Inpher

SecurAI

Enhancing Large Language Model Inference with Confidential Computing

Executive Summary

In the rapidly evolving landscape of Generative Artificial Intelligence (AI) organizations are exploring applications to enhance productivity and unlock substantial business benefits. However, utilizing AI for applications like code development, content creation, anomaly detection, automation, healthcare analytics or personalization often involves handling sensitive data and intellectual property. The visibility of this data specifically through prompts and completions shared with a model service provider raises serious governance concerns, often hindering organizations from fully leveraging AI capabilities.

In this paper, we will look at the advantages of using generative AI, in this case ChatGPT or other Large Language Models (LLMs), ethically and responsibly by leveraging Trusted Execution Environments (TEEs) and Inpher SecurAI. Inpher SecurAI ensures that both the prompt and the completion are secured during model inference, thus enabling large organizations to solve the sensitive data challenge.

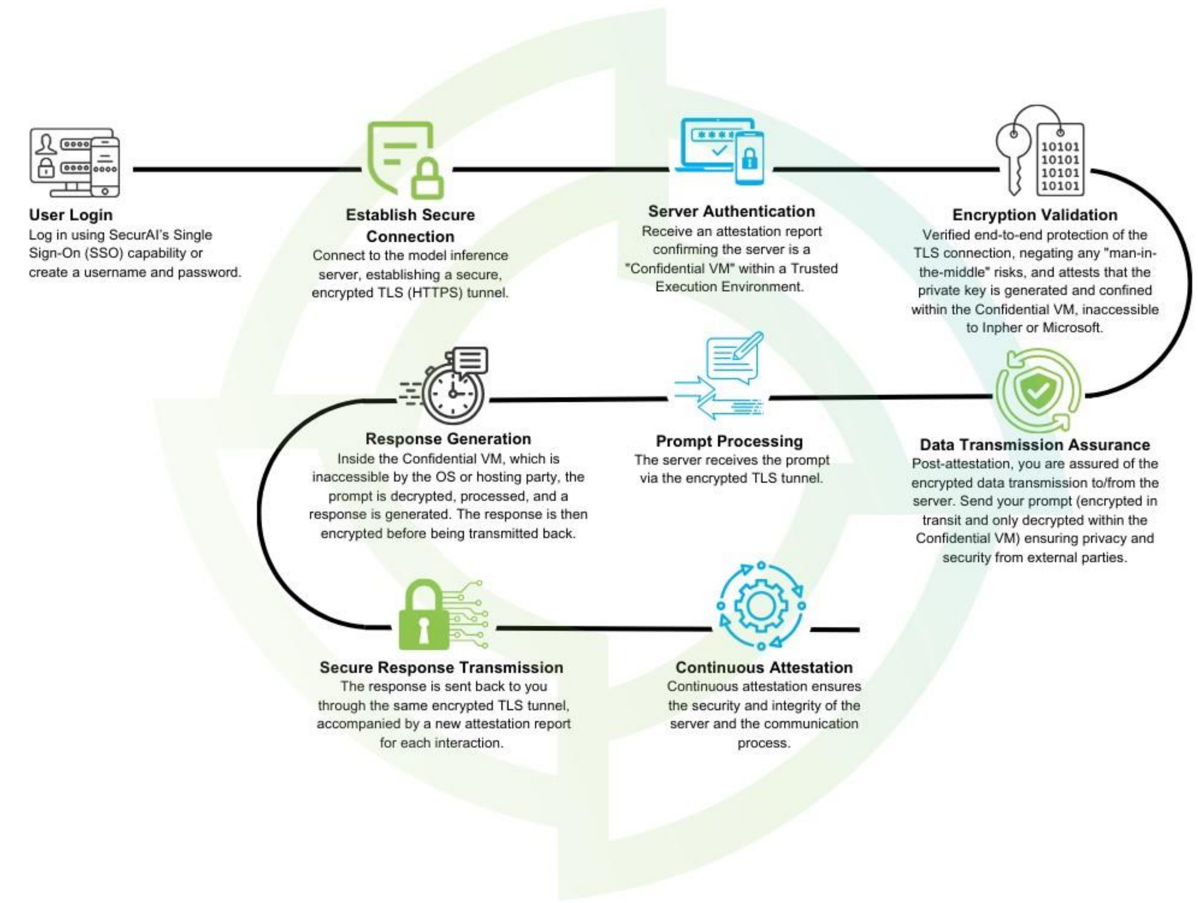
Inpher SecurAI and Trusted Execution Environments

