QiD Vanguard

Every prompt your company sends, governed.

Your people keep using the AI tools they already like, including a personal ChatGPT account on a company laptop. Behind the screen every prompt is checked, personal details are swapped for placeholders, and anything too sensitive to leave the country is answered by a model inside your own building.

AT A GLANCE

4
Ways your staff reach AI, all governed by one set of rules

100+
AI models reachable through one address, swappable without touching your apps

0
Personal data written to disk anywhere in the system

Yours
Runs in your cloud or data centre. We never host or see your traffic.

THE PROBLEM

Your staff are already using AI. You just cannot see it.

Blocking AI outright does not work, because people use their phones instead, and you lose both the productivity and the visibility. Allowing it without controls means customer names, ID numbers and contract terms leave the building in a text box.

Most gateways only govern the applications your own developers built to call them. That is the smallest part of the problem. The bigger part is a person, in a browser tab, pasting a spreadsheet.

What makes this different

QiD Vanguard governs **four** routes to AI, not one: the company chat, your own applications, coding tools on a laptop, and a consumer AI site open in a browser. All four are held to the same rules, written once.

The three outcomes, on every message

ALLOW Nothing sensitive. Sent as typed. Most messages.

MASK Personal details swapped for placeholders the model still understands. The real values never leave.

REROUTE Too sensitive to send abroad, so your own in-house model answers it. The person still gets help.

REFUSE Passwords and access keys are stopped outright. A masked key is still a leaked key.

Nobody has to learn anything. Staff keep the same chat window, the same coding tool, the same browser. The controls sit behind the screen, not in front of it.

What the AI provider actually receives

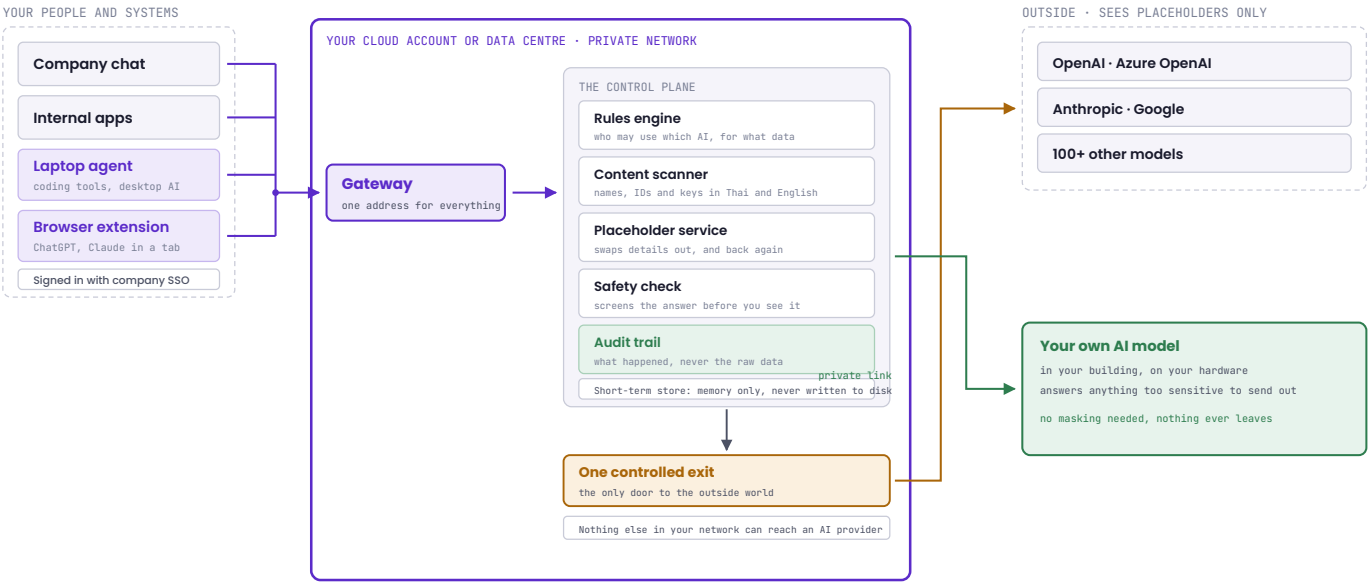
```
the person typed Draft a reply to John Doe at john.doe@example.com about the invoice. the AI provider received Draft a reply to [PERSON_2cb009] at [EMAIL_ADDRESS_69499f] about the invoice. the person read Certainly, I'll email John Doe at john.doe@example.com about the invoice.
```

