Privacy-Preserving LLM Inference with Hardware-Attested Trusted Execution Environments

Dzianis Vashchuk[1]

[1]Vibe Technologies, LLC, dzianisvv@gmail.com

January 2026

Abstract

We present an open-source infrastructure for deploying Large Language Model (LLM) inference within Trusted Execution Environments (TEEs) with cryptographic remote attestation. Our implementation runs self-hosted DeepSeek models on Azure Confidential VMs with Intel TDX, providing hardware-enforced memory encryption and verifiable privacy guarantees. We introduce a remote attestation API that enables clients to cryptographically verify TEE execution before submitting sensitive prompts. Our production deployment demonstrates practical feasibility with 12 tokens/second on CPU TEE and projects 150+ tokens/second on GPU TEE with NVIDIA H100 Confidential Computing. The complete infrastructure, including Terraform configurations and attestation services, is available at https://github.com/VibeTechnologies/TrustedGenAi.

1 Introduction

The widespread adoption of Large Language Models (LLMs) for sensitive applications—including browser automation, code generation, and document analysis—raises fundamental privacy concerns. Users must trust that their prompts are not logged, their data is not used for training, and their credentials remain confidential. Current cloud-hosted LLM APIs provide no cryptographic guarantees about data handling; users rely entirely on provider policies and legal agreements.

Trusted Execution Environments (TEEs) offer a hardware-based solution by providing isolated execution contexts where data remains encrypted even from the cloud operator. However, deploying production LLM inference within TEEs presents unique challenges: model size constraints, performance overhead, and the complexity of remote attestation for end users.

1.1 Contributions

We make the following contributions:

1. **Production TEE-LLM Infrastructure**: We deploy and validate an end-to-end LLM inference system on Azure Confidential VMs with Intel TDX, demonstrating practical feasibility for privacy-critical workloads.

2. **Remote Attestation API**: We implement a REST API that provides cryptographic proof of TEE execution, enabling clients to verify hardware isolation before submitting sensitive data.

3. **Open-Source Reference Implementation**: We release complete infrastructure code, including Terraform configurations, attestation services, and client integration examples.

4. **Performance Benchmarks**: We provide empirical measurements of inference performance on both CPU TEE (Intel TDX) and projections for GPU TEE (NVIDIA H100 Confidential Computing).

2 Background

2.1 Trusted Execution Environments

A Trusted Execution Environment is a secure, isolated processing environment that guarantees confidentiality and integrity of code and data. The Confidential Computing Consortium defines TEEs as hardware-based, attested environments that protect data in use [1].

2.1.1 Intel Trust Domain Extensions (TDX)

Intel TDX [2] provides hardware-isolated virtual machines called Trust Domains with the following properties:

- Memory encryption using CPU-managed keys (AES-256-XTS)
- Isolation from hypervisor, other VMs, and host OS
- Hardware-rooted remote attestation
- Minimal performance overhead (typically <5% for compute-bound workloads)

Azure offers Intel TDX via the DCesv5 VM series (e.g., Standard_DC4es_v5).

2.1.2 AMD Secure Encrypted Virtualization (SEV-SNP)

AMD SEV-SNP [3] provides similar guarantees with:

- Per-VM encryption keys managed by AMD Secure Processor
- Secure Nested Paging preventing hypervisor memory remapping attacks
- No guest modifications required (transparent to applications)

Azure offers AMD SEV-SNP via the DCasv5 series and NCCads_H100_v5 for GPU workloads.

2.2 Remote Attestation

Remote attestation enables a client to cryptographically verify that code is running within a genuine TEE. The attestation flow consists of:

1. TEE generates a hardware-signed attestation report containing platform measurements
2. Report is signed by manufacturer (Intel/AMD) or cloud provider (Azure)
3. Client verifies signature chain and platform configuration
4. If valid, client trusts subsequent interactions

Azure Attestation provides PKCS7-signed documents containing VM identity, TPM Platform Configuration Register (PCR) values, and TEE activation proof.