Our approach utilizes TEEs that protect virtual machines (VMs) from various threats. We employ cutting edge techniques like memory encryption, secure boot, secure nested paging, secure encapsulation and remote attestation. Hosted on Microsoft Azure, our service utilizes Azure Confidential Computing, today based on AMD SEV-SNP technology, reinforcing our commitment to robust security and a roadmap to continue pushing the edge of security and utility.

© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

Operating Model and Trusted Components

In this section, we will describe the operating model and trusted components of SecurAI. First, we provide an overview of the confidential computing framework. Next, we compare the two types of TEEs, including VM-level TEEs, which SecurAI employs. Then, we briefly explore Trusted Platform Modules, our AMD SEV-SNP Architecture, the Intel Trust Domain Extension, and Microsoft's Azure Confidential Computing. Finally, we define Remote Attestation and then explain the process of Attested Transport Layer Security as well as Guest Attestation with Microsoft Azure Attestation Services, used by SecurAI.



User Login
Log in using SecurAI's Single Sign-On (SSO) capability or create a username and password.

Establish Secure Connection
Connect to the model inference server, establishing a secure, encrypted TLS (HTTPS) tunnel.

Server Authentication
Receive an attestation report confirming the server is a "Confidential VM" within a Trusted Execution Environment.

Encryption Validation
Verified end-to-end protection of the TLS connection, negating any "man-in-the-middle" risks, and attests that the private key is generated and confined within the Confidential VM, inaccessible to Inpher or Microsoft.

Data Transmission Assurance
Post-attestation, you are assured of the encrypted data transmission to/from the server. Send your prompt (encrypted in transit and only decrypted within the Confidential VM) ensuring privacy and security from external parties.

Prompt Processing
The server receives the prompt via the encrypted TLS tunnel.

Response Generation
Inside the Confidential VM, which is inaccessible by the OS or hosting party, the prompt is decrypted, processed, and a response is generated. The response is then encrypted before being transmitted back.

Secure Response Transmission
The response is sent back to you through the same encrypted TLS tunnel, accompanied by a new attestation report for each interaction.

Continuous Attestation
Continuous attestation ensures the security and integrity of the server and the communication process.

SecurAI workflow

© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

1. Understanding Confidential Computing

Confidential computing is a security framework designed to protect sensitive data during processing within applications, servers, or cloud environments.

Key features of confidential computing include *secure boot* (the system boots into a defined and trusted configuration), *curtained memory* (memory that cannot be accessed by other OS processes), *sealed storage* (software keeps cryptographically secure secrets), *secure I/O* (prevents keystroke logger attacks), *integrity measurements* (computing hashes and fingerprints of executable code, configuration data and other system state information) and other vital safeguards.

2. VM-level vs. Processor-level Trusted Execution Environments

There are two major types of TEEs: processor-level TEEs and VM-level TEEs.

2.1. Processor-level TEEs

Processor-level TEEs integrate security features directly in the hardware. They are advantageous as they provide fine-grained security control by isolating individual processes directly on the hardware level. However, they pose challenges in compatibility and application development. A typical example of a processor-level TEE is Intel Software Guard Extensions (SGX).