The placeholder is the product. It is not a black box over the name: the model still knows it is a person, and refers to the same one later in the conversation. Answers stay useful, and the real name never leaves your network.

DEPLOYMENT

What gets installed, and where

Ten services, deployed into your own cloud account or data centre. We never host your traffic and no prompt leaves your network un-inspected.



Runs on Kubernetes. Scales with demand. No component writes personal data to disk.

The single exit is what makes the guarantee enforceable. AI providers are reachable from one address and nowhere else, so a forgotten script or a misconfigured service cannot quietly go around the gateway.

Nothing to re-learn

Same chat window, same coding tool, same browser.

Your account, your data

Deployed into your cloud. We never see a prompt.

Swap models freely

Change provider or fail over. A setting, not a project.

One audit answer

Who used which AI, with what kind of data, and what the rules decided.

THE JOURNEY

Four places people type. Every path, end to end.

The same rules apply wherever the message starts. What changes is how it is reached, and what the person sees when something is stopped.

Company chat `our chat app · signed in with company SSO`

1 The person's team is read from their sign-in, so the rules know exactly who is asking.

2 The message is scanned for names, ID numbers, account numbers and keys.

3 Details are swapped for placeholders, and the real values are held in memory only.

4 The AI answers, seeing placeholders. The reply is screened for harmful content.

5 Placeholders are swapped back, so the reply reads naturally.

FULL CONTROL Masking, rules and a complete audit record.

Your own applications `point them at one address`

1 Apps call one address instead of an AI provider directly. No provider keys live in your code.

2–5 Identical to the chat path: scan, mask, answer, screen, restore.

FULL CONTROL Swapping AI provider becomes a settings change, not a development project.

Coding tools and desktop AI `laptop agent`

1 The agent quietly repoints supported tools at the gateway. The developer notices nothing.

2 From there it is an ordinary governed request, with the same scanning, same masking, same audit.

3 Tools that cannot be redirected are blocked, with a link to the approved alternative.

FULL CONTROL where the tool can be redirected BLOCKED where it cannot

ChatGPT or Claude in a browser tab `browser extension`

1 The extension reads what was typed or pasted before the browser encrypts it. Nothing is intercepted or decrypted.

2 A fast local check runs on the laptop. Ordinary text never leaves it.

3 If something sensitive is found, the person is stopped and handed the approved chat, **with their text carried across**, so the safe route costs them nothing.

WARN & REDIRECT Sensitive content stopped at the source; ordinary use is untouched.

Why the browser matters most. Almost every real leak starts with a person pasting into a browser tab, not with an application making an API call. A gateway that only governs your own applications misses the part of the problem that actually happens.

WORKED EXAMPLES

Four messages, four different outcomes

Same person, same requested AI. Only the content changes, and the system decides on that alone.

An ordinary question — DELIVERED UNCHANGED

TYPED	What is the capital of Thailand?
SENT	What is the capital of Thailand?

Nothing sensitive, so nothing is changed. Most messages take this path.

A customer name and email — DELIVERED, MASKED

TYPED	Draft a reply to **John Doe** at **john.doe@example.com**.
SENT	Draft a reply to [PERSON_2cb009] at [EMAIL_ADDRESS_69499f].
READ BACK	Certainly, I'll email **John Doe** at **john.doe@example.com**.

The provider wrote the reply without ever seeing who it was for. The person read the real name.

