Duality

CONFIDENTIAL INFERENCE FOR MODEL PROVIDERS

Offer your customers verifiable, secure LLM inference and agentic workflows, on infrastructure you choose, without exposing their prompts, their data or your model.

Challenge: Your Best Customers Cannot Send You Their Data

Banks, health systems, governments and defense organizations hold the data where frontier models add the most value, yet cannot let prompts, retrieved documents or outputs leave their control. Retrieval and agentic workflows raise the stakes: the more context an agent reaches, the more sensitive the prompt.

At the same time, your model weights are among your most valuable assets. Deploying them into a customer's cloud or a sovereign region puts that intellectual property in an environment you do not operate, and contractual promises are no longer enough.

The Solution: Confidential Inference, Verified Before Anything Is Released

Duality Technologies runs inference inside hardware-backed Trusted Execution Environments (TEEs) and verifies the environment, the application and the model through remote attestation before any key, prompt or model asset is released. Governance policies decide which applications, models, users and datasets may take part, and every session leaves an audit trail.

The result: your customers get cloud-scale AI with evidence they can check, and you reach regulated workloads that would otherwise stay out of reach.

Key Benefits For You

Reach Regulated Customers

Serve finance, healthcare, government and defense buyers who need proof that prompts and outputs stay protected from cloud operators.

Deploy Without Handing Over Your Model

Run your model in a customer's cloud or a sovereign region while its weights stay inside the attested environment.

Let Customers Verify, Not Trust

Attestation covers the exact workload and model running in the TEE, so customers check what is running before their data is released.

Govern Agentic Workflows

Policies set which applications, models, users and datasets can interact, giving security teams controls they can approve.

HEADQUARTERS
5 Marine View Plaza, Hoboken, NJ 07030

CONTACT US
info@dualitytech.com

LINKEDIN
Duality Technologies

WEBSITE
dualitytech.com

CONFIDENTIAL INFERENCE FOR MODEL PROVIDERS

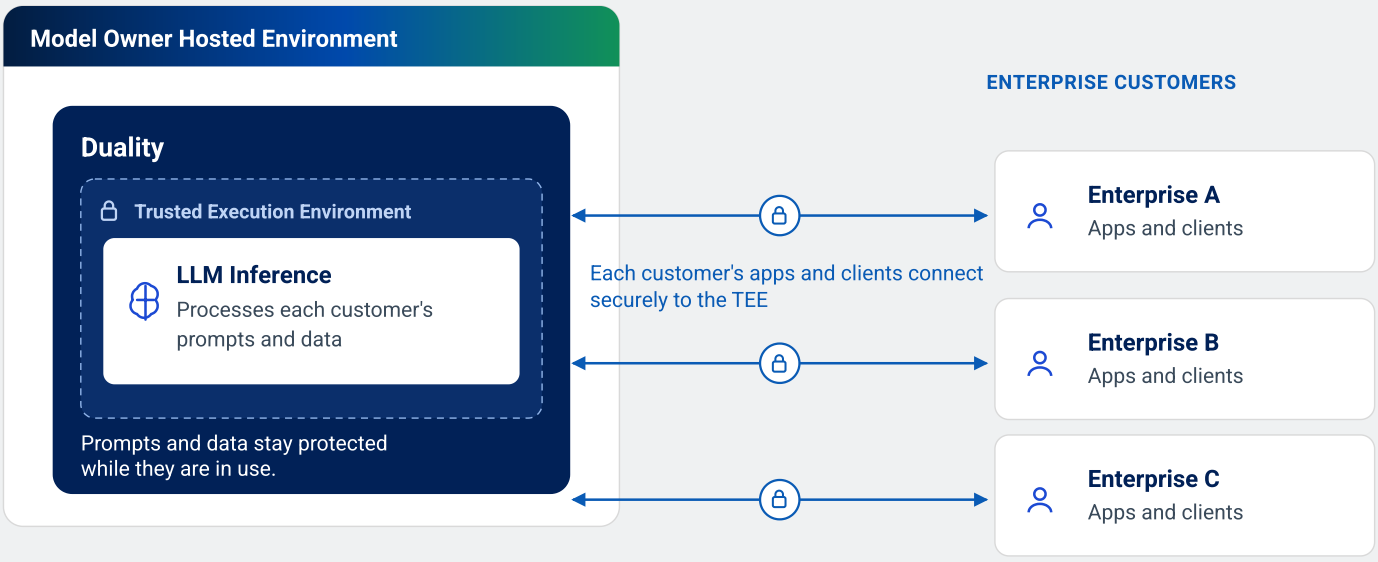
Extend Confidential Computing Beyond Infrastructure Security

A cloud vendor's TEE protects the infrastructure. Duality adds the verification, key handling, governance and audit that regulated customers ask a model provider to demonstrate.

	Feature / Benefit	With Duality
01	**Attestation**	Validates the attestation report, the confidential VM parameters and the authenticity of the workload container before any key is released.
02	**Key Exchange**	AES-256 session keys wrapped to the enclave with ML-KEM-1024 key encapsulation, which is quantum-resistant.
03	**Result Verification**	Custom result-verification script checks outputs before they are returned, to help prevent sensitive information leaking through results.
04	**Governance**	Policies over which applications, models, users and datasets may participate in a session.
05	**Auditing**	Audit trail that includes attestation details and policy-enforcement outcomes.
06	**Cloud Flexibility**	Move workloads between clouds and support multi-cloud strategies without redesigning security controls.
07	**Collaboration Framework**	Roles, participants, policies and assets managed in one project model, so a model provider, a customer and a compute host each hold a defined role.
08	**Always-On Model Services**	Register a containerized application once and run it as a long-running service on a confidential-GPU TEE in your own cloud account, with mutual-TLS client trust and browser-verifiable attestation.

Attestation signatures from the cloud vendor are not quantum-resistant; they carry no secrets and are valid for seconds to minutes.

Many Customers, One Protected Hosted Service



HEADQUARTERS
5 Marine View Plaza, Hoboken, NJ 07030

CONTACT US
info@dualitytech.com

LINKEDIN
Duality Technologies

WEBSITE
dualitytech.com

CONFIDENTIAL INFERENCE FOR MODEL PROVIDERS

How It Works, and Where It Fits

How It Works

1

Verify the Environment

Duality validates the TEE, the application and the model before any key or sensitive asset is released.

2

Enforce Governance Policies

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

3

Run Inference Securely

Prompts, retrieved documents, model weights and inference processing stay inside the protected environment.

4

Deliver Results With Confidence

Results are encrypted to the authorized recipient, with audit records of attestation and policy enforcement.

Three Ways to Offer It

SECURE INFERENCE

- Offer a confidential tier of your API to regulated customers.
- Customers bring prompts and retrieved enterprise knowledge; neither the host nor your operators see them.

CUSTOMER-CLOUD DEPLOYMENT

- Run your model in a customer's cloud account or a sovereign region.
- Weights are released only to an attested environment that policy approves.

AGENTIC WORKFLOW PLATFORMS

- Run governed workflows over sensitive enterprise knowledge.
- Policy defines which applications, models and datasets take part; audit shows what ran.

Deployment

Deploys on all cloud vendors, across public cloud, sovereign cloud and multi-cloud environments. Any model, traditional ML or LLM, can run for inference on structured or unstructured data. Duality is not a model vendor: the models, and the customer relationship, stay yours.



HEADQUARTERS
5 Marine View Plaza, Hoboken, NJ 07030

CONTACT US
info@dualitytech.com

LINKEDIN
Duality Technologies

WEBSITE
dualitytech.com