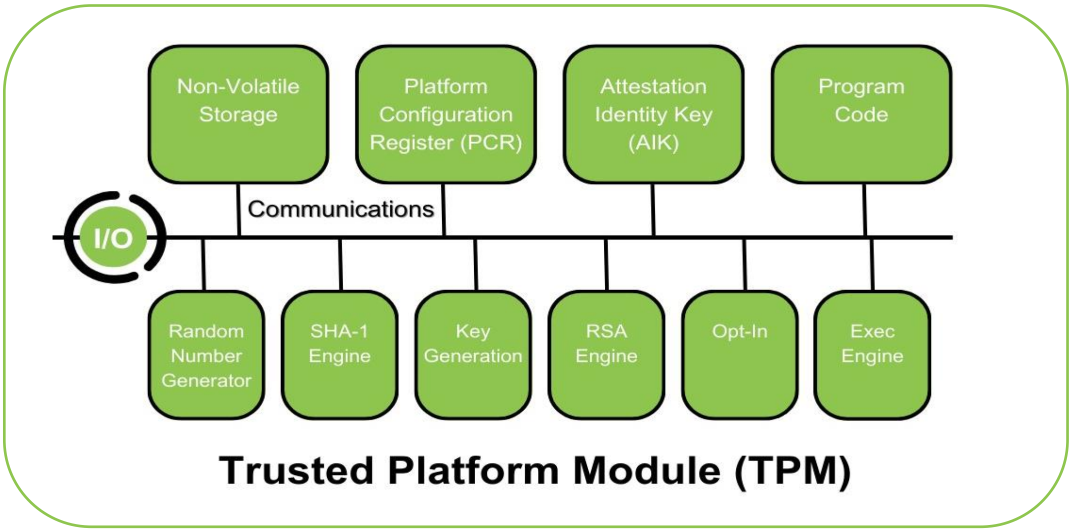
2.2. VM-level TEEs

VM-level TEEs offer broader security control by isolating entire VMs. Hypervisors create and manage isolated VMs operating independently of each other and enforce security by ensuring that VMs cannot interfere with each other and cannot read each other's data. Virtualization overcomes some of the compatibility and application development challenges present for processor-level TEEs. Examples of VM-level TEEs are AMD Secure Encrypted Virtualization with Secure Nested Paging (SEV-SNP) Architecture and Intel Trusted Domain Extensions (Intel TDX).

© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

3. Platform Modules

A Trusted Platform Module (TPM) is an international standard for a secure cryptoprocessor or a microcontroller that secures hardware through an embedded (integrated) cryptographic key. The term TPM is interchangeably used for any chip conforming to these standards. The typical TPM architecture includes non-volatile storage, platform configuration registers (PCRs), attestation identity keys and basic cryptographic functionality for random number generation, key generation, hashing and support for basic encryption and digital signature algorithms.



© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

4. AMD Secure Encrypted Virtualization with Secure Nested Paging Architecture

AMD’s SEV architecture was the first x86 architecture allowing for memory isolation of the virtual machine from the hypervisor. Before its introduction, even if a hypervisor was not able to read VM memory, it was still able to see values in the CPU registers during computation, thus allowing for possible sensitive data identification, vulnerabilities and leaks during computation. The subsequent SEV-ES technology (SEV with Encrypted State) addressed this problem by encrypting the values of the registers during computation, thus preventing the hypervisor from seeing the data actively used by the VM. The latest SEV-SNP technology provides additional integrity guarantees: namely, if a VM is able to read an encrypted memory page, it must read the value it last wrote. It allows VMs to run in a secure enclave known as SNP while protecting them from other VMs and the hypervisor itself. Compared to SEV-ES, SEV-SNP provides strong guarantees against data corruption, re-mapping and aliasing-based attacks.

5. Intel Trust Domain Extension

Similar to AMD SEV-SNP, Intel TDX (TDX) is a technology developed by Intel for Intel CPUs that further protects the entire hypervisor and VMs it manages by encrypting the entire memory space with encryption keys directly managed by the hardware. While SecurAI is currently based on AMD SEV-SNP, the underlying architecture is exchangeable, so Intel TDX is also an option when preferred.

Unlike AMD SEV-SNP, where the hypervisor is trusted and the security model does not protect against malicious hypervisors, Intel TDX encrypts memory spaces from the hypervisor, providing a higher-level of protection against malicious threats from both internal and external sources.

