

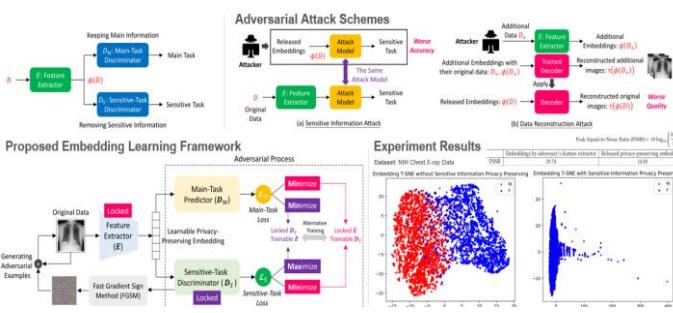
# 子計畫一：具可信度與可靠度的次世代AI客服 (PI: 林守德教授, co-PI: 張智星教授, 陳縕儂教授)

本子計畫研究次世代AI客服應具備幾項要素：1. 系統的隱私保護能力（左）2. 系統解釋推論的能力（中）3. 對於潛在攻擊的偵測（右）

## Learning Privacy Preserving Embeddings for Image Data

Target: image D that is going to be published, along with its label  $\rightarrow$  release privacy-preserving embeddings, denoted by  $\varphi(D)$ .

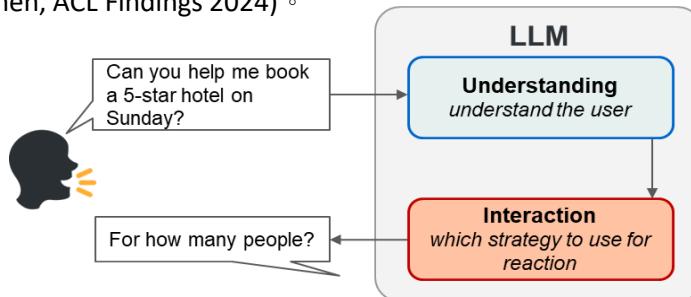
Goal: prevent the original image data and its corresponding sensitive information hidden behind from being unveiled by the adversary.



Li, Chu-Chen, Cheng-Te Li, and Shou-De Lin.  
"Learning Privacy-Preserving Embeddings for Image Data to Be Published." ACM Transactions on Intelligent Systems and Technology 14.6 (2023)

## Injecting Expert Strategy for Explainability

**Goal:** 在LLM的對話互動中引入專家的互動策略 (理解使用者與決定互動方式), 開發者可理解模型為何選擇此互動方式，上升可解釋性與可控性 (Chang & Chen, ACL Findings 2024)。



A: Actually, I just finished a mystery novel that had a great plot twist. It's one of my favorite genres. Do you read mystery books?

U: Sometimes, but mostly I prefer sci-fi and fantasy. Have you read any good books in those genres?

A: (The user implicitly mentioned the intent of FindRestaurants. I should smoothly pivot the conversation to FindRestaurants topic)

Speaking of sci-fi and fantasy, have you ever been to a themed restaurant?

U: No, I haven't. Is that a thing?

A: (The user did not change the topic of FindRestaurants. I should continue the topic)

Yes! There are restaurants that are decorated to look like settings from popular sci-fi and fantasy movies and TV shows.

U: That sounds really cool! Do you know of any nearby? ...

### Contributions

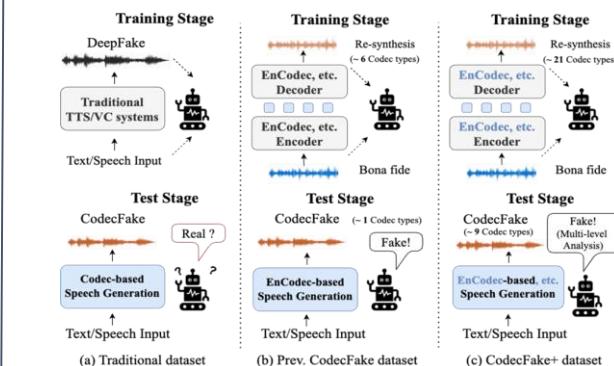
- 在多種不同使用者的模擬對話中，皆可透過此技術達成更自然也更好的行銷對話手法。
- 提出之SalesBot 2.0資料集與訓練之模型，皆開源於GitHub供後續研究使用。

Wen-Yu Chang and Yun-Nung Chen, "Injecting Salesperson's Dialogue Strategies in Large Language Models with Chain-of-Thought Reasoning," in *Findings of the Association for Computational Linguistics: ACL 2024 (Findings of ACL 2024)*, pages 3798–3812, Bangkok, Thailand, August 11–16, 2024. ACL.

## Towards Generalized Audio Deepfake Detection

**Background:** The emerging LLM-based speech generation method poses a great threat to the traditional speech deepfake detection model.

**Goal:** Propose a general method for detecting and understanding audio generated by LLM-based speech generation methods.



### Contributions

- Propose a large-scale neural audio codec-based deepfake speech dataset
- Propose a generalized source tracing framework based on codec taxonomy

Xuanjun Chen, et al., "CodecFake+: A Large-Scale Neural Audio Codec-Based Deepfake Speech Dataset," Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

Xuanjun Chen, et al., "Codec-Based Deepfake Source Tracing via Neural Audio Codec Taxonomy," INTERSPEECH, 2025.

Xuanjun Chen, et al., "Towards Generalized Source Tracing for Codec-Based Deepfake Speech," Submitted to ASRU 2025.