A Thai national ID number — ANSWERED IN-HOUSE

TYPED	ลูกค้า บัตรประชาชน 1101701558227
SENT	never sent to an overseas provider

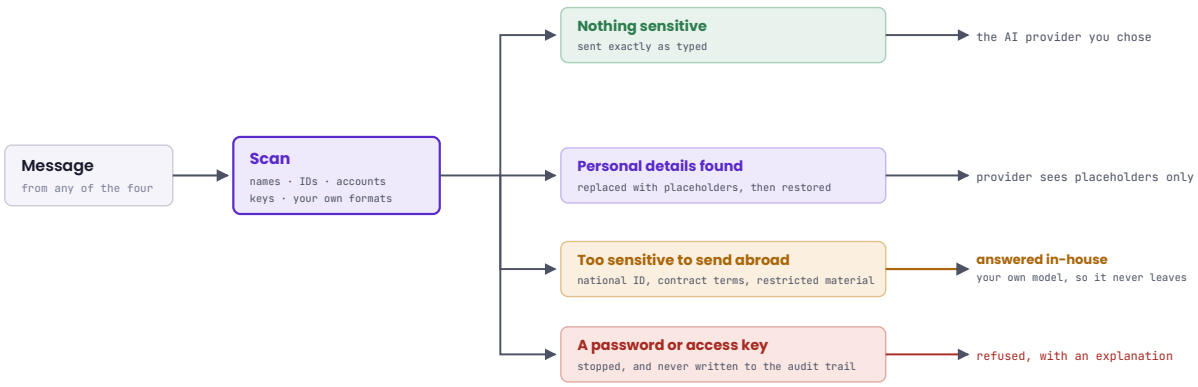
Too sensitive to leave the country, so it did not. Your own model answered it and the person still got help.

A cloud access key — REFUSED

TYPED	deploy with **AKIAIOSFODNN7EXAMPLE**
SENT	nothing sent

Credentials are refused, never masked. The key is not written to the audit trail either.

How the decision is made



The third branch is the one competitors do not have. Most products can only allow or refuse. Rerouting protects the data and answers the question, so staff have no reason to look for a way around the control.

EVIDENCE

What the audit trail keeps


```
"action": "mask", "masked_prompt": "Draft a reply to [PERSON_2cb009] at [EMAIL_ADDRESS_69499f].",
"sensitive_items": { "PERSON": 1, "EMAIL_ADDRESS": 1 }, "fingerprint": "sha256 of the original, proves what
was sent and reveals nothing"
```


The record shows what happened without becoming a copy of the data it was protecting. It refuses entries that still contain a real identifier, so if masking ever silently stopped working, the audit trail would reject the entry rather than quietly store the leak.

The third example is the one to remember. Refusing people is easy and it backfires, because anyone refused finds another way. Answering the question on your own hardware protects the data and keeps the person productive. That is the difference between a control people work around and one they never notice.

RUNNING IT YOURSELF

Your rules, your dashboard, your evidence

Define your own formats

Paste five real employee IDs and the builder proposes the pattern. No regular expressions, no code change, no waiting on us.

Test before you ship

A playground shows what a rule matches, what it would do to a whole request, and how many messages a day it would newly block.

See what is happening

Active users, request volume, what was masked, blocked or rerouted, and spend per team against budget.

Prove it to an auditor

Ninety days of records showing what the controls did, holding masked content and fingerprints rather than the data itself.

The playground measures false positives, not just matches. A rule that catches every employee ID and also every purchase order number gets the whole control switched off within a week, and that failure is invisible until it has already happened. So a candidate rule is replayed against your own recent traffic and reports what it would newly block, before anyone can save it.

When an investigation genuinely needs the original

Refusing outright sounds principled and ends with someone exporting to a spreadsheet, which is worse. Instead there is a narrow path: **two people must approve**, and they must be different people. Access lasts 24 hours, the security team is notified on approval, and the approval itself is recorded permanently and cannot be reached by the same mechanism it authorises.

