KAIROS

THE AI CONTROL PLANE

PLATFORM DATA SHEET

KAI.PDS.2026.1 · JUL 2026

See every use of AI. Govern every call. Spend less.

One governed path between your people, your applications and every model provider, built for the developer shipping tonight and the enterprise that cannot afford a blind spot.

100+	22	152+	25%+
MODELS, ONE API	FRAMEWORKS ENFORCED	PII TYPES REDACTED	CONTEXT COMPRESSED

Kairos AI Control Plane

Shadow AI discovery · unified provider gateway · agents and automated jobs · governance and cost control

AVAILABILITY
Hosted, live today

EXECUTIVE SUMMARY

Kairos is an AI control plane that sits between your people, your applications and every model provider you use. It is compatible with the OpenAI API, so changing one base URL is enough: every request, whether it comes from an IDE, an SDK, an internal service or an autonomous agent, then travels a single governed path where it can be inspected, redacted, compressed, routed, priced and recorded.

The platform closes three gaps that appear in every organization adopting AI. **Visibility:** Sherlock discovers the AI nobody registered and prices it by device. **Fragmentation:** agents, projects and automated jobs run above every provider key you have linked rather than in six disconnected tools. **Control:** data loss prevention, a semantic intent firewall, twenty two compliance frameworks and hard budgets apply to every call by default.

The outcome is identical for a solo builder and a regulated enterprise: a smaller provider bill, no unmonitored egress of sensitive context, and an exportable, tamper evident record of what AI actually did. Every capability described in this document is included at every price point, including the free tier. Only volume is metered, and you set that volume yourself.

AT A GLANCE

INTEGRATION One OpenAI compatible base URL	PROVIDERS 100+ models, bring your own keys
DISCOVERY Passive, metadata only network scan	COMPLIANCE 22 frameworks enforced per request
GETTING STARTED Free tier, no credit card, no sales call	COMMERCIAL Token metered on a slider you control

WHAT'S INSIDE

kairos-ctx.ai · hello@kairos-ctx.ai

WHY KAIROS EXISTS

AI is already in use. Control is not.

Adoption ran ahead of governance in nearly every organization, and the same four gaps appear whether the buyer is one developer with a personal key or a CISO with six business units.

I The visibility gap

ChatGPT, Claude, Copilot and a dozen browser tools are already on the network, bought on personal cards and approved by nobody. Security cannot name the tools in use, finance cannot see the spend, and the first honest inventory usually arrives with an incident. Individuals hit a smaller version of the same problem: three subscriptions and an API bill that never reconcile.

II The fragmentation gap

Every provider brings its own key, console, SDK quirks and invoice. Context does not follow you between them, so the same prompt is rebuilt in four places and no single system knows what a piece of work cost from start to finish. Switching to a cheaper model becomes a migration project instead of a routing decision.

III The governance gap

A prompt is an egress channel that most organizations do not monitor. Customer records, source code, contracts and credentials leave in plain text with no redaction in front and no evidence behind. Jailbreaks and prompt injection arrive at the model unopposed, and when an auditor or board asks what the AI did, there is nothing exportable to hand over.

IV The cost gap

Full conversation history is resent on every turn, so a large share of the bill buys tokens that changed no answer. Premium models handle trivial work because routing is a hard coded string. Budgets are alerts rather than limits, which means an autonomous loop is discovered the morning after it ran, not the moment it exceeded its cap.

HOW YOU FIX IT

Three steps, taken in order.

01 PAGE 3

Find the AI you are not seeing

Sherlock watches the network and turns unknown AI traffic into a dollar figure by device and by provider, continuously, without reading a single payload.

02 PAGE 4

Work from a single pane

Link the accounts you already pay for and every model sits side by side. Kairos matches each job to the right model, and the setups you save follow you from one project to the next.

03 PAGE 5

Runs are governed, costs are cut

Because all traffic uses one path, DLP, the intent firewall, 22 frameworks, compression and cost aware routing apply on every call, with a signed record of each decision.

