

NeodyAI AI Gateway

Using AI agents with confidential data — safely, and under your model risk controls

Technical overview · October 2026 · Patent pending

This overview describes the NeodyAI AI Gateway and the guarantees it is designed to provide. It is being delivered in phases; contact us for availability for your deployment. A detailed architecture briefing is available under NDA.

The problem

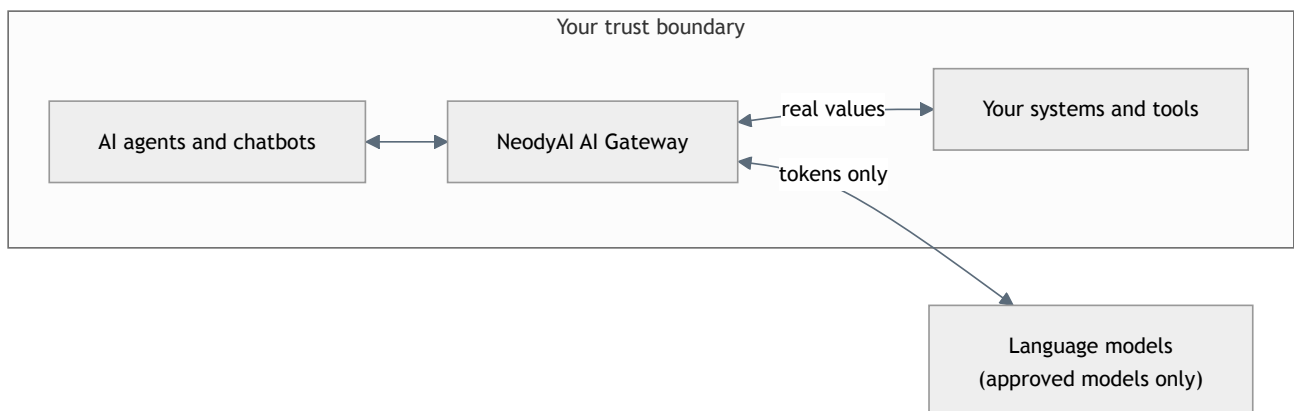
AI agents are useful because they combine two things:

- **A language model** — usually run by a third party, outside your organization — that reasons about the work.
- **Your systems** — CRM, payments, claims, core banking, documents — where the work actually happens.

The model does not need your customers' real card numbers, account numbers, health information or names to reason about a task. Your systems cannot act without them. And if you are a regulated institution, every model that influences a decision has to be inventoried, validated, monitored and changed under control — including a third-party model that its vendor can update at any time.

