUAL
TECH AWARDS
HONG KONG APPLIED SCIENCE AND
TECHNOLOGY RESEARCH INSTITUTE


am730 NEWS
2018.12.6 THU
中文手寫文字識別系統 1秒識別每頁內容
機械人隨行協助


新系統4秒將手寫字轉電腦文檔
另外，本港不少行業面臨人手荒，香港應用科技研究院研發人工智能中文手寫文字識別系統，只需四秒便可完成掃描文件及將文件上的手寫文字轉為電腦文檔，取代人手逐字輸入。系統可識別七千多個字符，研究院智能軟件與系統經理鄧羽真表示，研發期間，人員將多種字體輸入進系統以供學習，現時銀行、保險、速遞等行業需要大量人手輸入相關客戶資料，單是一間中型保險公司便需要五十人，相信系統可減少人手成本及提高效率，目前已有銀行正運用系統。
■鄧羽真相信，系統可減少人手成本及提高效率。（何青露攝）
載Apps體驗


ASIA
FIN
TECH
awards


inTech and
nnovation Award
022/2023
iFTA
CFT


FITMI第十二屆理事會就職典禮暨香港創新科技成就大獎
亞洲國際創新發明大獎2019頒獎典禮


ASTR
Spirit of Hong Kong
DIPLÔME DIPLÔME
SALON INTERNATIONAL DES INVENTIONS GENÈVE
SALON INTERNATIONAL DES INVENTIONS GENÈVE

ASTRI

ASTRI

開發及應用AI模型與隱私保護

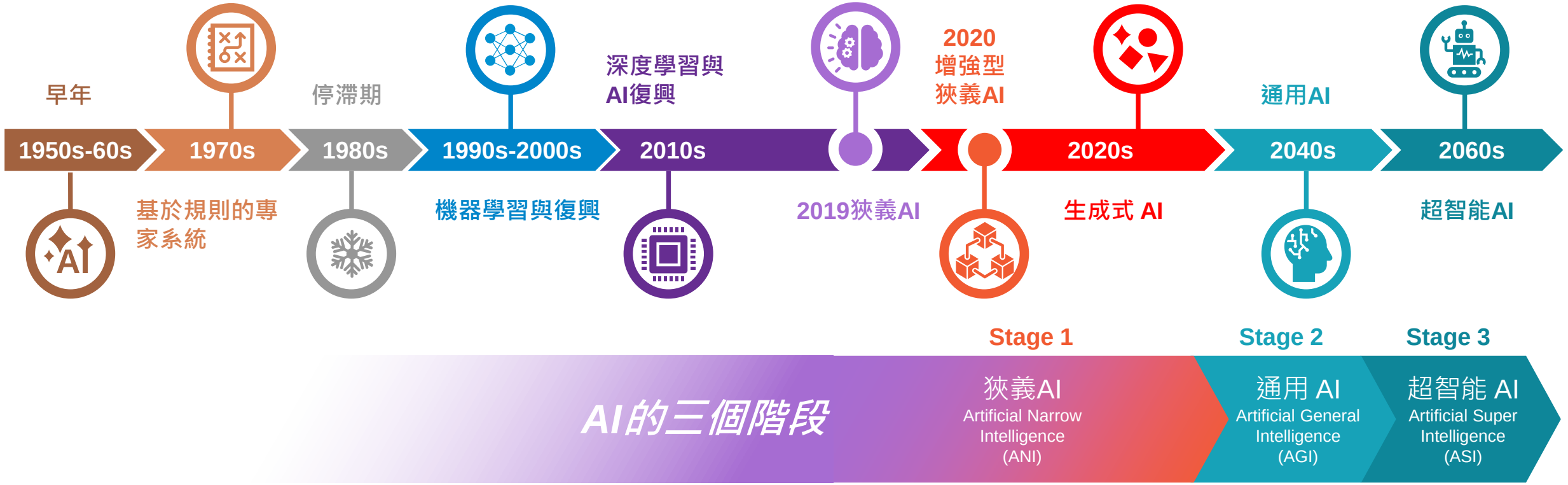
Can you still trust what you see online?

簡介

- AI演變與發展歷程
- 隱私保護的重要性與挑戰
- 實施 AI 隱私保護方案
- 應對 AI 隱私挑戰

ASTRI

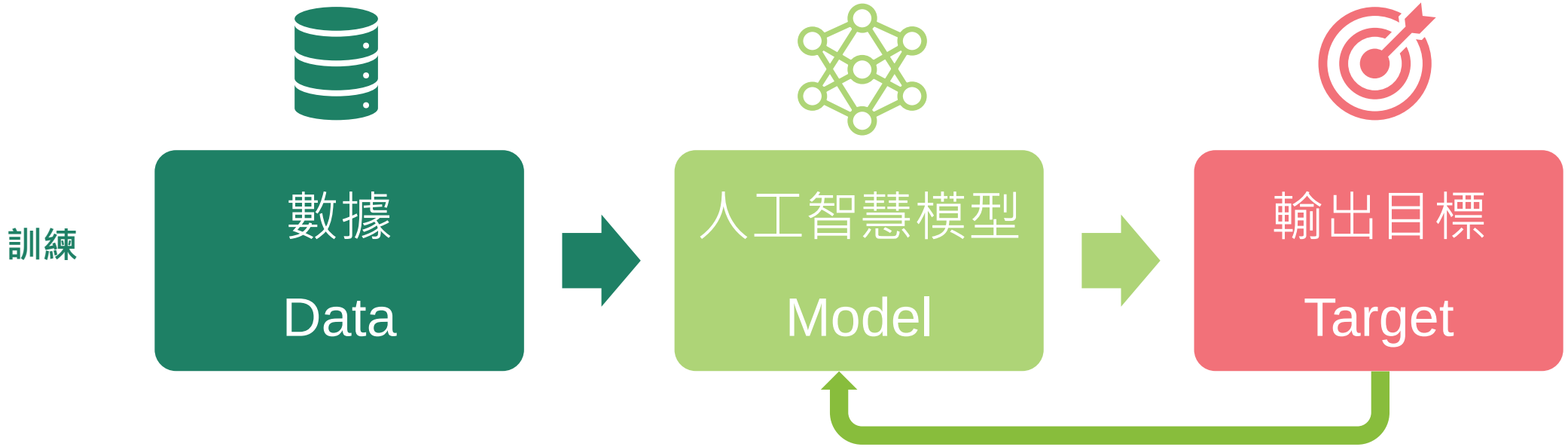
AI演變與發展歷程



人工智能（AI）的發展徹底改變了各個行業，在數據分析、自動化和決策方面提供了前所未有的能力。然而，隨著人工智能系統越來越融入我們的日常生活，大量個人資料的處理引發了嚴重的隱私問題。