It also cannot reach ordinary traffic. Originals are kept only for requests that were blocked or flagged, so for everything else there is nothing to retrieve even with approval. That is a property of what was written down, not a promise about who is allowed to read it.

SUPPORT MATRIX

What we support today, and how each one is reached

Three enforcement points cover three kinds of destination. Which one applies is decided by what the destination is, not by preference.

AI models, reached through the gateway

PROVIDER	REACHED BY	WHAT HAPPENS TO YOUR DATA
OpenAI	gateway API	MASKED personal details replaced before it leaves
Azure OpenAI	gateway API	MASKED regional residency, so internal content may reach it
Anthropic	gateway API	MASKED
Your own model, vLLM	gateway API	INSPECTED no masking needed, nothing leaves your network
Your own model, Ollama or Unsloth	gateway API	INSPECTED
100+ others	gateway API	Google, Bedrock, Mistral, Cohere and the rest, through the same address

Consumer AI in a browser tab

SITE	REACHED BY	WHAT HAPPENS
claude.ai	browser extension	GUARDED checked before Send, personal accounts included
chatgpt.com	browser extension	GUARDED logged-out use blocked outright
gemini.google.com	browser extension	BLOCKED until an enterprise agreement is in place
perplexity.ai	browser extension	BLOCKED no zero-retention terms
openrouter.ai	browser extension	BLOCKED the real destination cannot be determined

REQUIREMENTS

What you provide, and what we bring

Desktop and coding tools

APPLICATION	REACHED BY	WHAT HAPPENS
Cursor	desktop agent	FULL CONTROL repointed at the gateway
Claude Code	desktop agent	FULL CONTROL
Continue	desktop agent	FULL CONTROL
Aider	desktop agent	FULL CONTROL
Anything on the OpenAI SDK	desktop agent	FULL CONTROL Python and Node
Claude Desktop	desktop agent	GUARDED honours a system proxy
ChatGPT Desktop	desktop agent	BLOCKED pins its certificate, so it cannot be inspected

An application not on this list defaults to blocked rather than allowed. New AI tools appear faster than any review process, so the list is the allowance and everything else waits for review.

You provide

Somewhere to run it	Kubernetes your cloud account or data centre
Identity provider	OIDC or OAuth2 Entra, Okta, Google Workspace
Device management	Intune, Jamf or similar to deploy the agent and extension
Managed browser	Chrome or Edge enterprise policy for force-install
One firewall rule	Outbound allowlist AI providers reachable only from the gateway
Your AI accounts	Whichever you use we never resell model access
Your confidential formats	Employee ID and similar entered in the rule builder, not by us

We bring

The gateway	Ten services deployed into your account
Chat application	Branded, SSO the sanctioned alternative
Desktop agent	Windows and macOS signed, MDM-deployable
Browser extension	Chrome and Edge scoped to AI sites only
Detection	English and Thai including identity numbers with checksums
Rule builder	With a playground so your team owns its own formats
Dashboard and audit	90 day retention adoption, blocks, spend, health

Optional but recommended: a model of your own. Sensitivity-based routing needs somewhere to send content that must not leave the country. Without an in-house model that content can only be refused, which works but costs you the goodwill that makes staff accept the control.

COVERAGE

Where people use AI, and what covers it

WHERE THEY WORK	HOW IT IS COVERED	WHAT YOU GET
Company chat	Our chat app, company SSO	FULL masking, rules, audit
Your own applications	Point them at one address	FULL masking, rules, audit
Coding tools, desktop AI	Laptop agent repoints them	FULL becomes an ordinary governed request
ChatGPT or Claude in a tab	Browser extension checks before Send, personal accounts included	WARN & REDIRECT sensitive content stopped
Apps that cannot be inspected	Blocked, with a link to the approved tool	BLOCKED
Personal phone, home computer	Outside any company control	POLICY ONLY

We say the last row out loud. No product covers a personal phone on mobile data, and any vendor claiming otherwise is selling a gap you will discover later. What this covers is every company-managed device and every company system.