3 Threat Model

We consider an adversary with the following capabilities:

- **Malicious cloud operator**: Full control over hypervisor, physical access to hardware, ability to inspect VM memory in standard deployments
- **Compromised service operator**: Access to application deployment, configuration, and logs
- **Network adversary**: Ability to intercept and modify network traffic (mitigated by TLS)

Out of scope: Side-channel attacks on TEE implementations, supply chain attacks on hardware, and denial-of-service attacks.

3.1 Security Goals

1. **Prompt Confidentiality**: User prompts are never accessible to operators or cloud providers
2. **Response Integrity**: Model outputs cannot be tampered with by external parties
3. **Verifiable Execution**: Clients can cryptographically verify TEE deployment
4. **Operator Blindness**: Service operators cannot access user data

4 System Architecture

4.1 Overview

TrustedGenAi deploys an OpenAI-compatible LLM API within an Azure Confidential VM. The architecture consists of three components:

1. **LiteLLM Proxy**: OpenAI-compatible API gateway supporting multiple model backends
2. **Ollama/vLLM**: Local model inference engine running DeepSeek models
3. **Attestation API**: REST endpoint providing cryptographic TEE verification

```
+-----------------------------------------------------------+
|                    Client Application                     |
+---------------------------+-------------------------------+
                            |
                            v HTTPS (TLS 1.3)
+-----------------------------------------------------------+
|                  Cloudflare Edge (TLS Termination)        |
+---------------------------+-------------------------------+
                            |
                            v Cloudflare Tunnel
+-----------------------------------------------------------+
|         Azure Confidential VM (Intel TDX / AMD SEV-SNP)   |
|                                                           |
|  +---------------------+  +----------------------------+  |
|  | LiteLLM (port 4000) |  | Attestation API (port 4001)|  |
|  | - OpenAI API        |  | - Azure PKCS7 proof        |  |
|  | - API key auth      |  | - TPM PCR values           |  |
|  +---------+-----------+  | - TEE dmesg proof          |  |
|            |              +----------------------------+  |
|            v                                              |
|  +---------------------+                                  |
|  | Ollama (port 11434) |                                  |
|  | - DeepSeek-R1       |                                  |
|  +---------------------+                                  |
|                                                           |
|  [Hardware: Memory Encryption via Intel TDX]              |
+-----------------------------------------------------------+
```

Figure 1: TrustedGenAi System Architecture

4.2 Attestation API Design

The attestation endpoint returns a JSON document containing multiple verification layers:

Listing 1: Attestation API Response

```
{
  "platform": "Intel-TDX",
  "vm_size": "Standard_DC4es_v5",
  "tee_verified": true,
  "azure_attestation": {
    "encoding": "pkcs7",
    "signature": "<base64 Microsoft-signed document>"
  },
  "tpm_pcr_sha256": {
    "0": "0x2ADE8023...",
    "7": "0xF8C9E2A1..."
  },
  "tee_dmesg": [
    "Memory Encryption Features active: Intel TDX"
  ]
}
```

Verification Layers:

1. `platform`: TEE technology (Intel-TDX or AMD-SEV-SNP)
2. `azure_attestation`: PKCS7 document signed by Microsoft Azure, containing VM identity and timestamp
3. `tpm_pcr_sha256`: TPM Platform Configuration Registers for software integrity verification
4. `tee_dmesg`: Linux kernel messages proving TEE activation

4.3 Client Verification Flow

Clients should verify attestation before submitting sensitive prompts:

Listing 2: Client Verification Example

```
async function verifyAndChat(prompt) {
  const TEE_API = ’https://tee.vibebrowser.app’;

  // Step 1: Fetch and verify attestation
  const attestation = await fetch(‘${TEE_API}/attestation‘)
    .then(r => r.json());

  if (!attestation.tee_verified) {
    throw new Error(’TEE verification failed’);
  }

  if (!attestation.tee_dmesg.some(
      l => l.includes(’Intel TDX’) || l.includes(’SEV-SNP’))) {
    throw new Error(’TEE not active’);
  }

  // Step 2: Submit prompt to verified TEE
  return fetch(‘${TEE_API}/v1/chat/completions‘, {
    method: ’POST’,
    headers: {
      ’Content-Type’: ’application/json’,
      ’Authorization’: ’Bearer <api-key>’
    },
    body: JSON.stringify({
      model: ’deepseek-r1’,
      messages: [{ role: ’user’, content: prompt }]
    })
  }).then(r => r.json());
}
```

