Duality

CONFIDENTIAL LLM INFERENCE FOR NEOCLOUDS

Protect your customers' prompts and enterprise data on your GPUs, and deliver a secure LLM inference service they can verify.

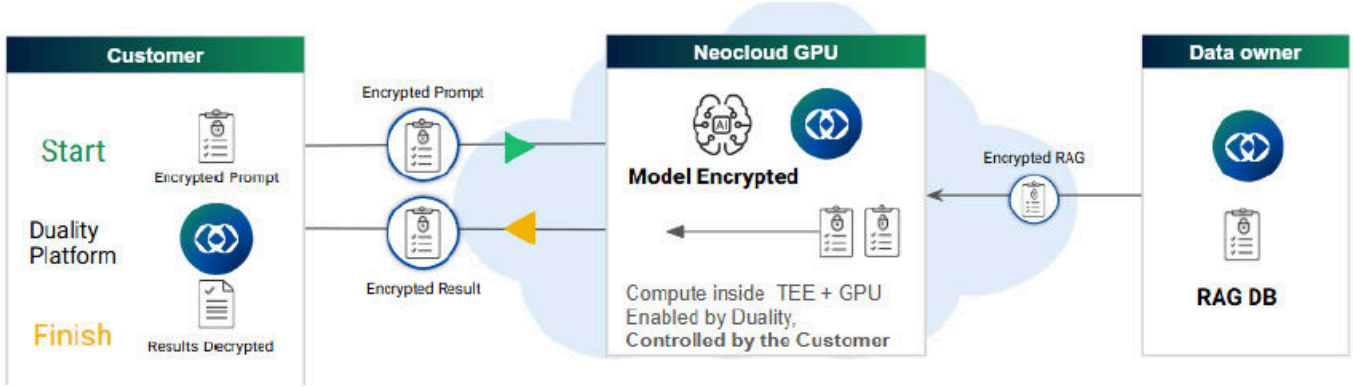
The Opportunity: Win the Customers Who Can't Yet Say Yes

Neoclouds provide the GPU capacity modern AI demands. But regulated and IP-sensitive organizations, from healthcare and financial services to government and defense, hesitate to run LLMs on shared infrastructure. Their prompts, retrieved enterprise documents (RAG), and outputs contain exactly the data they cannot expose to an infrastructure operator.

The challenge grows with RAG, where internal knowledge is incorporated directly into prompts. Contractual assurances and standard access controls are not enough: customers want proof that no one, including your administrators and other tenants, can see what runs on your hardware.

Without that proof, high-value workloads stay on-prem or with providers who can offer it.

The Solution: Protect Customer Prompts on Your GPUs



Duality lets you offer secure LLM inference on your GPU infrastructure, with customer prompts, enterprise knowledge, model inputs, and outputs protected from your own staff and platform. Using Trusted Execution Environments (TEEs), remote attestation, and built-in governance, Duality verifies each workload before releasing sensitive assets and ensures only approved applications and models can access protected data.

The result: your customers keep control of their data, and you gain a verifiable, differentiated security story for your inference service.

HEADQUARTERS
5 Marine View Plaza Hoboken, NJ 07030

CONTACT US
info@dualitytech.com
Duality Technologies
dualitytech.com

Go Beyond Standard Confidential Computing

Feature / Benefit	Typical Cloud-Vendor TEE	Confidential Inference with Duality
01 Attestation	Verified through the cloud KMS. Limited controls, vendor-specific parameters, and trust in cloud administrators required.	Verify not only the TEE, but the exact workload and model running inside it. Works air-gapped and multi-cloud.
02 Key Exchange	Decryption keys are transmitted from the KMS to the TEE over HTTPS.	Quantum-resistant key encapsulation. Ephemeral keys are never transmitted or exposed.
03 Result Verification	Not available.	Extensible verification helps prevent sensitive information from leaking through outputs.
04 Governance	Cloud IAM controls access to infrastructure.	Granular governance over workloads, data, users, and execution policies.
05 Auditing	Basic KMS logging.	Detailed audit trails including attestation details and policy-enforcement outcomes.
06 Scale-Out TEE Clusters	Confidential instances with limited attestation, sovereignty, and GPU support.	Secure interconnect between attested TEE instances with Layer-3 encryption and GPU support.
07 Federated Workloads	Not available.	Built-in federated workflows with secure aggregation inside the TEE.
08 FHE in the TEE	Not available.	Supports FHE-based queries inside the TEE for an additional layer of data protection.
09 Cloud Flexibility	Tied to a single cloud provider.	Seamless migration across clouds and support for cross-cloud collaboration.
10 Collaboration Framework	Not available.	Turnkey framework for managing participants, roles, policies, and datasets across multi-party collaborations.

HEADQUARTERS
5 Marine View Plaza Hoboken, NJ 07030

CONTACT US
info@dualitytech.com
Duality Technologies
dualitytech.com

Key Benefits for Neoclouds

Unlock Regulated, High-Value Demand

Serve customers who can't send sensitive prompts, RAG content, or outputs to shared GPU infrastructure today.

Give Customers Control

Customers define exactly which applications, models, users, and datasets can interact on your platform.

Prove What Is Running

Attestation confirms the integrity of workloads and models before any data is released.

Protect Prompts From the Operator

Prompts, retrieved content, and outputs stay shielded from your administrators and other tenants.

Differentiate With a Premium Tier

Move the conversation from price per GPU-hour to verifiable data protection and accelerate customer security reviews.

Stay Multi-Cloud, Avoid Lock-In

Support multi-cloud and cross-environment deployments without redesigning security controls.

How it Works

1. **Verify the Environment**

Duality validates the integrity of the workload, application, and AI model before sensitive assets are released.

2. **Enforce Governance Policies**

Access controls determine which applications, users, models, and datasets are authorized to participate.

3. **Run Inference Securely**

Prompts, retrieved documents, model weights, and inference processing occur within the protected execution environment on your GPUs.

4. **Deliver Results with Confidence**

Results are securely returned to authorized users, accompanied by detailed auditing and verification records.

Ready to offer confidential inference on your GPU cloud?

Talk to Duality about adding verifiable prompt protection to your inference service: info@dualitytech.com

Proudly Working With

Capgemini · National Security Strategic Investment Fund · ORACLE · Red Hat · NIH · DARPA · NHS · intel · aws · Google Cloud · NVIDIA · Azure · Dana-Farber Cancer Institute · Deloitte. · GOV.UK

HEADQUARTERS
5 Marine View Plaza Hoboken, NJ 07030

CONTACT US
info@dualitytech.com
Duality Technologies
dualitytech.com