6. Microsoft Azure Confidential Computing Services

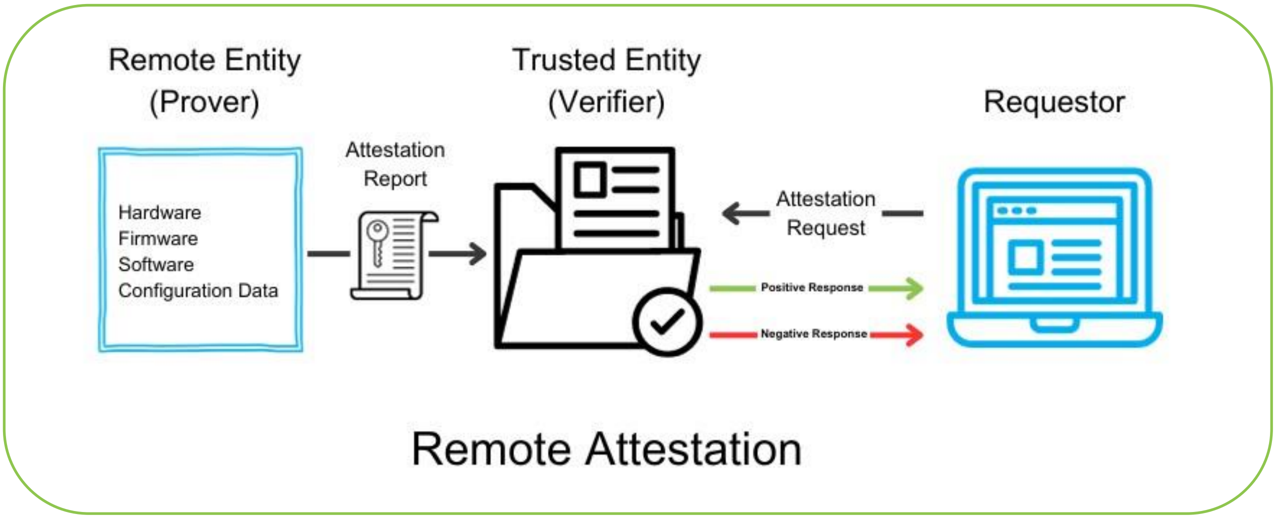
Employed by SecurAI, Microsoft Azure Confidential Computing (ACC) services is one of a growing list of Cloud Service Providers (CSPs) to support TPM utilization and offering confidential Virtual Machines (VMs) running on AMD SEV-SNP hardware with forthcoming support for Intel TDX, NVIDIA H100 confidential GPUs and more.

© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

7. Remote Attestation

Remote Attestation (RA) in confidential computing allows distributed systems to validate each other's trustworthiness through a security process where one system provides reliable statements to another system about the software it is running. The RA process allows for a trusted entity or a verifier to then assess the trustworthiness of a remote entity (a prover) and validate its claims.

In an RA workflow, the prover generates a set of measurements about its current state (i.e., measurements about its hardware, firmware, software or configuration data). These measurements are then encapsulated into an attestation report that is cryptographically signed by a key embedded into the hardware by the original hardware manufacturer (e.g., AMD, Intel or NVIDIA). A separate system, the verifier, issues an attestation request that prompts the prover to prove its trustworthiness by providing the generated evidence. After the prover communicates the evidence to the verifier, the verifier validates a digital signature on the attestation report and checks that the measurements are consistent with a predefined good state of configuration. The verifier can then make a trust decision by checking that the prover's measurements align with the verifier's expectations. In the case of discrepancies, the prover is considered untrustworthy. A relying party in this process would be a customer application that is accessing the prover.



© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

8. Attested Transport Layer Security

Attested Transport Layer Security (TLS) ensures that data is securely communicated via protocol between the user and the model inference server (running on a TEE) such that plaintext data is provably not visible to either the system administrator of the TEE or to any other external party (including Inpher).

Inpher uses a trusted platform module (TPM) to generate and store a fingerprint of the TLS certificate so that during remote attestation, it takes a measurement of the fingerprint, signs it with the hardware-embedded key, and provides the attestation report to the verifier.