5 Implementation

5.1 Infrastructure as Code

We provide Terraform configurations for reproducible deployment:

Listing 3: Deployment Commands

```
# Clone repository
git clone https://github.com/VibeTechnologies/TrustedGenAi
cd TrustedGenAi/terraform

# Deploy CPU TEE (Intel TDX)
terraform init
terraform apply -var="enable_cpu_tee=true"

# Deploy GPU TEE (NVIDIA H100 CC)
terraform apply -var="enable_gpu_tee=true"
```

5.2 CPU TEE Deployment

Our production CPU TEE deployment uses:

Table 1: CPU TEE Configuration

Parameter	**Value**
VM Size	Standard_DC4es_v5
vCPUs	4
Memory	16 GB
TEE Type	Intel TDX
OS	Ubuntu 22.04 LTS (Confidential)
Model	DeepSeek-R1 1.5B
Cost	$216/month

5.3 GPU TEE with NVIDIA Confidential Computing

For production workloads requiring high throughput, GPU-accelerated TEE provides the optimal solution. NVIDIA H100 Tensor Core GPUs support Confidential Computing mode, extending the TEE boundary from CPU to GPU with hardware-based memory encryption [4].

5.3.1 NVIDIA H100 Confidential Computing Architecture

The NVIDIA H100 GPU implements Confidential Computing through:

- **GPU Memory Encryption**: HBM3 memory is encrypted with keys managed by the GPU's security processor
- **Secure Channel**: Encrypted PCIe communication between CPU TEE and GPU TEE
- **GPU Attestation**: Hardware-rooted attestation reports verifying GPU confidential mode
- **Isolation**: GPU memory isolated from host OS, hypervisor, and other VMs

5.3.2 Cloud Provider Availability

Table 2: GPU TEE Availability by Cloud Provider

Provider	**VM Series**	**GPU**	**TEE Type**
Azure	NCCads_H100_v5	1-4x H100 NVL	AMD SEV-SNP + NVIDIA CC
Google Cloud	A3 Confidential	8x H100	Intel TDX + NVIDIA CC

5.3.3 DeepSeek Deployment on GPU TEE

DeepSeek models can be efficiently deployed on GPU TEE:

Table 3: DeepSeek Model Fit on NVIDIA H100 (94GB HBM3)

Model	Precision	VRAM	Fits on 1x H100
DeepSeek-R1-Distill-1.5B	FP16	3 GB	Yes
DeepSeek-R1-Distill-7B	FP16	14 GB	Yes
DeepSeek-R1-Distill-14B	FP16	28 GB	Yes
DeepSeek-R1-Distill-32B	FP16	64 GB	Yes
DeepSeek-R1-Distill-70B	FP8	70 GB	Yes
DeepSeek-V3 (671B MoE)	FP8	80-90 GB	Yes (37B active)

Note: GPU TEE has not been deployed in our production environment; the following specifications are projections based on hardware capabilities and vendor documentation.

Table 4: GPU TEE Configuration (Projected)

Parameter	Value
VM Size	Standard_NCC40ads_H100_v5
vCPUs	40 (AMD EPYC Genoa)
Memory	320 GB
GPU	1x NVIDIA H100 NVL (94 GB HBM3)
TEE Type	AMD SEV-SNP + NVIDIA CC
Model	DeepSeek-R1-Distill-70B (FP8)
Projected Throughput	150-300 tokens/sec
Cost	$6,300/month

5.4 TPU TEE: Current Limitations

Google Cloud TPUs (Tensor Processing Units) currently **do not support Confidential Computing**. Unlike NVIDIA GPUs with hardware-level encryption, TPUs lack:

- Hardware memory encryption for TPU HBM
- Secure channel establishment with CPU TEE
- Hardware-rooted attestation for TPU workloads

Implication: For privacy-critical LLM inference requiring hardware attestation, NVIDIA H100 Confidential Computing is the only available GPU TEE option. TPU workloads cannot currently provide cryptographic privacy guarantees equivalent to CPU/GPU TEE deployments.

