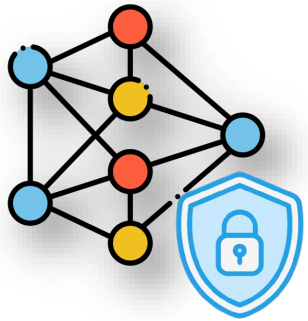
Cryptography Meets AI

Mingxun Zhou

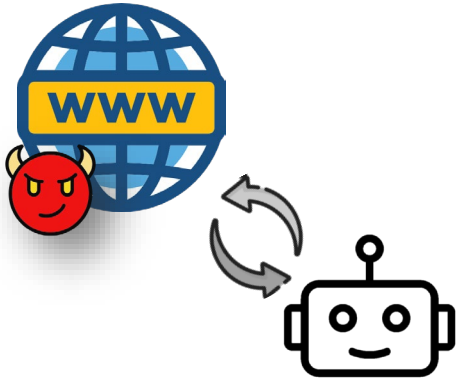
CSE, HKUST

Many New Challenges in Generative / Agentic AI Era



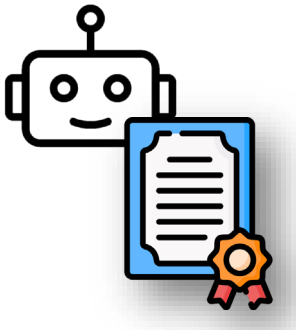
Core Model Security

- Instruction/Data Confusion
- Hallucination
- Jailbreaking
- Adversarial Robustness
- ...



Privacy & Data Security

- Information Leakage
- Credential Compromise
- Phishing Attack
- Control Hijacking
- ...



Trustworthy Output

- Harmful Content Regulation
- AI Output Sourcing
- Fact Check
- Copyright Enforcement
- ...

Many New Challenges in Generative / Agentic AI Era

Our Lab's Active Research:

Responsible AI Ecosystem with Cryptography Principles

- *Provable **Security & Privacy***
- ***Verifiability** by Design*
- *Exploring **Fundamental Limitations***

Our Lab's Active Research:
Responsible AI Ecosystem with Cryptography Principles

Core Model Security

- **Hallucination Reduction**
  - **Post Training w/ Proof System** (on going)
- **Anti-Prompt Injection**
  - **Message-Authentication-Code based Solution** (on-going)

Privacy & Data Security

- **Zero-Leak External Info Access**
  - **Piano PIR** *(S&P 2024),* **Quarter PIR** *(Eurocrypt 2025)* **Pacmann** *(ICLR 2025)*
  - 3 more in top venues
- **Private Information Processing**
  - **Misbehavior Tracing** *(CCS 2025)*
  - **Differential Obliviousness** *(Eurocrypt 2024, ITCS 2025)*

Trustworthy Output

- **LLM Watermarking / Fingerprinting**
  - **PMark** *(ICLR 2026)*
  - 2 in submissions
- **Human / LLM Output Classification**
  - **Fundamental Limitation of Supervised Detector** (in submission)
  - **Proof of originality** (2 projects on-going)

Our Lab's Active Research:
Responsible AI Ecosystem with Cryptography Principles

Core Model Security

- **Hallucination Reduction**
  - **Post Training w/ Proof System** (on going)
- **Anti-Prompt Injection**
  - **Message-Authentication-Code based Solution** (on-going)

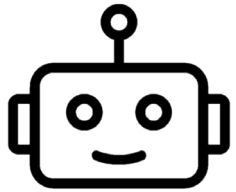
Privacy & Data Security

- **Zero-Leak External Info Access**
  - **Piano PIR** *(S&P 2024),* **Quarter PIR** *(Eurocrypt 2025)* **Pacmann** *(ICLR 2025)*
  - 3 more in top venues
- **Private Information Processing**
  - **Misbehavior Tracing** *(CCS 2025)*
  - **Differential Obliviousness** *(Eurocrypt 2024, ITCS 2025)*

Trustworthy Output

- **LLM Watermarking / Fingerprinting**
  - **PMark** *(ICLR 2026)*
  - 2 in submissions
- **Human / LLM Output Classification**
  - **Fundamental Limitation of Supervised Detector** (in submission)
  - **Proof of originality** (2 projects on-going)

External Information Access Leaks Privacy



Agent

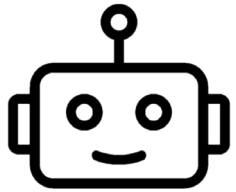
"Tylenol safety during pregnancy"

"Generally safe…"

Got your medical info!

Private Information Retrieval (PIR) [CGKS95]

Provably Zero-Leak Access



Agent

"Tylenol safety during pregnancy"

"Generally safe…"

???

Our Vision:

“PIR as a zero-leak external knowledge connector for agents”

2022	2023	2024	2025	2026
Theoretical Sublinear 1-server PIR Eurocrypt '23 First theoretical construction	**Piano PIR** S&P '24 **1st gen** 100x faster	**QuarterPIR** **Pacmann** Eurocrypt '24 · ICLR '25 **2nd gen** **50x less comm** Private Vector DB	**Zelda PIR** **GZSP PIR** S&P '25 · Eurocrypt '25 **3rd gen** Unconditional Security	**Primus** In submission **4th gen** **40x less storage** Optimal + practical

Primus: 4th Gen Practical PIR

Hyper-competitive Performance

DATABASE SIZE

2 Billion

Scalable design

SETUP TIME

~1 Hour

Ready to Go

QUERY TIME

<10 ms

Blazing Fast

COMMUNICATION

<0.5 MB

Per Query

EXTRA STORAGE

~50 MB

Lightweight

Pacmann: Private Vector-Based Searching Engine

ICLR 2025

VECTOR SPACE

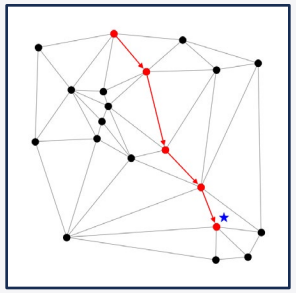
100 Million

128-dim

SETUP TIME

~6 Hours

Preprocessing



Interactive Travel

Graph Structure

TOTAL QUERY TIME

3.0 Seconds

Fast Retrieval

RECALL

>90%

recall@10

PLUG & PLAY

RAG, AI Agents

& Beyond

Many Interesting Projects Brewing :)

Self-Anchored LLM Watermarking

- First to achieve >85% detection under full-rewriting attack

Black-Box LLM Identity Discovery

- Near-perfect detection with <1% false positive rate

Proof of Image Originality

- Zero-knowledge proof techniques, even with editing

mingxunz@ust.hk

Welcome to collaborate!