An attestation report can be verified either locally (by the system administrator of the enterprise user of the secure inference server) or using an attestation service such as Microsoft Azure Attestation service.

The following section provides an overview of the technologies and cryptographic methods needed for attested TLS.

8.1 Platform Configuration Registers (PCRs)

Platform Configuration Registers (PCRs) are used by a TPM to store hashes of measurements of system components in order to verify the integrity of the system. PCR values are used in remote attestation reports to allow a remote party to verify that the platform has not been tampered with.

PCRs are ordered in an array whose order and size may depend on the TPM specification. During a system startup, PCR registers are initialized with known values to represent a clean state. Platform components such as the Basic Input/Output System (BIOS), bootloader, firmware and other critical software and hardware components extend their own hashes into designated PCRs.

Measurements of platform components are extended into PCR registers using TPM-specific commands. Extending a measurement into a PCR register hashes the new measurements as well as the existing value of the PCR to ensure that the PCR retains the history of the measurements. More precisely, a caller cannot directly overwrite a PCR value, but can update a PCR extension with the below logic, ensuring a PCR represents the software stat as well as its history.

```
PCR new value = Digest of (PCR old value || data to extend)
```

© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

A typical PCR allocation is demonstrated in the table below:

PCR Number	Allocation
0	BIOS
1	BIOS Configuration
2	Option ROMs
3	Option ROM Configuration
4	Master Boot Record
5	Master Boot Record Configuration
6	State Transition and Wake Events
7	Platform Manufacturer Measurements
8-15	Static Operating System
23	Application Support

Inpher’s specific approach to attested TLS is to extend PCR with the hash (fingerprint) of the TLS certificate generated on the TPM. A remote attestation report signed with the hardware-embedded secret key then includes the measurement of that register.

8.2 Fingerprints for TLS Certificates

Inpher uses the sha256 hash of a server-generated TLS certificate (a .pem file) to get a fingerprint of the certificate to be used for the attestation. This fingerprint is then passed as an input to a Guest attestation C++ library described in the next section.

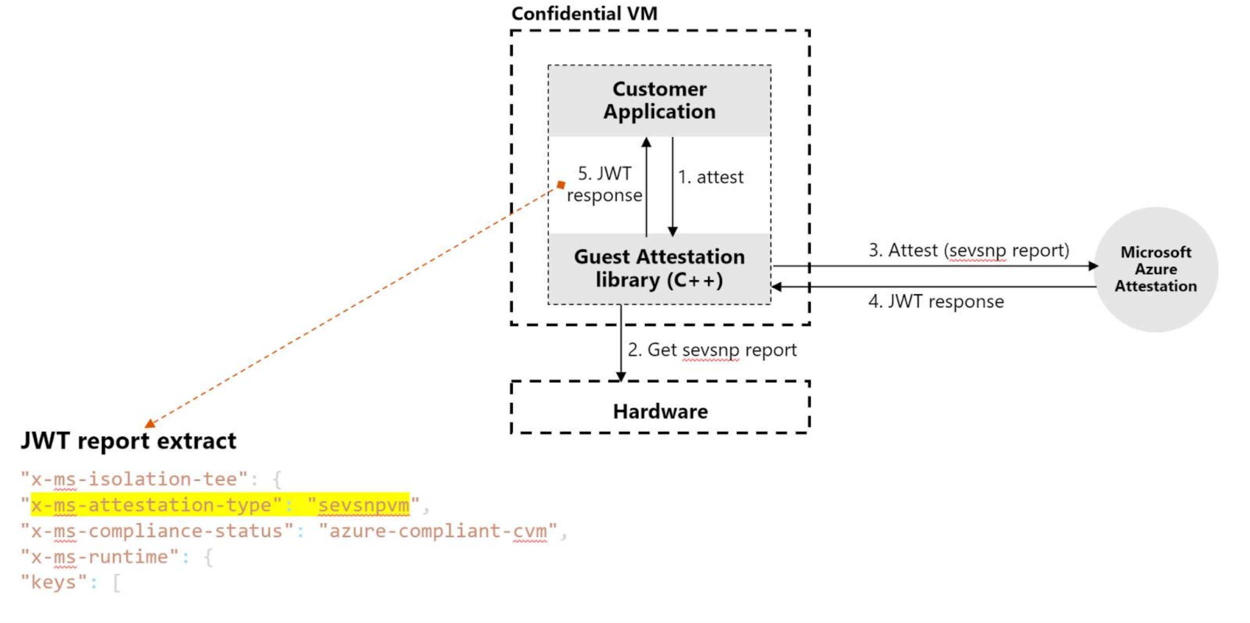
© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