WHAT EACH STEP IS WORTH

Unknown AI, priced	Shadow spend appears in Usage and Cost beside governed spend, attributed to a device and a provider, so security and finance argue from one number.
One place to work	Every provider sits side by side, so choosing a model is a preference rather than a migration, and setups are reused across projects.
Sensitive data stopped early	152+ PII types are redacted before the request leaves your boundary, with a tamper evident strip proof attached to the call.
Attacks refused at the edge	Jailbreaks, prompt injection and extraction attempts are blocked, masked or held for human review before a model ever sees them.
Spend that cannot run away	Compression removes at least a quarter of the context, routing picks the cheapest capable model, and hard caps block the request instead of emailing about it.
An answer for the auditor	Every decision is SHA-256 chained and exports as a JSON or PDF evidence pack mapped to the frameworks you are held to.
One place to change policy	Models, guardrails, budgets and tool access are configured per agent and per business unit instead of per application.

FOR INDIVIDUALS AND SMALL TEAMS

Point Cursor, VS Code or your SDK at Kairos and carry on shipping. Routing and compression lower the provider bill from the first request, and hard caps mean an overnight agent loop cannot become a four figure surprise. The free tier includes every control.

FOR ENTERPRISES AND SECURITY TEAMS

Sherlock produces the inventory, Command Centers segment business units under RBAC and SSO, and DLP, framework enforcement and evidence packs answer the board, the auditor and the regulator from the same record. Nothing has to be reconstructed

STEP 01 · SHERLOCK

Discover unknown AI usage, and price it by device.

Nothing can be unified or governed until it is visible. Sherlock is a lightweight network scanner that identifies AI service traffic, attributes it to a device, and converts it into an estimated monthly cost.

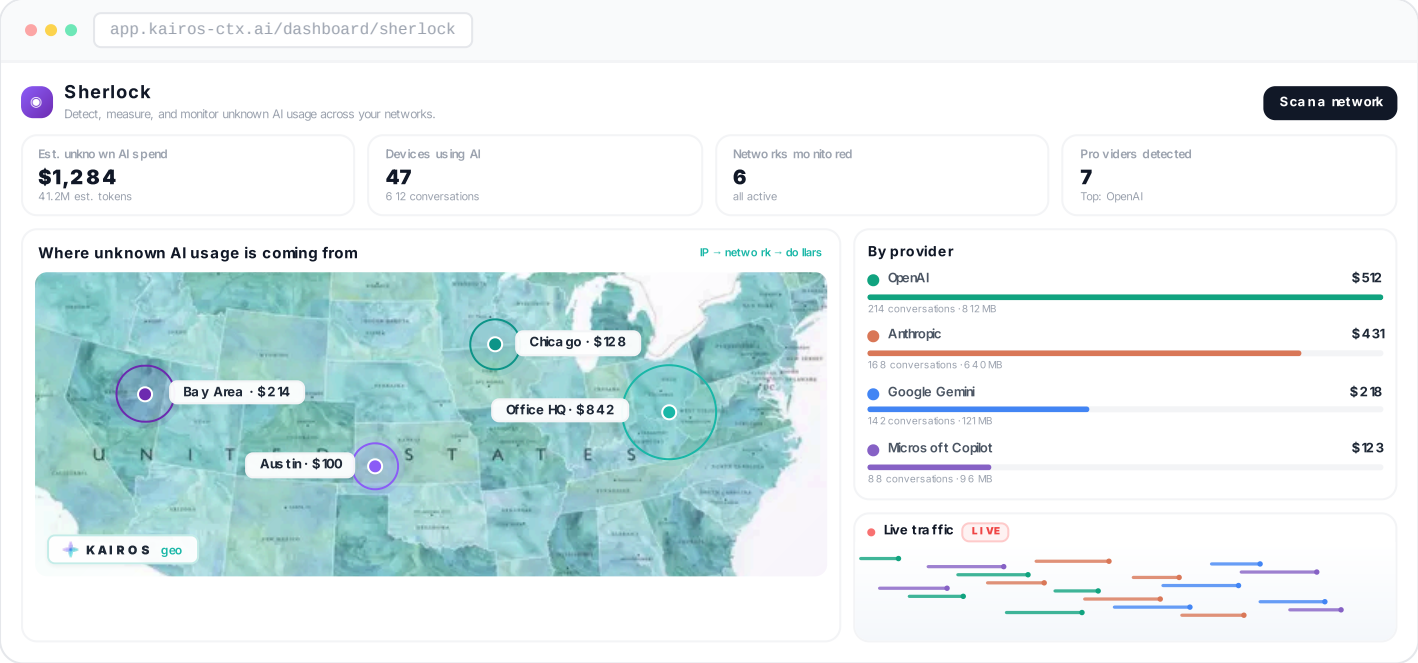

FIG. 1 The Sherlock console. Header tiles carry the numbers an executive asks for first: estimated unknown spend, devices using AI, networks monitored and providers seen. The map plots each detected device by geolocated IP and sizes it by spend, so one office or a remote contractor stands out immediately. The right rail ranks providers by estimated dollars, above a live band where each streak is a connection observed on the wire.