Limits worth knowing before you buy

Personal devices — `Not covered`
a phone on mobile data is outside any company control

Thai personal names — `Not yet detected`
Thai identity numbers are; names need a language model

Browser coverage — `Chromium only`
Chrome and Edge; others are blocked rather than guarded

Certificate-pinned apps — `Blocked, not inspected`
they cannot be examined, so they are refused

AI inside other software — `Vendor settings`
Notion or Slack AI is turned off in their admin, not here

Printed here rather than discovered later. A customer who finds a limit you hid stops believing the rest of the document.

CAPABILITIES

What runs on every message

CAPABILITY	IN PLAIN TERMS	STATUS
Access rules	Finance may only use the in-house model. Legal may not use one particular vendor at all. Set once, applied everywhere.	AVAILABLE
Personal data masking	Names, emails, phone and card numbers swapped for placeholders the model still understands.	AVAILABLE
Thai identity numbers	National ID, tax ID, phone and bank account, all validated by check digit, so ordinary reference numbers are not flagged.	AVAILABLE
Credential blocking	Passwords and access keys refused outright, never masked.	AVAILABLE
Sensitivity-based routing	Content too sensitive for an overseas provider is answered in-house automatically.	AVAILABLE
Conversation memory	The same person keeps the same placeholder all conversation, so follow-up questions work.	AVAILABLE
Audit trail and dashboard	Records what happened without recording the sensitive data, and refuses entries that still contain it. 90 day retention, with a dashboard for adoption, blocks and spend.	AVAILABLE
Answer safety screening	Checks the AI reply for harmful content and withholds it before it reaches the person.	AVAILABLE
Your confidential formats	Employee IDs, contract numbers, project codenames. Defined by your own team in a rule builder, tested in a live playground, with no code change.	SELF-SERVICE
Thai personal names	Thai ID numbers are found today. Thai names in free text are not yet.	ON THE ROADMAP

The last row is printed here deliberately. A customer who finds a gap you hid stops believing everything above it.

SPECIFICATIONS

Operating characteristics

Client interface	`OpenAI-compatible` most tools work unchanged
Sign-in	`Your corporate identity provider` OIDC / OAuth2, role-based
Added delay	`About a tenth of a second` before the answer starts
Personal data at rest	`None` memory only, encrypted, deleted on close
Deployment	`Your cloud or data centre` we never host your traffic
Scaling	`Kubernetes` scales with demand
Languages	`English and Thai` including mixed text
Compliance posture	`GDPR and PDPA` supports your obligations

What this changes for you

Use public AI without exposing customers. Personal details are replaced before the request leaves your network. The answer still reads naturally, because context is preserved rather than deleted.

Change AI providers without changing applications. Teams build against one address. Swapping provider, adding your own model, or failing over during an outage is a routing decision.

Answer the audit question in one place. One trail across every application, instead of a per-team reconstruction after the fact.

Keep spend visible before it surprises you. Per-team budgets and rate limits, with cost attributed to the team that generated it.

Deploying it

STAGE	TYPICAL DURATION	WHAT HAPPENS
Observe	2 weeks	Deployed in listen-only mode. You get a measured picture of who is using which AI, and with what kind of data. Nothing is blocked.
Route	4 weeks	The company chat goes live and tools are repointed. Most AI use is now governed, and staff notice nothing except a better chat app.
Protect	3 weeks	Masking, credential blocking and sensitivity-based routing switch on.
Extend	ongoing	Your own confidential formats, answer screening, and tuning against your real data.

`The order is deliberate. Measuring first means the rules you write are based on what your staff actually do, not on what anyone assumed. It also means the business sees value before it sees a restriction.`

Put every AI conversation behind one gate.

Your staff keep the tools they like. You get the rules, the evidence, and the confidence to say yes to AI.

`QiD Vanguard · Enterprise AI Gateway · Specifications subject to change`