

*To: Paula Murphy
Lancaster, PA
Lancaster, PA*



Model-Side Confidential Inference

Leveraging OpenSSL for End-to-End Encrypted AI Inference Pipelines

Tarique Aman Aziz
Manager, Software Engineering
Data and AI, Red Hat

Navinya Shende
Senior Principal Software
Engineer
Data and AI, Red Hat

AI 101

1

Generative AI Shift

2

What is AI inference?

3

What are AI Agents and MCP?

TLS and at-rest encryption aren't enough!

Prompts still leak inside the AI pipeline.

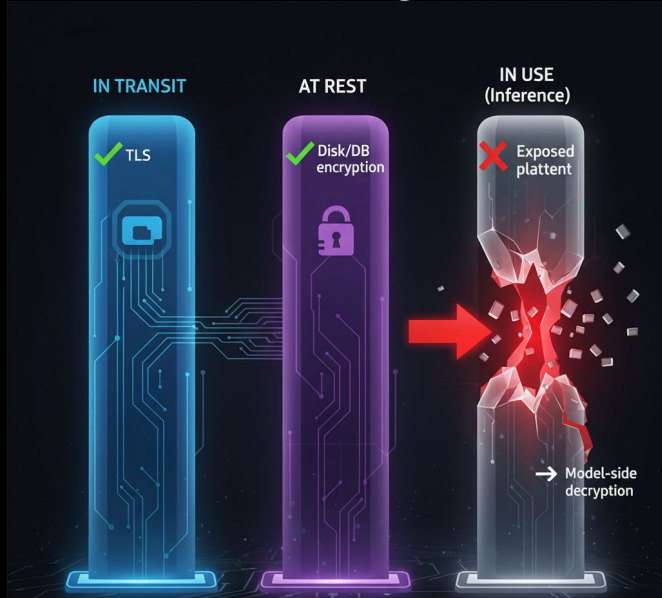
- Patient medical records in prompt
- Unreleased earnings data
- Confidential contracts or code
- Sensitive intelligence data

Prompts plaintext exposure

- API gateway
- Reverse proxy
- Logs and monitoring systems
- Agent middleware
- Retrievers
- Inside AI Inference Runtimes

TLS protects in-flight traffic (✓)
Disk encryption protects storage (✓)
Inside the AI stack → plaintext everywhere (✗)

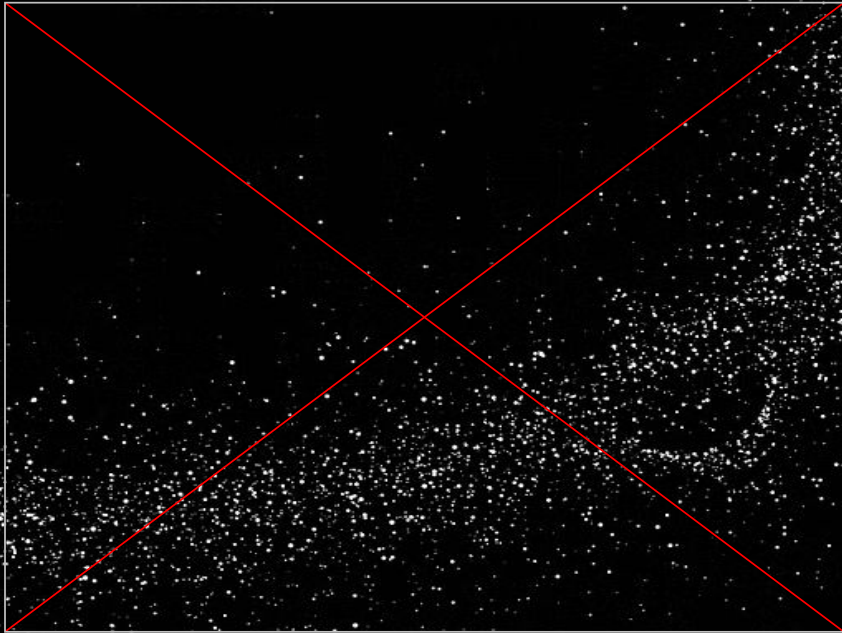
Why Existing Security Isn't Enough?



3 MIN • PRIVACY & COMPLIANCE

ChatGPT chats 'leaked' in Google search results

The Idea: Model-Side Confidential Inference



Concept

- User/client encrypts prompt with OpenSSL.
- Encrypted prompt travels across network + orchestration stack.
- Only decrypted inside model runtime memory.
- Model tokenizes → processes → generates output.
- Response re-encrypted before leaving runtime.

Why OpenSSL?

1



Widely available across runtimes and integrates with C, Python, Rust, Java

2



Algorithms we can use:
AES-256-CBC, AES-GCM

3



Standards-based,
trusted, interoperable

4



Future-proof

Demo

Future Directions

- Full homomorphic encryption
 - MCP Layer encryption
-

Code Repository



Connect With Us



Tarique Aman Aziz



Navinya Shende