ASTRI

什麼是機器學習或人工智慧？



- 數據驅動(data-driven) 的方法來構建機器學習模型
- 這是一個基本的監督學習 (supervised learning)
- 當未指定輸出目標時，它將是無監督學習 (unsupervised learning)

ASTRI

AI 發展中隱私保護的重要性

ASTRI

AI 發展中隱私保護面臨的挑戰

AI 的監管

- 確保不同司法管轄區的合規性
- 應對多樣化且不斷發展的隱私法規

20 biggest GDPR fines so far

1. Meta GDPR fine- €1.2 billion

In May 2023, in a groundbreaking decision in the past five years of GDPR enforcement, the **Irish Data Protection Commission (DPC)** imposed a historic fine of **€1.2 billion** on US tech giant Meta.

不斷變化的威脅和風險

- 持續防禦新出現的威脅或新型網路攻擊
- 可用的緩解策略和應急計劃

Daryna Antoniuk

March 27th, 2024

Cybercrime

Technology

News

Thousands of companies using Ray framework exposed to cyberattacks, researchers say

隱私保護的挑戰

SCIENCE

Not so anonymous: Medicare data can be used to identify individual patients, researchers say

ABC Science / By technology reporter Ariel Bogle

Posted Mon 18 Dec 2017 at 10:22am, updated Mon 18 Dec 2017 at 11:20am

處理個人數據

- 合法的資料收集和處理
- 安全資料存儲，例如匿名、加密

Facial recognition: 20 million euros penalty against CLEARVIEW AI

20 October 2022

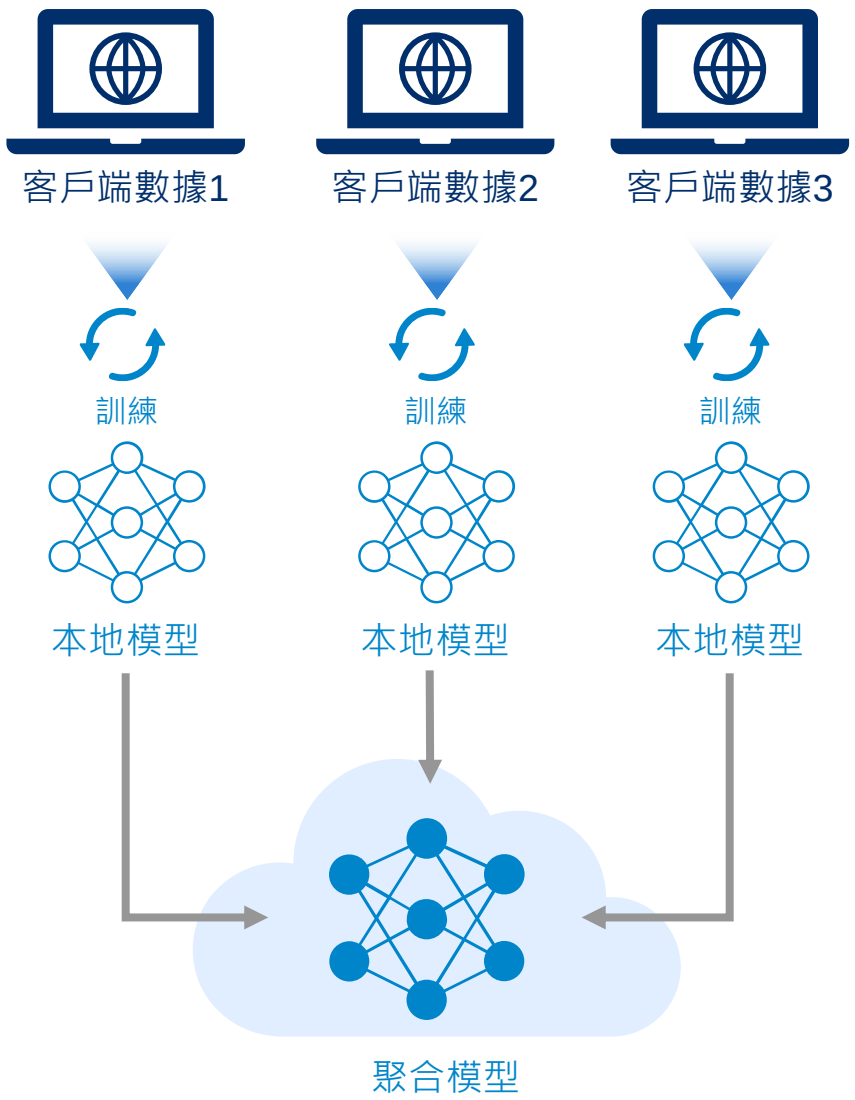
Following a formal notice which remained unaddressed, the CNIL imposed a penalty of 20 million euros and ordered CLEARVIEW AI to stop collecting and using data on individuals in France without a legal basis and to delete the data already collected.

隱私與實用的平衡

- 在保護隱私的同時保持模型準確性
- 解決隱私和效能之間的權衡

ASTRI

實施AI隱私保護方案：聯邦學習 (Federated Learning)



什麼是聯邦學習 (Federated Learning)

- ▷ **一種分布式機器學習方法**
- ▷ **模型分享而非原始數據**
- ▷ **控製本地節點之間數據訪問權**
- ▷ **隱私保護和數據安全**

關鍵技術優勢:

- ✓ **隱私保護** – 兼顧處理「數據孤島」和隱私保護問題
- ✓ **降低責任** – 公司避免集中式數據的處理或泄露風險
- ✓ **合法合規** – 符合《個人資料(私隱)條例》等限製個人資料收集和儲存的法規要求
- ✓ **增強模型** – 通過「數據可用不可見」, 提升準模型確性和性能

ASTRI

實施AI隱私保護方案：聯邦學習 (Federated Learning)

用例：聯邦學習改善貸款申請

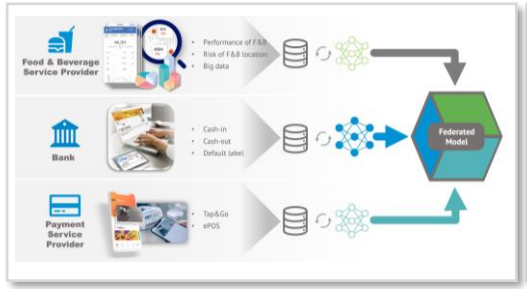
實施AI隱私保護方案：差分隱私 (Differential Privacy)