9. Guest Attestation and Microsoft Azure Attestation Service

Inpher uses Microsoft's Guest Attestation for Confidential VMs. Guest Attestation (GA) is the process of confirming that a confidential VM runs on an expected trusted hardware platform (i.e. AMD SEV-SNP) and that certain security features are enabled for isolation and integrity. In particular, GA checks that a confidential VM has an enabled secure boot which protects lower layers of the VM such as the firmware, the boot loader and the kernel from malware (rootkits and bootkits).

Inpher's use of GA leverages the Microsoft Azure Attestation Service for remote attestation and use of TEEs such as Intel SGX, Virtualization-based security (VBS) enclaves, TPMs, Trusted launch for Azure VMs and Azure confidential VMs. Microsoft provides a GA library to support generating these reports; sample use cases with this library are available here.

Binaries built on top of the Azure `AttestationClient` library takes as input a cryptographic nonce, generates the report on the TPM of the Azure confidential VM, signs the report with the attestation key and returns the signed report as a Json Web Token (JWT) string. The typical attestation workflow is described in the figure below:



```
"x-ms-isolation-tee": {
"x-ms-attestation-type": "sevsnpvm",
"x-ms-compliance-status": "azure-compliant-cvm",
"x-ms-runtime": {
"keys": [
```

© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

When a customer application attestation request is served by the Microsoft GA, the C++ library uses the standard TPM 2.0 API to generate the SEV-SNP report with measurements from the PCR registers. The report is signed with the hardware-embedded key. This SEV-SNP report is then sent to Microsoft Azure Attestation service for verification. The resulting JWT response is communicated back to the customer application via the GA library. In this scenario, the Microsoft Azure Attestation service is the verifier, the prover is the hardware, and the relying party is SecurAI, the customer application.

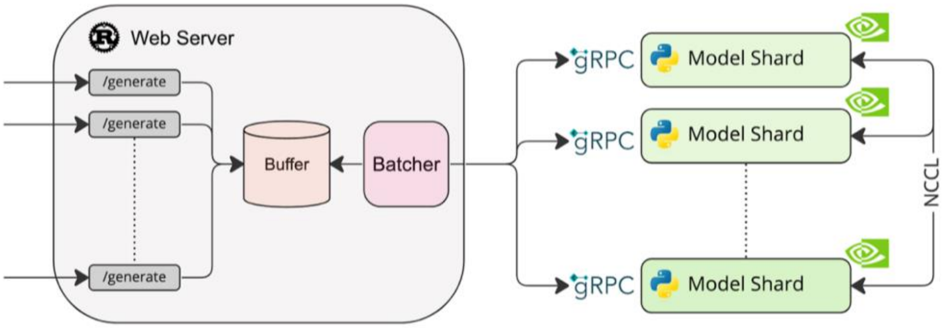
In order for Inpher to integrate and sign (by the TPM) the fingerprint of the TLS certificate, we pass an extended nonce that consists of the original user-provided nonce and the TLS certificate fingerprint.

Securing Prompt and Completion for LLM Inference

SecurAI consists of a front end (UI) and a backend (model inference server).

Inpher uses a standard chat UI front end from HuggingFace that can connect with the model inference server hosted on the same enclave (or separate enclave, as the model inference server and the chat UI could communicate via attested TLS). The main advantages of chat UI are authentication support and extensibility of the UI framework, as well as supporting multiple chats per user. The mongoDB database is used by the chat UI application for the persistent storage of the chat history. Since the database runs inside the confidential VM, it cannot be accessed by the cloud provider or Inpher.