Future Outlook: As confidential computing adoption grows, TPU TEE support may emerge. Organizations requiring TPU performance for large-scale LLM inference must currently choose between:

1. Performance (TPU without TEE) with policy-based privacy guarantees
2. Privacy (GPU TEE with H100) with hardware-verified guarantees

6 Evaluation

6.1 Performance Benchmarks

We measured inference performance across deployment configurations:

Table 5: Inference Performance by Configuration

Config	Model	Tokens/s	Latency	Cost/1M tokens
CPU TEE	deepseek-r1:1.5b	12	83ms/tok	\$5.00
CPU TEE	deepseek-r1:7b	0.7	1.4s/tok	\$85.00
GPU TEE (proj.)	DeepSeek-R1-7B	150	7ms/tok	\$0.40

6.2 Attestation Latency

Attestation API response time: 150-300ms (includes Azure metadata service call).

6.3 End-to-End Verification

We validated the deployment with integration tests:

- **Backend Tests**: 4/4 passing (attestation, models, chat, health)
- **Extension Integration**: 5/5 passing (provider config, HTTPS, wrapper, connectivity, build)
- **TEE Verification**: Intel TDX confirmed via `dmesg` and Azure attestation

7 Security Analysis

7.1 Trust Comparison

Table 6: Trust Requirements by Deployment Model

Trust Assumption	Cloud API	Self-Hosted	TEE
Trust cloud infrastructure	Yes	Yes	Hardware-verified
Trust service operator	Yes	Yes	No
Trust model provider	Yes	No	No
Cryptographic verification	No	No	Yes

7.2 Limitations

1. **TLS Termination**: Cloudflare terminates TLS before the TEE. For maximum security, clients should establish TLS directly to the TEE.
2. **Side Channels**: TEE implementations may be vulnerable to side-channel attacks [17]. Our threat model excludes these.
3. **Model Size Constraints**: CPU TEE limits practical model size to approximately 7B parameters due to memory and performance constraints.
4. **Region Availability**: GPU TEE (NCCads_H100_v5) is limited to East US 2 and West Europe regions.

8 Related Work

Confidential Computing for ML: Prior work has explored TEE-based machine learning [19, 20], primarily focusing on training privacy. Our work addresses inference privacy with remote attestation for end users.

Private LLM Inference: Approaches include differential privacy [22], secure multi-party computation [23], and homomorphic encryption [24]. TEEs offer lower latency at the cost of trusting hardware manufacturers.

Decentralized AI: Projects like Marlin Oyster provide on-chain attestation for TEE workloads. Our work focuses on centralized deployment with client-verifiable attestation.

9 Conclusion

We presented TrustedGenAi, a production-ready infrastructure for privacy-preserving LLM inference using Trusted Execution Environments. Our implementation demonstrates that TEE-based LLM deployment is practical today, with acceptable performance for many use cases and a clear path to GPU acceleration.

The key insight is that remote attestation transforms the trust model: instead of trusting service operators, users verify hardware-signed cryptographic proofs. This enables privacy-critical applications that were previously infeasible with cloud-hosted LLMs.

Open Source: Complete infrastructure code is available at:
https://github.com/VibeTechnologies/TrustedGenAi

9.1 Future Work

1. **Direct TLS to TEE**: Eliminate Cloudflare from the trust path
2. **On-Chain Attestation**: Publish attestation proofs to blockchain for auditability
3. **Multi-Party TEE**: Distribute inference across multiple TEE nodes
4. **Larger Models**: Deploy DeepSeek-V3 on multi-GPU TEE clusters

References

[1] Confidential Computing Consortium. *A Technical Analysis of Confidential Computing*. 2022. https://confidentialcomputing.io/

[2] Intel Corporation. *Intel Trust Domain Extensions (Intel TDX) Module Architecture*. 2023. https://www.intel.com/content/www/us/en/developer/tools/trust-domain-extensions/documentation.html