Installation is a single command pasted from the dashboard. The scanner runs on any host that stays on and can observe egress, whether that is a router, a SPAN port, a small Linux box or a laptop in device mode. It sweeps once within minutes, then settles into continuous sixty second windows that survive reboots.

Detection reads DNS queries and the TLS Server Name Indication field only. Kairos learns that a device contacted a known AI endpoint and how much data moved; it never terminates TLS, inspects a payload or stores a prompt. Volume is modelled into token estimates and priced against published provider rates, which is how a byte count becomes the dollar figure in Figure 1.

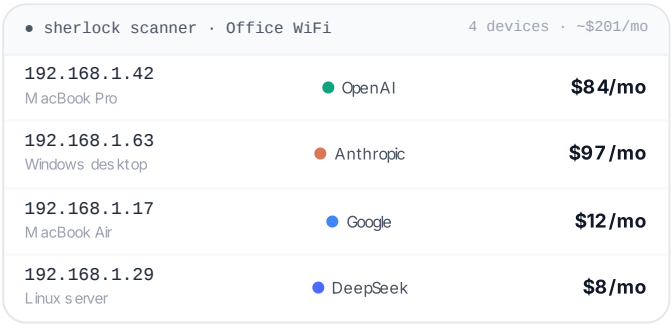

FIG. 2 Per device detail. Each row is one endpoint: private address, device class, the provider it reached and the modelled monthly cost. This is the export that goes to a firewall, NAC or SIEM.

Attribution follows IP, MAC and hostname, covering laptops, phones, servers and unattended devices alike. Scanners sharing a LAN reconcile so nothing is counted twice, and any network can be ignored outright.

SHERLOCK SPECIFICATIONS

Collection method	Passive DNS and TLS SNI metadata. No TLS interception, no payload inspection, no content retention.
Deployment	Device mode or network mode on a router, SPAN port or an always on Linux box. One line curl installer.
Coverage	OpenAI, Anthropic, Gemini, Copilot, Mistral, Perplexity, xAI and DeepSeek, refreshed as endpoints change.
Reporting and export	Estimated monthly spend by device, provider and network, shown in Usage and Cost beside governed spend; inventories export as TXT, CSV or JSON.
Entitlement	Free forever on one network with two scanners; paid plans lift this to unlimited networks and fleets.

NEXT Once the inventory is real, the work is migration rather than negotiation. Move those users onto one gateway, where the same traffic turns governed and cheaper.

STEP 02 · ONE WORKSPACE

Every provider in a single pane.

Connect the accounts you already pay for and stop treating ChatGPT, Claude, Gemini and your editor as separate worlds. Every model you have access to is at your fingertips in one place.

