Duality

DUALITY SECURE AI

RUN AI ON SENSITIVE DATA. VERIFY BEFORE DECRYPT.

Confidential AI infrastructure for regulated, distributed environments.

The Trust Gap in Enterprise AI

Enterprise AI has a trust gap at its center. The organizations that hold the most valuable data, in government, defense, finance and healthcare, are also the ones that cannot let it cross a security boundary. At the same time, the owners of proprietary models will not place their weights inside infrastructure someone else administers. A large share of regulated data still sits on-premises or in private clouds, out of reach of the best available models. The data is not allowed to travel to the model, and the model cannot safely travel to the data.

Duality closes that gap. The Duality platform's Secure Collaborative AI capability brings together hardware-isolated Trusted Execution Environments (TEEs), remote attestation checked against the attestation authority, per-owner encryption keys that never leave their owner's control, and a governance layer that decides which workload may touch which asset. Nothing is decrypted until the environment has proven what it is. Data, models, keys and results stay hidden from the cloud provider, the host administrator and every other participant in the project, including while the AI workload is actively running.

How It Works

Enclave-Based AI Execution

Each AI session executes inside a hardware-backed enclave whose memory and execution state are invisible to the operating system, the hypervisor, the cloud provider and any privileged user on the host.

Duality runs on the confidential-computing platforms customers already have: AWS, GCP, Azure or on-prem. The enclave, the attestation exchange and the key hand-off are managed by the platform rather than built by hand on raw cloud primitives.

Attestation-Gated Key Release

Isolation on its own does not tell a data owner that an environment deserves their data. Before any asset is released, the owner's Duality node asks the enclave to prove itself. The enclave returns an attestation report signed by the cloud provider's attestation authority or, in the case of Red Hat, by the Trustee running at the asset owner. The node validates that signature, checks the confidential-VM parameters, and compares the digest of the running workload image against the digest the owner approved when the policy was granted.

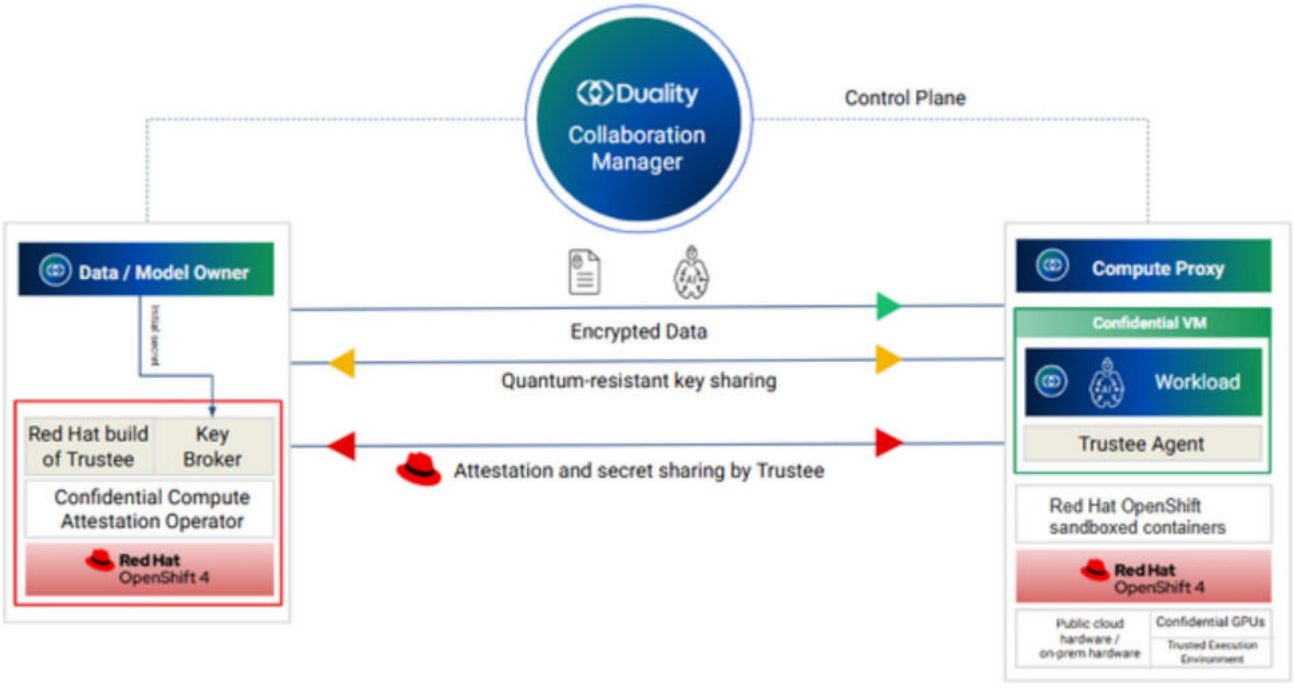
Only when every check passes does the node encrypt its session key to the enclave's public key and send it across. A modified image, a different host configuration or a report that fails signature validation means attestation fails and the key is never sent.