多級動態差分隱私

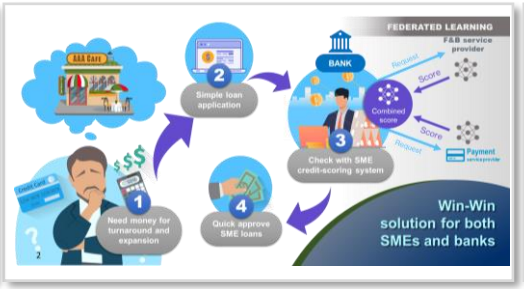
DP1
第一級
DP聚合模組

DP2
第二級
DP 機器學習算法

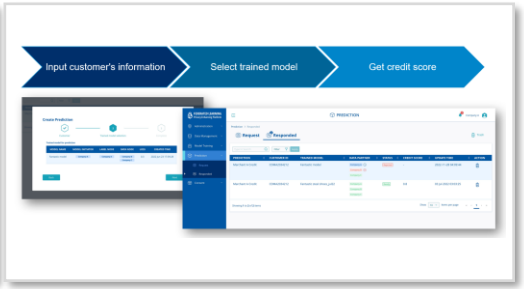
DP3
第三級
DP聚合模組



聯邦學習改善貸款申請



聯邦學習替代信用評分



ASTRI 聯邦學習平臺

研究背景與動機

- 聯邦學習作為一種保障數據安全的建模方法，在保險、金融等行業中的應用前景十分廣泛. 因為這類行業昔論受到更為嚴格的監管和隱私保護法規的約束
- 我們利用聯邦學習來協助對中小企業的貸款申請進行信用評分
- 然而，在聯邦學習中，模型交換是必要的一步，但隱私泄露的問題仍可能發生

我們的發明旨在解決上述挑戰，加強聯邦學習應用中的隱私保護。

創新及影響

- 動態差分隱私
- 多級動態差分隱私
- 保護機制
- 預製式用戶介面

相關產品

我們的產品目前已被國內外的銀行使用，用於協助信用評分和貸款申請的評估工作

ASTRI

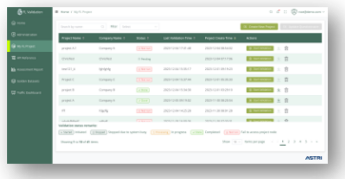
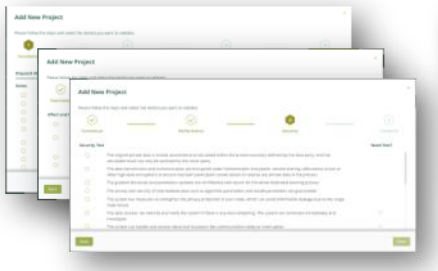
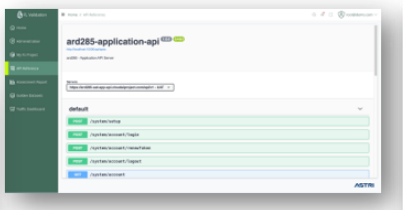
聯邦學習 (FL) 自動驗證系統

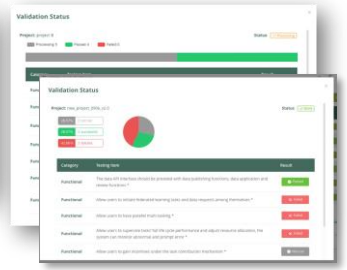
- 隱私保護是首要動機。但目前大多數聯邦學習（（FL）產品對於用戶來說仍然是「黑匣子」，數據合作夥伴在向「黑匣子」**提供數據時擔心隱私問題**。
- FL平臺與用戶、數據合作夥伴之間的溝通**不規範、無參考性**，影響了FL的兼容性和發展。
- 開發一個**自動驗證系統**，構建系列FL基準API，根據其填寫的技術調查問卷來評估所測試的 FL 平臺，並生成**評估報告**。

FL驗證申請人

評估報告

註冊	創建FL項目	資質檢查	基準API	自動遠程評估系統	報告
▶ 提交公司基本資料	▶ FL項目名稱 ▶ 服務器信息	調查問卷： ▶ 功能測試 ▶ 性能測試 ▶ 安全測試	▶ API文檔 ▶ 參考API實現	▶ 功能測試 ▶ 性能測試	▶ 在線報告 ▶ 報告下載 ▶ 購買完整報告

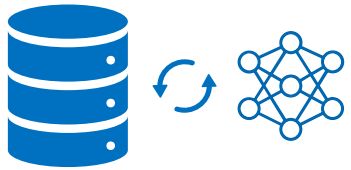


實施AI隱私保護方案：合成數據 (Synthetic Data)

隱私資料

- 收集個人資料用於機器學習可能會導致**隱私洩露**？
- 如何遵守個人資料隱私**法規**？

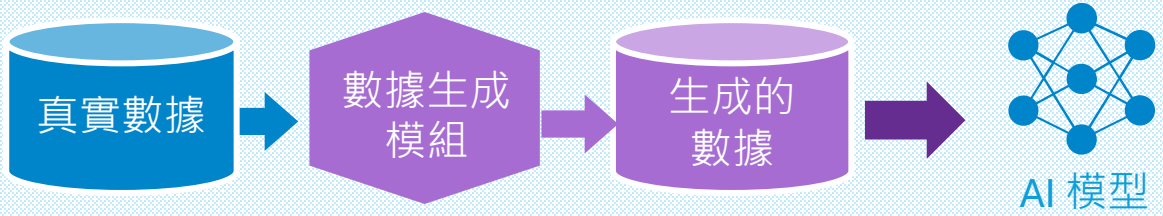
使用**真實數據**進行機器學習訓練



偏差

- 如何避免可能導致**道德問題**的訓練資料偏差？
- **少數群體**的準確度較低。

合成數據 (SYNTHETIC DATA)



