DSATM

Dialogue Security Anti-Tampering Mechanism

網路與多媒體實驗 第四組
吳名洋 B09901045
林穎軒 B11508006
陳嘉謙 B11504045

MOTIVATION

In recent years, generative AI has been used by an increasing number of people.

However, occasional instances malicious tampering with AI or human responses may occur. Yet, there is no method to ensure the integrity of conversation. This could make clarifying liability difficult.

Moreover, AI and user data are usually stored in central web server. This model makes it prone to cyber attack, resulting in leak of sensitive info.

INTRODUCTION

We create a tamper-proof mechanism for human-AI dialogue using serialized encryption and signature, which we drew inspiration from blockchain technology .

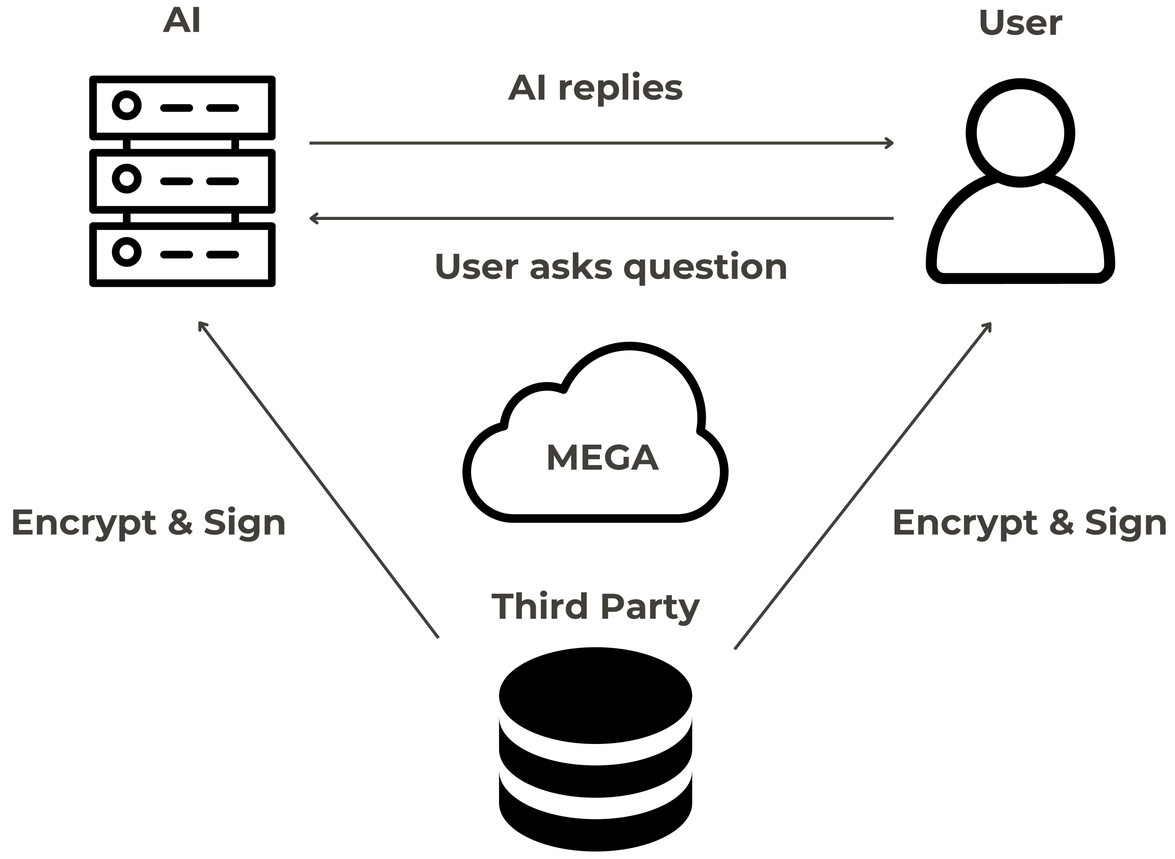
Both side's prompts are hashed. Then encrypted by key of third-party and appended onto the other's prompts as signature, and so on. This mechanism have 3 major advantages:

- Integrity of dialogue record
- Secrecy of conversation
- Tampering source traceability

TECHNOLOGIES

Platform: MEGA
Generative AI: Llama3 (Meta)
Cryptography: RSA

TECHNICAL STRUCTURE



MECHANISM