The exchange is built on quantum-resistant primitives. Assets are protected with AES-256, and session keys travel under ML-KEM-1024 post-quantum key encapsulation, so material intercepted today cannot be decrypted by a future quantum computer.

AES-256
Protects every asset

ML-KEM-1024
Post-quantum key encapsulation for session keys

EXAMPLE: ARCHITECTURE WITH RED HAT CONFIDENTIAL CONTAINERS



Keys and Control, Owned by the Asset Owner

Every participant in a project holds an explicit role.

Asset Owners

Contribute data or models.

Analyzing Party

Receives the results.

Compute

Hosts the enclave.

Collaboration Manager

Orchestrates the project without ever seeing data or computation.

Each Asset Owner encrypts under their own key and completes their own attestation with the enclave. A data owner and a model owner can therefore participate in the same session without either one handing a key to the other, to the Compute host or to the cloud provider. The model runs against the regulated dataset inside the enclave: the model owner never sees the records, the data owner never sees the weights, and the party operating the infrastructure sees neither.

TRUST IN EVIDENCE, NOT CONTRACTS

Trust is placed in a verified environment and an approved workload, not in a contract between organizations. That is what lets a bank use a vendor's proprietary risk model, or a hospital network train on a pharmaceutical partner's model, without either side giving anything up.

No Duality Key Store in the Middle

Duality does not introduce its own key store between the customer and their data. The ephemeral key is created by the Duality task, kept in memory, and cleared immediately after it has been used and shared. The Compute host, the cloud provider and Duality itself never hold a participant's keys.

Keeping execution, attestation and key ownership separate means that controlling the infrastructure never translates into access to what runs on it, and confidential AI deployments line up with the security boundaries an organization has already drawn.

Confidential GPU-Accelerated Inference

Modern models need accelerators, and the confidentiality guarantee has to cover them. Duality hosts applications on NVIDIA GPUs in confidential-compute mode, so the GPU is part of the attested boundary. Attestation confirms that confidential-GPU mode is enabled before any key is released, and the model is loaded into enclave memory only after the owner's node has verified the running image.

A TYPICAL DEPLOYMENT

An organization holds a proprietary model and wants to offer it as a service to its customers without exposing its IP. With Duality it deploys that model as a persistent inference service on confidential GPUs in a public cloud or on-prem, and the other party sends prompts to it, while the TEE host can see neither the model nor the prompts. Clients can confirm, from a standard browser with just a browser extension, that they are talking to a genuine, untampered enclave before they send a single token.

Quantum-Resistant Security Model

When evaluating the security model, it is essential to understand the associated risks and challenges. In the context of quantum-resistant security, the primary risk is that an adversary could capture encrypted data today, store it, and later decrypt it using the computational power of a future quantum computer.