[3] AMD. *AMD SEV-SNP: Strengthening VM Isolation with Integrity Protection and More*. 2020. https://www.amd.com/system/files/TechDocs/SEV-SNP-strengthening-vm-isolation-with-integrity-protection-and-more.pdf

[4] NVIDIA. *Confidential Computing on NVIDIA H100 Tensor Core GPU*. 2024. https://developer.nvidia.com/confidential-computing

[5] NVIDIA. *NVIDIA H100 Tensor Core GPU Architecture Whitepaper*. 2022. https://resources.nvidia.com/en-us-hopper-architecture/nvidia-h100-tensor-c

[6] NVIDIA. *NVIDIA Confidential Computing: Protecting Data in Use on GPUs*. 2023. https://www.nvidia.com/en-us/data-center/solutions/confidential-computing/

[7] Microsoft. *Azure Confidential Virtual Machines*. 2024. https://learn.microsoft.com/azure/confidential-computing/confidential-vm-overview

[8] Microsoft. *Azure Attestation Overview*. 2024. https://learn.microsoft.com/azure/attestation/overview

[9] Microsoft. *NCCadsH100v5-series Confidential VMs*. 2024. https://learn.microsoft.com/azure/virtual-machines/nccadsh100-v5-series

[10] Google Cloud. *Confidential VM Overview*. 2024. https://cloud.google.com/confidential-computing/confidential-vm/docs/about-cvm

[11] Google Cloud. *Create a Confidential VM Instance with GPU*. 2024. https://cloud.google.com/confidential-computing/confidential-vm/docs/create-a-confidential-vm-instance-with-gpu

[12] BerriAI. *LiteLLM: Call all LLM APIs using the OpenAI format*. 2024. https://github.com/BerriAI/litellm

[13] DeepSeek AI. *DeepSeek-V3 Technical Report*. 2024. https://github.com/deepseek-ai/DeepSeek-V3

[14] DeepSeek AI. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. https://arxiv.org/abs/2501.12948

[15] Ollama. *Ollama: Run Large Language Models Locally*. 2024. https://github.com/ollama/ollama

[16] Kwon, W., et al. *Efficient Memory Management for Large Language Model Serving with PagedAttention*. SOSP 2023. https://arxiv.org/abs/2309.06180

[17] Van Bulck, J., et al. *Foreshadow: Extracting the Keys to the Intel SGX Kingdom*. USENIX Security 2018.

[18] Costan, V. and Devadas, S. *Intel SGX Explained*. IACR Cryptology ePrint Archive 2016. https://eprint.iacr.org/2016/086

[19] Tramer, F. and Boneh, D. *Slalom: Fast, Verifiable and Private Execution of Neural Networks in Trusted Hardware*. ICLR 2019.

[20] Kumar, N., et al. *Privado: Practical and Secure DNN Inference with TEEs*. arXiv:2023.

[21] Hanzlik, L., et al. *MLCapsule: Guarded Offline Deployment of Machine Learning as a Service*. CVPR 2021.

[22] Yu, D., et al. *Differentially Private Fine-tuning of Language Models*. ICLR 2022.

[23] Knott, B., et al. *CrypTen: Secure Multi-Party Computation Meets Machine Learning*. NeurIPS 2021.

[24] Chen, H., et al. *Homomorphic Encryption for Machine Learning*. ACM Computing Surveys 2022.

[25] Mohassel, P. and Zhang, Y. *SecureML: A System for Scalable Privacy-Preserving Machine Learning*. IEEE S&P 2017.

[26] Marlin Protocol. *Oyster: Verifiable Serverless Computing*. 2024. https://www.marlin.org/oyster

[27] Phala Network. *Phala: Trustless Cloud Computing on TEE*. 2024. https://phala.network/

[28] Oasis Labs. *Oasis Network: Privacy-Preserving Smart Contracts*. 2024. https://oasisprotocol.org/

[29] HashiCorp. *Terraform: Infrastructure as Code*. 2024. https://www.terraform.io/

[30] Cloudflare. *Cloudflare Tunnel: Secure Connections Without Public IPs*. 2024. https://developers.cloudflare.com/cloudflare-one/connections/connect-networks/