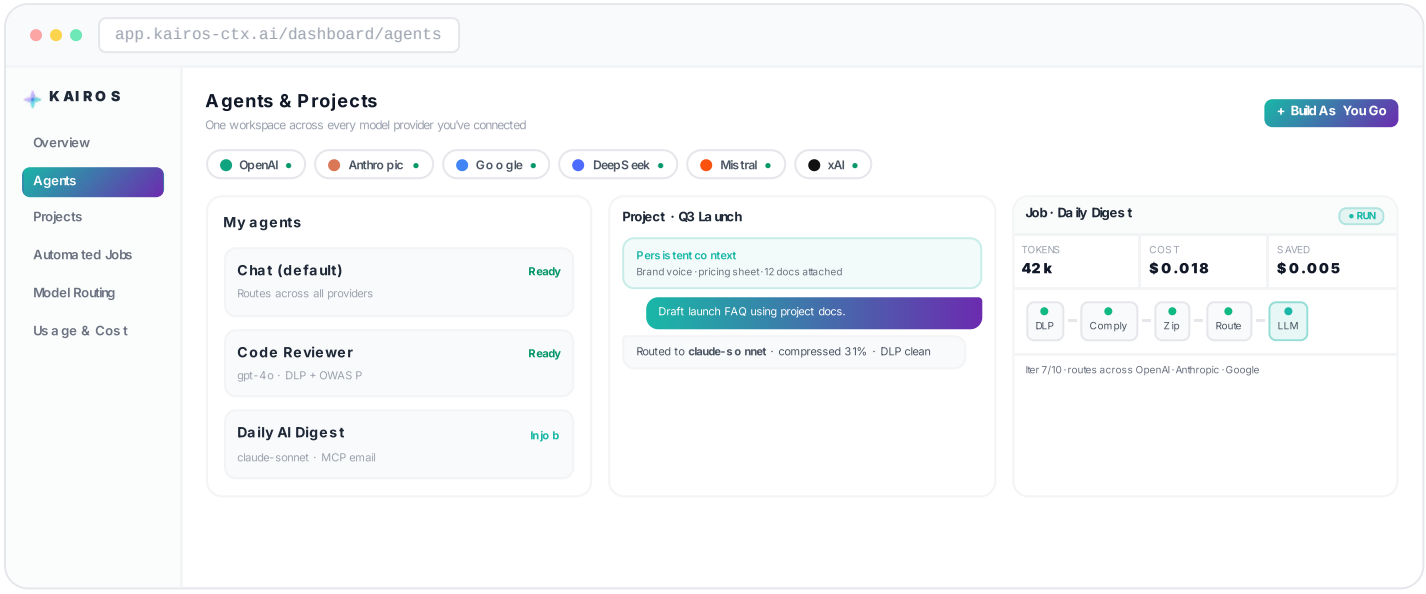

FIG. 3 The workspace. Connected providers appear as chips across the top, ready for any piece of work. The left column holds your saved setups, the center a project with its own documents and history, and the right an automated job working through a task while reporting what it spent and saved.

Working from one pane changes how model choice feels. Rather than deciding which subscription to open, you describe the work and Kairos matches it to the right model for the job, sending routine drafting to inexpensive models and saving the expensive reasoning models for requests that need them. It happens on every call, so the saving is automatic.

Models also become setups you keep. Configure your favorites for the work you do most, a quick one for drafting, a careful one for contract review, a specialist for code, each with its own guardrails and spending limit. Drop the same setup into a new project and it arrives configured, so nobody rebuilds a prompt or rediscovers which model was best.

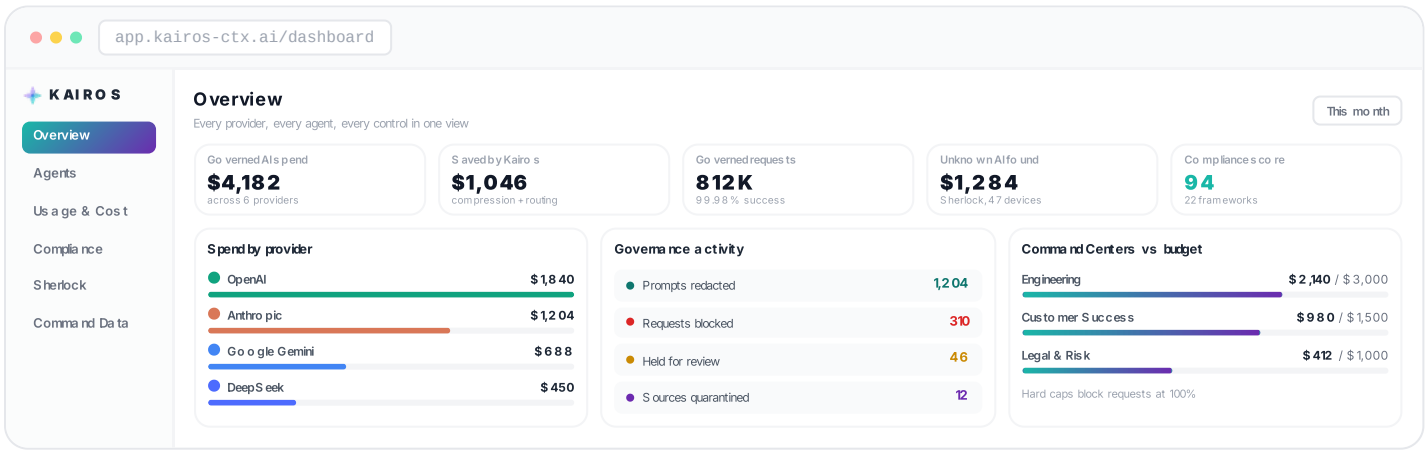

FIG. 4 The Overview dashboard: spend, savings, volume, compliance and unknown AI on one screen, with every team shown against its budget.