To mitigate this risk, Duality employs encryption protocols that are mathematically proven to be quantum-resistant: even a quantum computer would not be able to decrypt the protected data. These protocols are integrated throughout Duality's encryption processes, which involve a Trusted Execution Environment (TEE).



Data Encryption

Duality uses the AES-256 protocol, which is considered quantum-resilient.



Key Exchange

Duality employs the ML-KEM-1024 protocol, implemented via the liboqs library, for quantum-resistant key exchange.



Attestation

The report signature currently uses non-quantum-resistant algorithms. It carries no sensitive information and is valid only for seconds to minutes, so it presents no material quantum-related risk.

Governed and Auditable by Design

Confidentiality is only half the requirement; the other half is being able to show what happened. In Duality, access is expressed as policy. An Asset Owner states which workloads or applications may use an asset, by whom, under what conditions, how many times and until what date. A workload's hash is captured when the policy is approved and re-checked at execution, so the computation that runs is exactly the one that was approved. Revoking a policy, or reaching its termination date, removes the asset from enclave memory immediately.

Every attestation result, key release, policy decision, session lifecycle event and user action is written to the platform's built-in audit and logging stack. Security, compliance and operations teams can query that record for regulatory review, incident response and forensic investigation.

The record carries metadata only. Prompt and response content, raw data and model weights are never logged, exported or persisted outside the enclave.

Who It Is For

Duality gives enterprises, model providers, sovereign and private cloud operators, and integration partners a shared foundation on which state-of-the-art AI can meet the world's most sensitive data, without the data owner or the model owner giving up control.

Enterprises

Run proprietary foundation models and fine-tuned models against regulated data inside their own security domain, and serve their own models from confidential GPUs in a cloud they do not have to trust.

Model providers

Bring closed-weight models to workloads that previously demanded local or open-weight alternatives, while the weights stay encrypted under their own key.

Sovereign and private cloud operators

Host regulated data and proprietary models for their tenants without being able to read either, opening markets where operator access would breach security or sovereignty rules.

Integration and OEM partners

Deliver pre-integrated confidential-AI stacks that combine confidential compute, GPUs and Duality's governance and attestation layer.

Connected, Sovereign and Disconnected Deployments

The organizations with the strongest appetite for advanced AI frequently operate under the tightest infrastructure rules: behind enterprise firewalls, within a sovereign cloud, or on networks with no route to the internet at all. Duality is software, and it deploys where the data lives: public cloud, private cloud, on-premises or a mix of the three. Data owners can keep their node on-premises and contribute to sessions running on cloud enclaves without their data ever being stored outside their perimeter.

Connected environments

Attestation is validated live against the cloud provider's attestation authority, and keys are brokered through the participant's own KMS.

Restricted environments

Offline attestation validation uses a pre-provisioned signature validation key, so the same verify-before-decrypt model holds where no outbound connection is permitted.

Why Duality

By bringing together hardware isolation, remote attestation, participant-owned keys, policy-based governance and a complete audit record, Duality protects data in use, so organizations can put AI to work on data that could never leave their control.

Protect Data in Use	Extend protection from data at rest and in transit to data and models while they are being processed.
Verify Before Decrypt	Remote attestation is checked against the cloud provider's attestation authority and the approved workload digest before any key is released.
Every Owner Keeps Their Key	Data owners and model owners each encrypt under their own key, and nobody else in the project, including the infrastructure host, ever holds it.
Quantum-Resistant by Default	AES-256 for assets and ML-KEM-1024 post-quantum key exchange throughout the enclave flow.
Confidential GPUs	Persistent inference services on NVIDIA confidential GPUs, with browser-verifiable attestation for end clients.
Governed and Auditable	Policies define who, what, when and how many times; a tamper-evident audit trail records every decision without exposing protected assets.
Deploy Anywhere	Public cloud, private and sovereign cloud, on-premises and hybrid, with offline attestation for restricted networks.