隱私保護

- 在資料產生過程中保護隱私

優質數據

- 產生與原始數據具有相同統計特性的數據
- 低成本高效益的生成方式

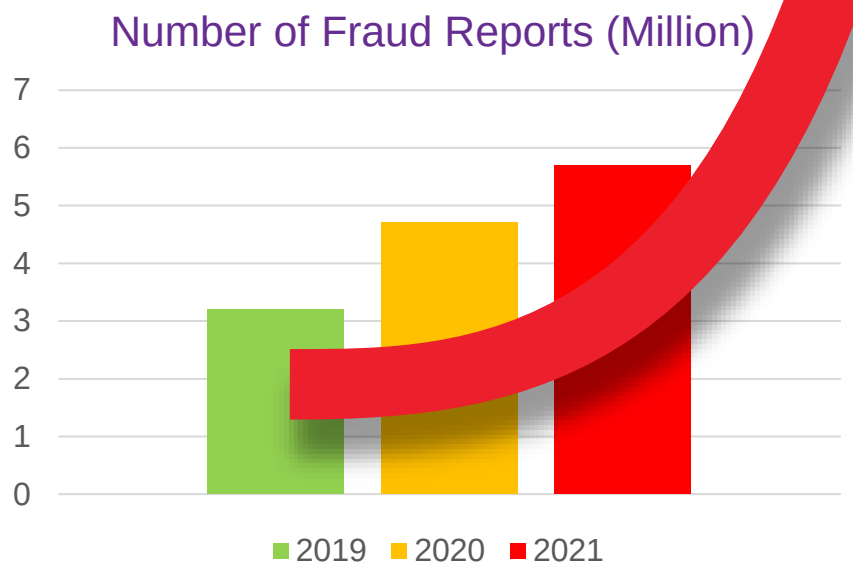
偏差修正

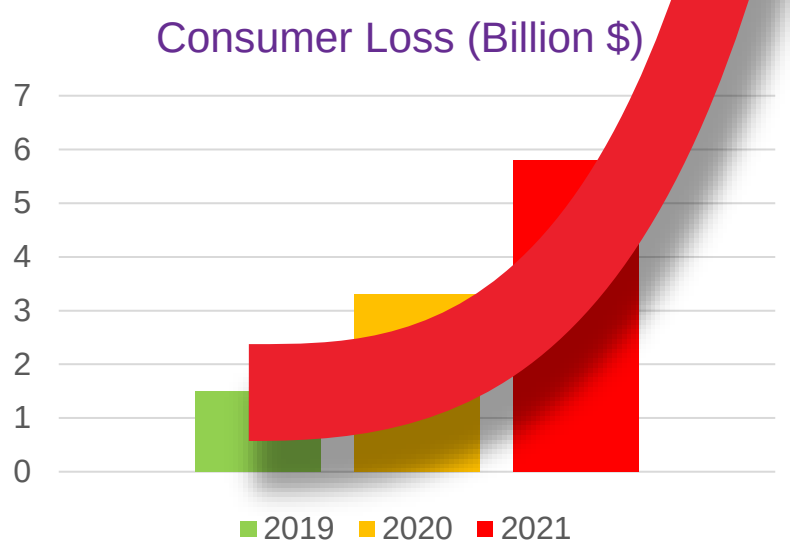
- **可程式**自訂數據統計特性
- 數據產生的**偏差修正**

AI相關欺詐

- Deepfake冒名頂替騙局.

- 根據Federal Trade Commission (FTC eports[2,3]:

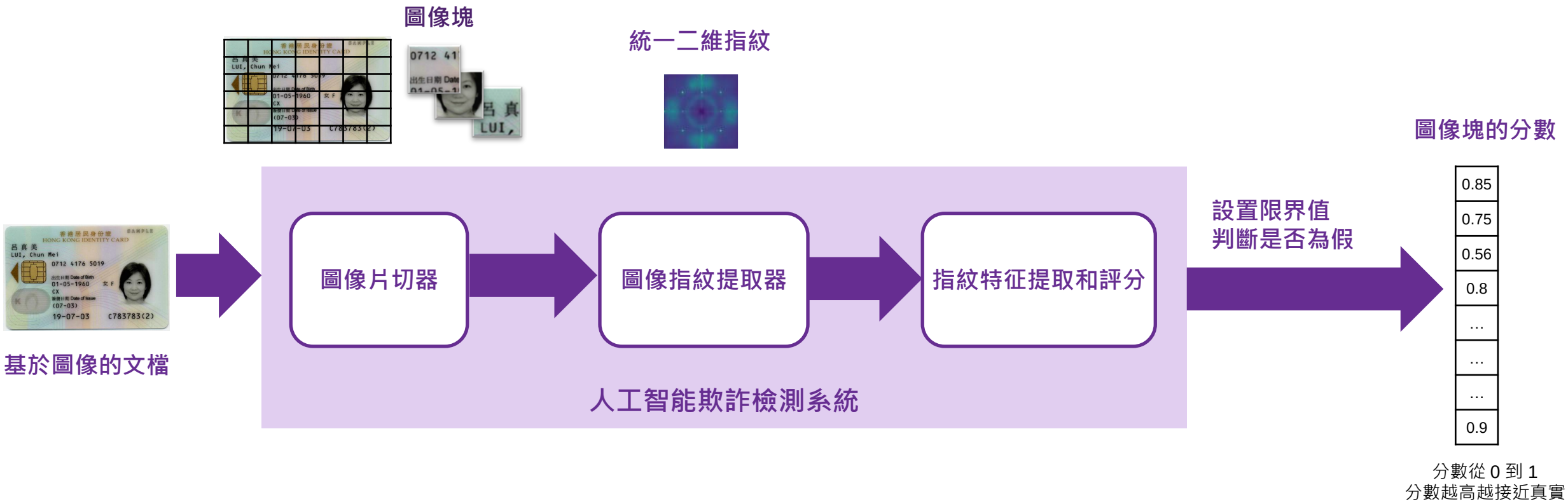




Ref: 1. https://www2.hkej.com/instantnews/current/article/3544322/%E6%B6%89AI%E6%8F%9B%E8%87%89%E5%86%92%E8%AA%8D%E8%BA%AB%E4%BB%BD%E7%B6%B2%E4%B8%8A%E5%80%9F%E8%B2%B8+6%E4%BA%BA%E8%A2%AB%E6%8D%95
2. https://www.ftc.gov/news-events/news/press-releases/2021/02/new-data-shows-ftc-received-22-million-fraud-reports-consumers-2020
3. https://www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0