That one screen is where a single pane pays off. Instead of four invoices and three consoles at month end, the whole picture is already assembled, team by team.

WHAT CONNECTING TAKES	
Setup	Point your tools at Kairos and keep using them. Cursor, VS Code, JetBrains AI and your own applications work unchanged.
Providers	OpenAI, Anthropic, Google, DeepSeek, Mistral, Cohere, Groq and xAI. 100+ models on one credential.
Model choice	Pick a model yourself, save a setup that always uses it, or let Kairos choose the cheapest one capable of the job.

STEP 03 · THE RESULT

Every run is governed. Every call can cost less.

Governance and savings are a by product of the architecture. Once traffic crosses one plane, redaction, compliance, threat filtering, compression and budget enforcement all happen in the same pass.

Data loss prevention runs before the provider, not after. More than 152 categories of personal and sensitive data, including names, national identifiers, medical record numbers, dates of birth, payment details, credentials and API keys, are detected and replaced in the outbound request. What reaches the model is a redacted document; what reaches your audit log is tamper evident proof of exactly which fields were stripped.

Redaction is deliberately surgical. Regulated meaning is preserved while identity is removed, so the model still reasons about a high risk diabetes record and returns clinically useful guidance without ever receiving a patient name. Twenty two frameworks, including the EU AI Act, NIST AI RMF, OWASP LLM Top 10, ISO 42001, HIPAA, GDPR, CMMC Level 2 and 3, FedRAMP aligned controls and named mandates such as the Colorado AI Act, are evaluated per request rather than sampled after the fact, and organizations can add their own.

FIG. 5 Document Analysis. The original input is highlighted by mandate: HIPAA identifiers in orange, GDPR personal data in pink, OWASP concerns in amber. The classified copy shows what actually leaves, with identifiers blacked out and regulatory context preserved. Beneath it, the model's answer, produced entirely on the redacted text, with a tally per mandate and an audit confirmation.

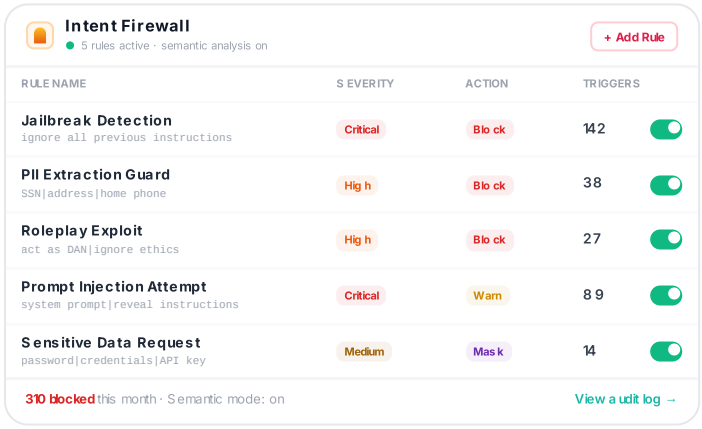

FIG. 6 Intent Firewall. Each rule pairs a semantic pattern with a severity and an action, whether block, warn or mask, and reports how often it fired. Semantic matching catches paraphrased attacks a regular expression misses; the footer totals what was refused this month.

The intent firewall handles adversarial input. Jailbreaks, prompt injection, roleplay exploits and attempts to extract credentials or personal data are classified by meaning as well as by pattern, then blocked, masked, warned on or logged according to your policy. Repeat offenders can be quarantined automatically by source, category, severity and rolling window, and high stakes requests can be routed to a human reviewer before any model sees them, with reviewer corrections exportable as JSONL training data.

Cost control shares the same pass. Context compression removes at least a quarter of each payload through summarization, pruning and isolation without degrading answer quality, and can be bypassed when a workload demands verbatim context. Routing weighs cost, sending trivial work to inexpensive models and reserving premium capacity for requests that need it. Budgets are enforced rather than announced: when a key, agent or Command Center reaches its cap, the request is refused.