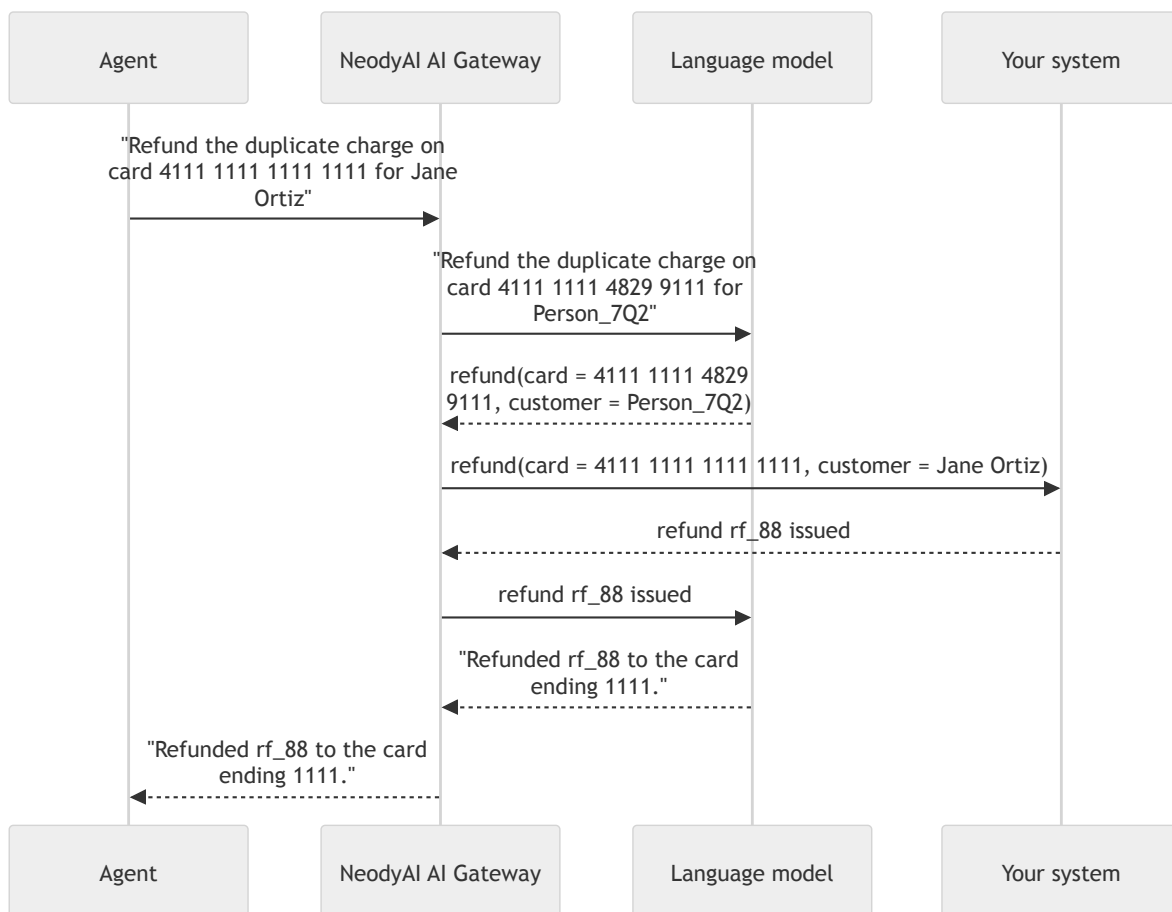
What the gateway does

The NeodyAI AI Gateway sits on every path between your agents, the language models they use, and your systems.



1. **Protected data never reaches the model.** Before anything is sent to a language model, the gateway finds protected values — card data, health information, personal information, account numbers, confidential business information, credentials — and replaces them with **tokens**.
2. **Your systems get the real values.** When the model asks for an action, the gateway restores the real values and passes them only to your approved systems, acting with the permissions of the person the agent works for.
3. **Results are protected on the way back.** Data returned by your systems is tokenized before the model sees it.
4. **Every call is governed and recorded.** Only approved models can be reached, each approved for specific kinds of data, and every call leaves an evidence record — without your data in it.

A round trip



The model worked with the shape of the data — a card number, a customer — and never with the real values.

Tokens the model can work with

- **They keep their shape.** A tokenized card number still looks and validates like a card number, so the model can reason about it and existing validation still works.
- **They are consistent within a conversation,** so the model can tell that two mentions refer to the same customer or account.
- **They mean nothing outside it.** Tokens cannot be reversed outside the gateway.

Finding protected data

Detection is **pluggable**. The gateway combines several methods, the most certain first:

- the field definitions of your own systems,
- validated patterns (for example card numbers with checksums, bank and routing numbers),
- matching against records you already hold,
- your own confidential term lists,
- recognition models for names and addresses in free text,
- and, if you choose, **the data loss prevention (DLP) system you already run.**

When detection is uncertain, the gateway protects the value. **No detection or tokenization service outside your trust boundary ever receives your protected data.**

Tokenization is pluggable too: use NeodyAI's, or **the tokenization product you already own.**

Keys

- Every agent has its own identity and keys; the agent's private key never leaves the agent.
- Every conversation has its own session keys, and traffic between an agent and the gateway is encrypted per call in addition to TLS.
- Keys are held in a managed key service; you can rotate or revoke them, and revocation ends sessions immediately.

Model risk management

The gateway is where every model call passes, so it is where model risk controls are enforced:

Expectation (e.g. SR 11-7, OCC 2011-12)	What the gateway provides
Model inventory	Only registered, approved models can be reached by any agent
Validation before use	New models and agents run alongside your people before they are allowed to act
Effective challenge	Approval by someone independent of the proposer, enforced by the system
Ongoing monitoring	Quality, cost and protected-data findings tracked per model and per agent
Change control	When a model changes, agents relying on it automatically lose autonomy until re-validated
Documented limits	Which data each model and agent may handle is enforced on every call
Evidence	Tamper-evident records of every call, exportable for examiners, with no customer data in them

The detection methods are managed the same way: inventoried, validated against minimum detection rates, and monitored.

What we claim, and what we do not

We say	We do not say
Protected data is tokenized before it reaches a language model	"End-to-end encrypted" — the gateway restores real values so your systems can act
The model provider never receives your protected values	"NeodyAI never sees your data" — in the cloud deployment the gateway processes values inside the trust boundary; an on-premises option protects the most sensitive data before it reaches NeodyAI
Detection rates are measured and reported	That every protected value can always be detected

Agents and the network

NeodyAI is built on **AgentMesh** and **Agent Messaging**, the open protocol for agents to work with each other. NeodyAI's agents have their own identities and can work with other agents — your suppliers', partners' and customers' — over Agent Messaging, with every message signed.

The same rule holds on the network as with models: **protected data does not leave your trust boundary untokenized**. Today NeodyAI does not send protected values on the network at all; every

request is checked before it goes and recorded as evidence. For regulated deployments, the open-source Agent Messaging Server can run in your own environment.

Deployment

- **NeodyAI cloud**, isolated per customer.
 - **Protection on your premises** for card data and the most sensitive classes.
 - **Your own agents** can use the gateway: register the agent and point its model and tool calls at the gateway.
-

NeodyAI · neody.ai · Patent pending. A detailed architecture briefing is available under NDA — contact us at neody.ai.