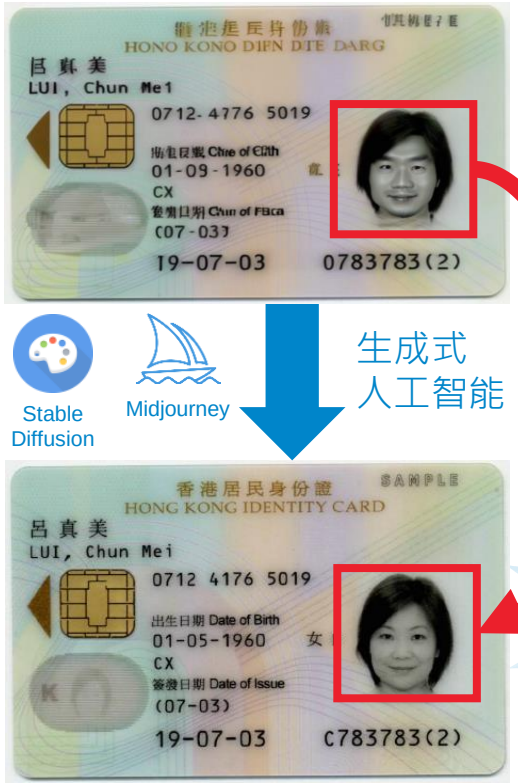
ASTRI

人工智能欺詐檢測系統結構



Ref.: R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Intriguing properties of synthetic images: from generative adversarial networks to diffusion models", CVPR Workshops 2023: 973-982.

生成式人工智能 檢測銀行借貸中的欺詐身份證明



Stable Diffusion

Midjourney

生成式
人工智能

由AI生成的假身份証

未使用我們的欺詐檢測系統的銀行

員工手動驗證

未識別出假身份證

金錢損失
增加風險
聲譽損害

使用我們的欺詐檢測系統的銀行

人工智能欺
詐檢測系統

Powered by ASTRI

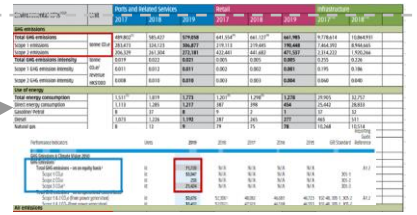
人工智能
檢測報告

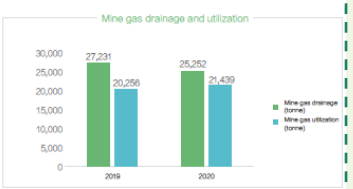
防損
拒假案件

ASTRI

AI ESG揭露報告分析

ESG reports





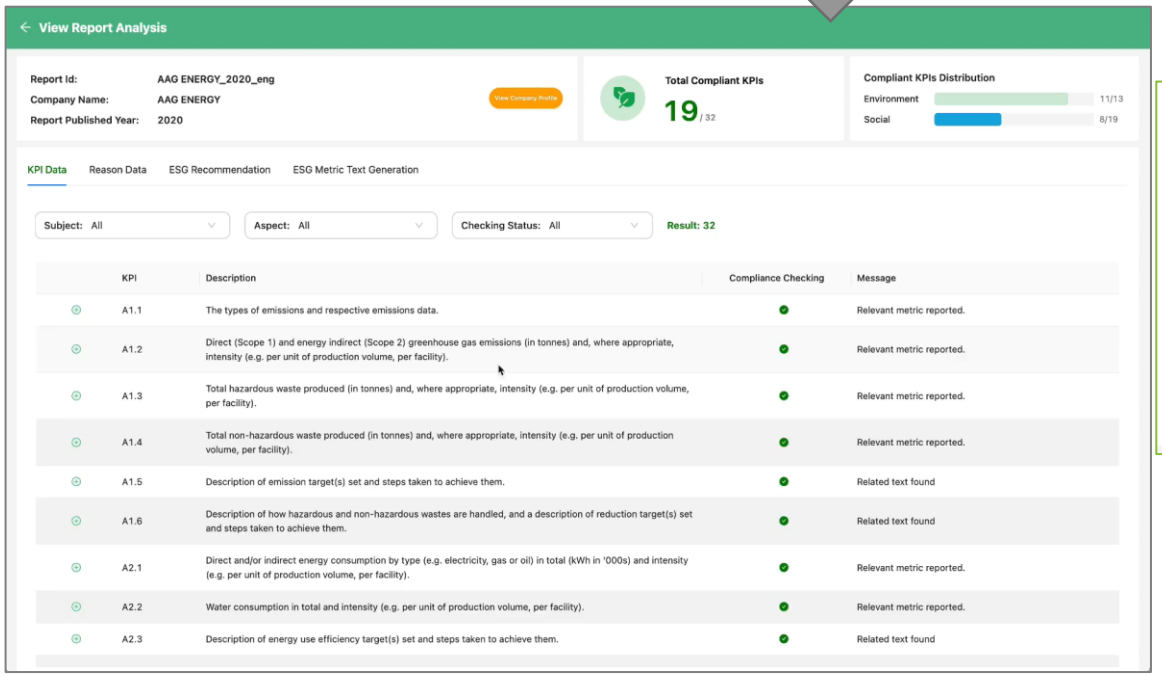
During the Reporting Period, the Group emitted a total of 2.31 kg sulphur oxides (SOx), 15.29 kg of nitrogen oxides (NOx), and 1.48 kg of particulate matters (PM).