There are different options for the model inference server depending on the underlying hardware or large language model. A typical option is a TGI – a Python, Rust, and gRPC text generation inference used by HuggingFace to power the Hugging Chat, the inference API and the inference endpoint. TGI is advantageous from various perspectives, such as GPU model support on a cluster of GPUs, CPU-mode only, and queuing as well as streaming tokens and model sharding, as illustrated in the diagram:



© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

Putting Everything Together

SecurAI combines all the concepts and components previously described into a verifiably attested and private LLM Inference solution. The chat UI, the model inference server and a nginx reverse proxy terminating the attested TLS connection are signed and dockerized and can be measured for integrity at boot time to ensure that a well-defined version of the SecurAI stack is running on the Microsoft Azure Confidential VM. At boot time, a TLS key pair is generated within the confidential VM and stored in a PCR which will be used to terminate the TLS connection with nginx.

Once completed, the chat UI is ready to receive user prompts. The user's browser establishes a TLS connection with the nginx reverse proxy and receives the TLS certificate along with the attestation report (verified by the MAA), where the TLS certificate fingerprint received by the browser can be compared with the fingerprint inside the attestation report to ensure that they match and that no interception of the TLS connection through a man-in-the-middle attack is possible. This establishes an attested TLS connection between the user's browser and the confidential VM.

The attestation report served by SecurAI and verified by MAA also contains other claims and measurements, such as the integrity of the software and the secure boot of the confidential VM. Each time the user submits a prompt, the prompt will be sent along with a nonce that will be used to generate a "fresh" attestation report to continuously monitor the integrity and confidentiality of the system. These attestation reports are then transmitted back to the user's browser where they are displayed in the chat UI front end.

© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).

Conclusion

With SecurAI, bringing ChatGPT and other LLMs into your organization can be achieved without compromising privacy. Grounded in the principles of Confidential Computing, SecurAI ensures that sensitive data and intellectual property are protected throughout the AI inference process. Employing TEEs, advanced encryption, and remote attestation creates a multi-layered defense that maintains integrity even in the face of potential threats. For organizations seeking to harness the predictive power of AI in a manner that aligns with ethical and regulatory demands, Inpher SecurAI can be your partner in navigating the complexities of AI security.

As AI technologies expand and improve at breathtaking speed, organizations need to know that the solutions they adopt today will evolve to meet new, unfolding needs. SecurAI has been developed with a vision for the future, including continuing to incorporate cutting-edge cryptographic techniques and ultimately supporting any infrastructure and any cloud environment.

Leverage LLMs and Generative AI Securely

Today we invite individuals and organizations to experience the security and efficiency of SecurAI. In an era where data breaches are costly and trust is hard-earned, partnering with Inpher means building your AI initiatives on a foundation of trust and reliability. Inpher SecurAI – a revolutionary approach to leveraging LLMs and generative AI privately, securely and with complete autonomy.

Try it for yourself today!

About Inpher

Inpher, Inc. is the leader in privacy-enhancing computation that empowers organizations to collaborate on sensitive data seamlessly and securely across teams and borders. Inpher's award-winning platform employs machine learning and AI in order to remove data barriers and silos while delivering the highest level of trust and precision in even the most complex data collaboration initiatives. Founded by world-renowned cryptographers and engineers, Inpher has long been recognized as a pioneer in the fields of secure Multiparty Computation (MPC), Fully Homomorphic Encryption (FHE), Federated Learning (FL), and other combinations of privacy-enhancing technologies (PETs). Inpher continues to deliver enterprise ready capabilities and real-world examples.

For more information on Inpher, please visit us at inpher.io and follow us on LinkedIn and Twitter.

© 2023 Inpher Inc. All rights reserved. Inpher, the Inpher logo and XOR are registered trademarks of Inpher, Inc. All other names or logos mentioned or used herein are for identification purposes only and may be the trademarks of their respective owner(s). Inpher, Inc. may not be associated with, or be sponsored or endorsed by, any such owner(s).