- 信息抽取
- 文檔審查
- 文檔摘要
- 文檔生成
- 文檔分類

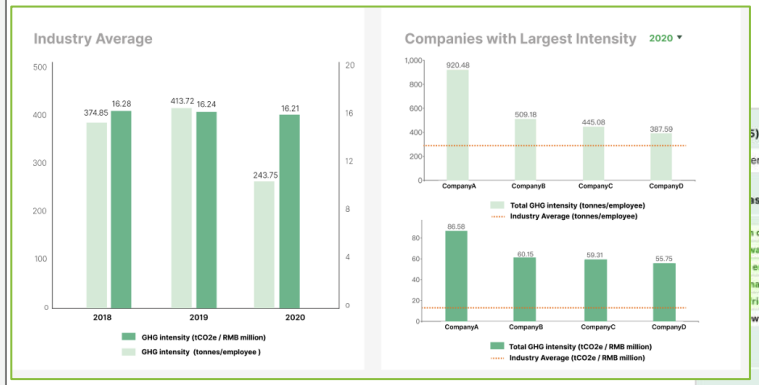
- 多模態特征融合：表格+文本+圖表
- 通用型
- 多樣性

AI ESG

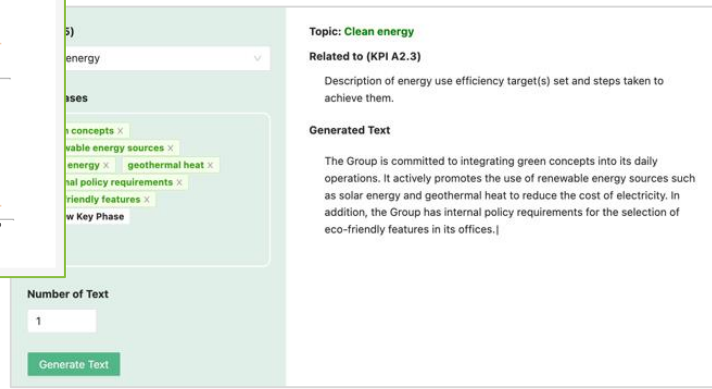
自動化、精簡、高效和有效



公司概況和合規性檢查



行業基准及同業比較

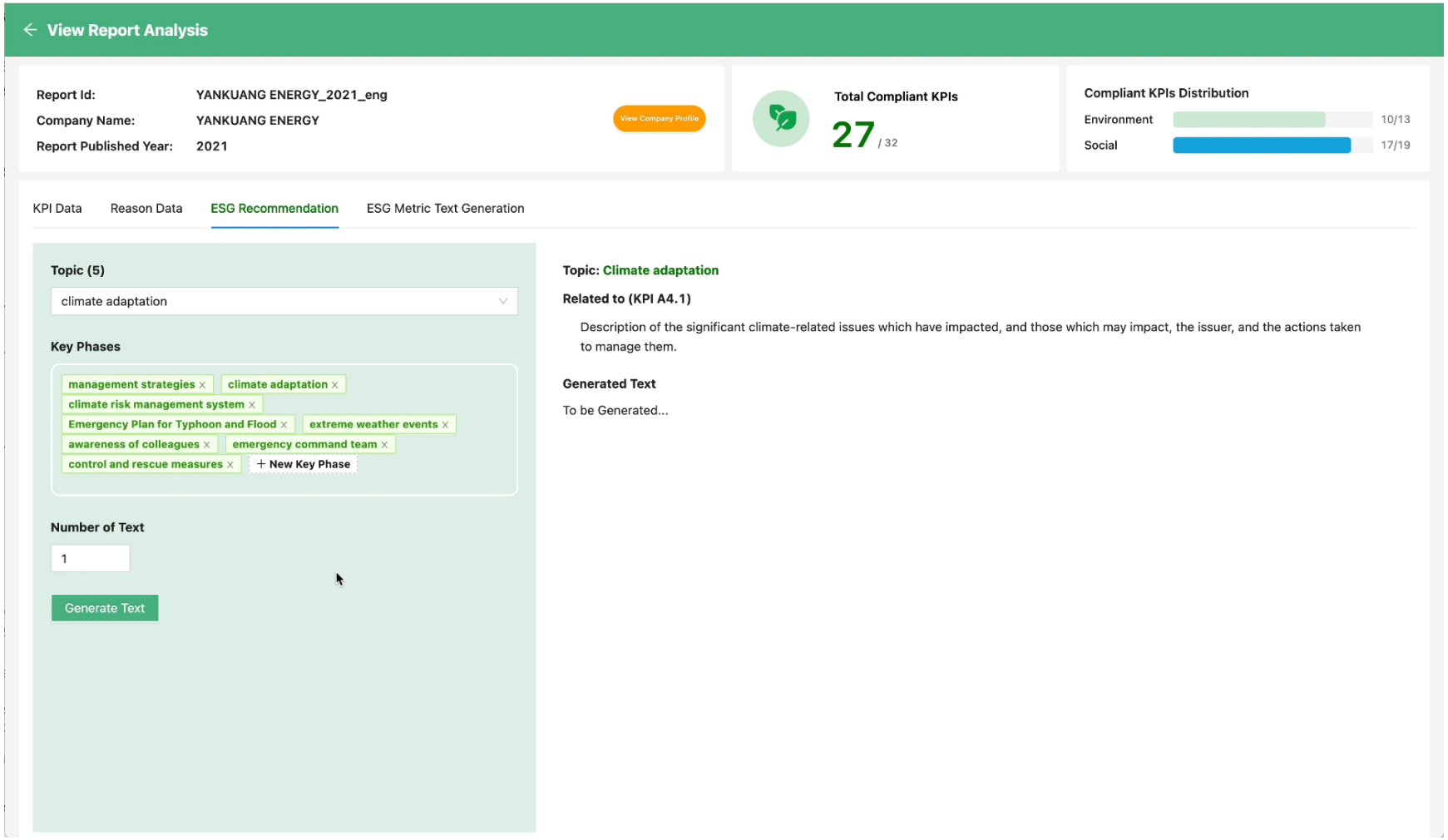


策略分析及改進建議

ASTRI

ESG改進建議生成 / 報告生成

- 識別公司ESG報告中缺失/表現不佳的部分並提出ESG改進建議。
- 幫助公司 進行報告撰寫



- 文檔摘要
- 報告生成
- 營銷文案、博客、社交媒體寫作等.

可控文本生成

AI內容生成的企業級應用

使用通用大模型 (e.g. ChatGPT) 的一些思考

數據隱私與安全

信息泄露風險

輸出不准確 / 錯誤答案

- 幻覺
- 生成虛假信息
- 社會偏見

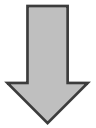
領域知識整合

- 如何將領域知識與模型相結合以控制其結果生成；
- 避免在專業領域產生錯誤的答案

AI 合規性 / 模型透明度

人工智能使用監管要求（例如金融領域）

企業需求：高度可控性，高可解釋性



需要**私有托管的本地**大模型，服務於特定領域

- 融合**領域知識、AI智能**和**人類反饋**
- 交互式**人機共創**
- **可控可解釋**

定制解決方案
服務於個性化和專業化的目的

ASTRI

ASTRI 生成式AI 解決方案概述

ChatDoc: 企業/領域 知識管理系統

- 基於對話的信息查詢
- ChatBI: 基於自然語言的數據分析
- AI 智能體: 流程自動化，風險管理，合規審查

客戶服務

本地化大模型：特定語言（粵語）及文化知識

- 聊天機器人
- 虛擬數字人

多模態內容生成：文本/圖像/視頻/語音

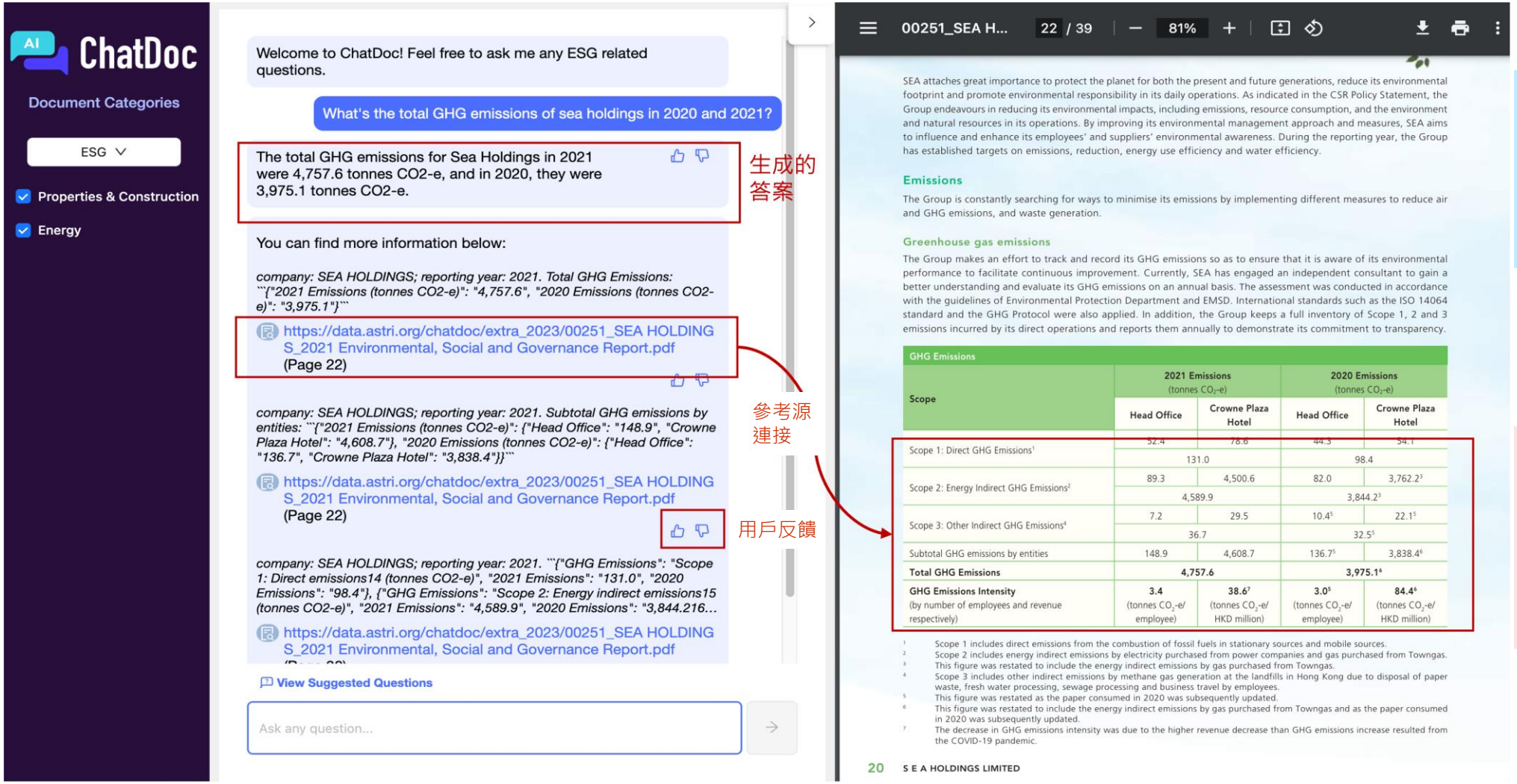
- 文檔/報告審核, 改寫, 生成
- 智能營銷（產品推廣、網站、社交媒體等）
  - AI海報生成
  - AI視頻生成
- 語音生成

風險管理

- 版權管理
- 欺詐檢測
- 隱私保護

ASTRI

ChatDoC = 文檔智能 + “ChatGPT” 生成式問答



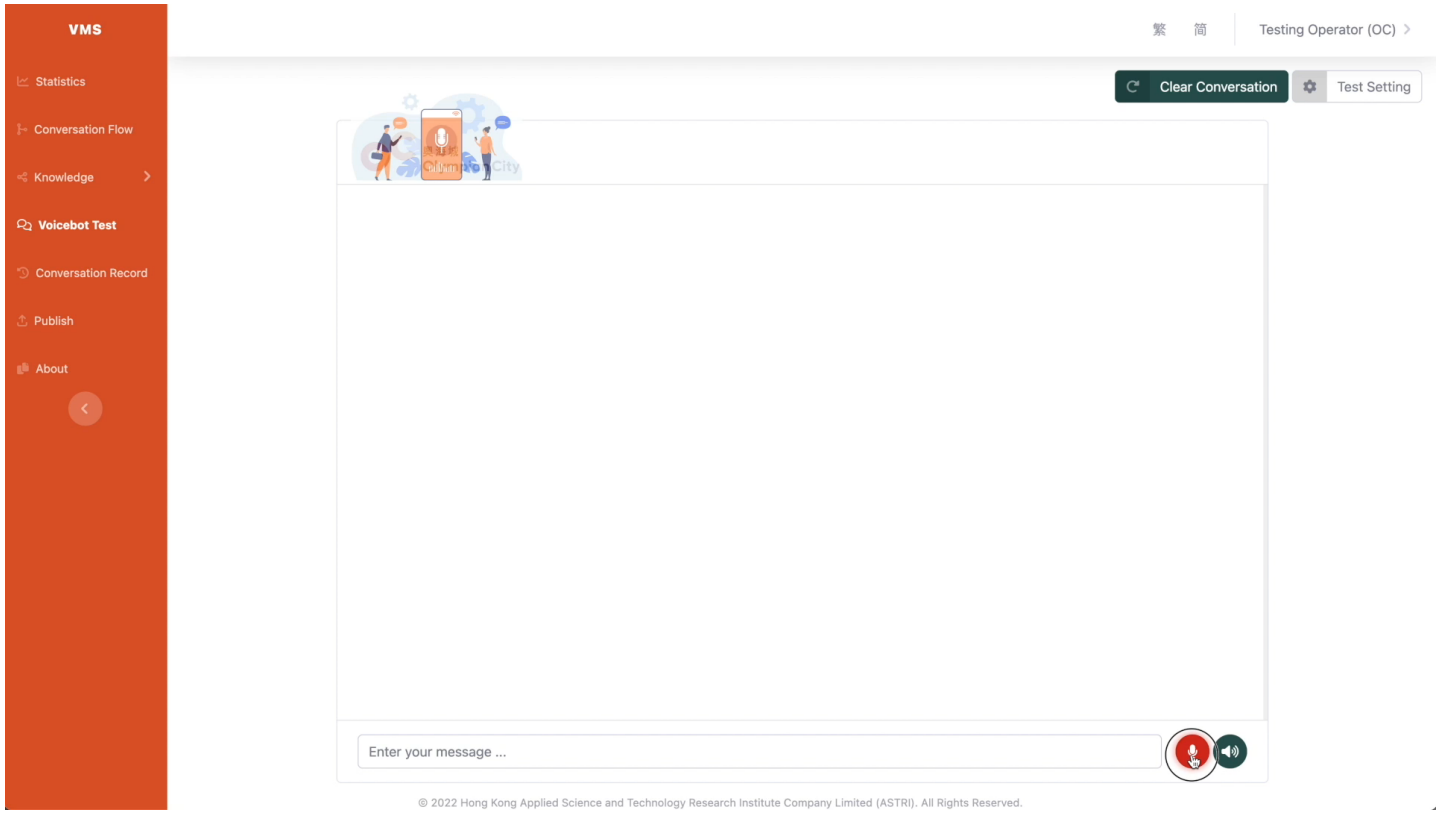
- 企業/領域專屬知識庫
- 通過 NLP 和對話查詢相關信息
- 提供參考源鏈接

- 減輕幻覺
- 避免事實錯誤
- 更準確、可驗證
- 知識更新快速簡單

垂直領域及企業解決方案： ChatESG / ChatLAW / ChatCorp / …

虛擬人客戶服務

- 提供語音智能對話系統voicebot
- 提供虛擬人視頻自動生成平台 (text-to-video)



voicebot

Avatar video generation

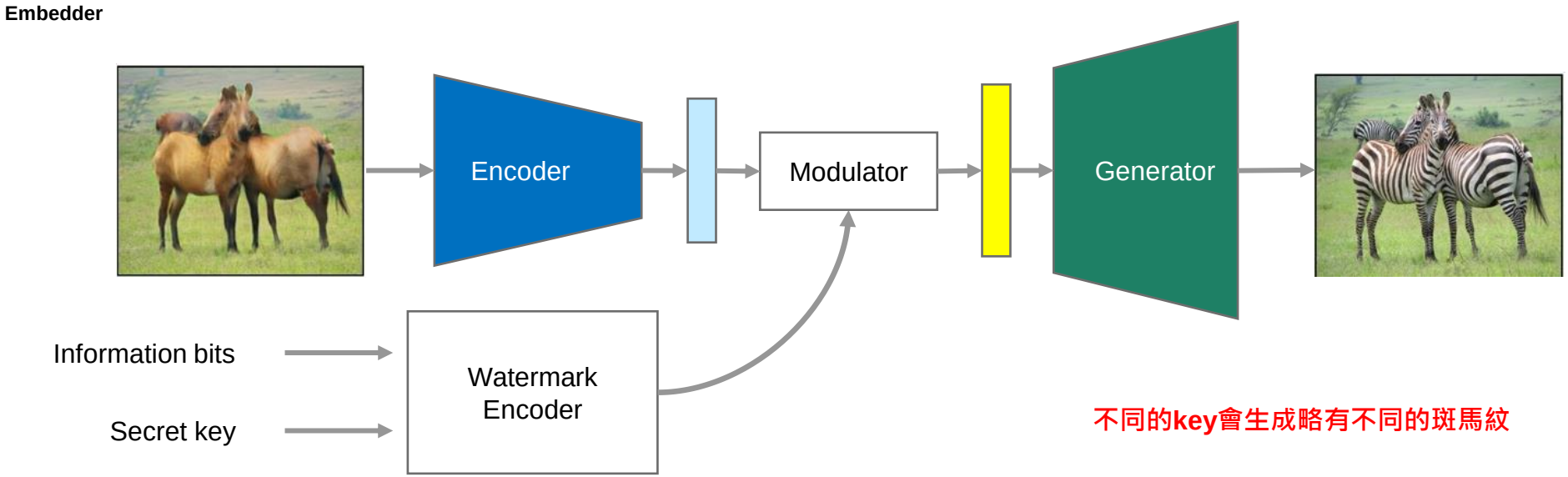
ASTRI

AI 版權管理

人工智能內容生成的水印系統

- 標記內容為AI生成，標記其所有權
- 通過對AIGC生成的內容進行特征修改來添加隱藏式水印
- 傳播水印是可擴展的——從單個 GAN 到多個級聯 GAN
- 特征難以被壓縮、RST等常規攻擊修改，相比傳統噪聲方法， 穩健性更高

GAN 生成內容的隱形水印示例



總結

平衡創新和隱私對於人工智能開發至關重要。加密和聯邦學習等先進技術提供了有前景的解決方案。優先考慮這兩方面使我們能夠創建值得信賴的人工智能系統，推動社會發展，同時保護個人資料。

法規合規性

- 遵守特定行業的法規，例如個人資料（私隱）條例（PDPO）
- 遵守六項保障資料原則 (six DPPs)
- 了解法規和標準的最新動態

增強隱私

- 採用隱私增強技術 (PET)，例如差分隱私
- 制定風險緩解策略和人工智能事件響應計劃
- 定期檢討和重新評估風險

應對AI隱私問題的挑戰

保護個人數據

- 確保僅收集和使用足夠但不過多的個人數據
- 正確記錄資料的處理
- 六項保障資料原則中指定的其他項目

透明度和準確性的權衡

- 與持份者進行有效溝通與互動
- 讓人工智能的決策和產生變得可解釋

ASTRI

謝謝

www.astri.org | TECH FOR IMPACT

姓名　　: Dr. Arvin TANG, 鄧羽真博士

電郵　　: arvintang@astri.org

電話　　: (852) 3406 0342

免責聲明

本演示文稿中包含的資料只供閣下參考，如日後有所改動，恕不另行通知。

該等資料的真實性、準確性和完整性並未得到保證，亦可能未包含有關香港應用科技研究院有限公司及/或其相關聯公司（統稱 "應科院"）的所有重要資料。 應科院對其所載的任何資料的真實性、準確性或完整性不作出陳述或保證，並且不承擔任何責任。

此外，該等資料可能包含預測和前瞻性的聲明，它們可能反映應科院對未來事件和財務表現的當前觀點。 該等觀點基於當前的假設，該等假設亦可能會隨著時間而改變。 應科院對於該等未來事件是否會發生、該等預測是否會被實現、或應科院的假設是否正確不作任何保證。

最後，本演示文稿不構成應科院的任何要約（包括與應科院的技術及/或服務有關的要約）。

ASTRI