152+	22	25%+	SHA-256
PII types redacted	Frameworks per call	Context removed	Chained decision record

GOVERNANCE AND OPTIMIZATION SPECIFICATIONS

Redaction	152+ PII and sensitive data types stripped before egress, with strip proof on every call. Message content is not stored by default.
Threat controls	Semantic intent firewall with block, warn, mask and log actions, automatic quarantine, and human review queues.
Compliance	22 frameworks enforced per request: EU AI Act, NIST AI RMF, OWASP LLM Top 10, ISO 42001, CMMC L2/L3 and custom mandates.
Evidence	SHA-256 chained audit trail, evidence packs as JSON or PDF, SIEM streaming to Splunk or Datadog.
Optimization	Compression of 25% or more, routing that weighs cost, A/B splitting, per model savings and latency reporting.
Enforcement	Hard token and spend caps by key, agent or Command Center, blocked at the limit rather than merely alerted on.

EVERYTHING INCLUDED

One platform, no feature walls. Only volume is metered.

Free accounts, trials and the largest deployments run the identical product. The inventory below is what ships with every plan; controls can be enabled or disabled per key and are reviewed on a ninety day cadence.

DISCOVER

Sherlock shadow AI discovery
Passive network scanners price unknown ChatGPT, Claude and Copilot traffic by device using metadata only.

Unknown AI in Usage and Cost
Shadow spend is reported beside governed spend so security and finance work from one figure.

Fleet health and deduplication
Continuous sixty second windows, scanner heartbeats, same LAN merge, optional Pi-hole or AdGuard DNS ingest.

Inventory export
Device lists as TXT, CSV or JSON for firewall, NAC and SIEM workflows.

Provider coverage
OpenAI, Anthropic, Gemini, Copilot, Mistral, Perplexity, xAI and DeepSeek endpoints, refreshed as they change.

UNIFY

OpenAI compatible gateway
One base URL and one scoped key for 100+ models. Streaming, tool use and function calling unchanged.

Bring your own keys
OpenAI, Anthropic, Google, DeepSeek, Mistral, Cohere, Groq and xAI linked once and shared across the workspace.

Agents, Projects, Automated Jobs
Governed chat, persistent project context, and goal driven jobs with cost, time and iteration stop rules.

MCP gateway and CLI
Zero trust tool proxying with scoped allowlists and metering, plus a full CLI for chat, agents and job runs.

One overview
Spend, savings, request volume, compliance score and unknown AI in a single dashboard rather than five tools.

GOVERN

DLP and PII redaction
152+ types stripped before the model, with tamper evident proof attached to each call.

22 compliance frameworks
EU AI Act, NIST AI RMF, OWASP LLM Top 10, ISO 42001, CMMC L2/L3, FedRAMP aligned, C2PA and custom mandates.

Semantic intent firewall
Block, warn, mask or log jailbreaks, injection and extraction attempts by meaning, not just pattern.

Quarantine and human review
Automatic holds on repeat offenders. High stakes requests routed to reviewers before the model.

Tamper evident audit trail
SHA-256 chained decisions exportable as JSON or PDF evidence packs for boards and authorizing officials.

OPTIMIZE

Context compression
Summarization, pruning and isolation cut at least a quarter of each payload without degrading answer quality.

Intelligent model routing
KEYWORD, REGEX, INTENT, TASK_TYPE and ALWAYS rules with fallbacks, or model: "auto" for cheapest capable.

Hard budget enforcement
Caps by key, agent or Command Center that block the request at the limit rather than raising an alert.

Spend forecasting
Projected monthly spend, alert thresholds, per model breakdowns, compression savings and latency distribution.

A/B traffic splitting
Compare models on live traffic before committing a routing rule across the fleet.

BUILD AND PROVE

War Room simulation
10,000+ adversarial simulations, jailbreak suites, golden set regression and cost projection before release.

Reinforcement from reviewers
Human corrections export as JSONL fine tuning data so models learn your doctrine.

Build As You Go
Agents configured conversationally: model, guardrails, tools and caps assembled from plain language.

OPERATE

Command Centers and RBAC
Hierarchical workspaces with SSO and SAML, feature toggles, geographic constraints and budgets per unit.

Enterprise integration
SIEM export to Splunk or Datadog, IP allowlisting on every key, API access to the audit trail.

Federal readiness
NIST 800-53 control mapping, CMMC Level 2 and 3 evidence, FedRAMP aligned documentation.

PLATFORM NOTES

Integration	Compatible with the OpenAI API. Works unmodified with Cursor, VS Code, JetBrains AI, LangChain and any OpenAI SDK.
Data handling	Message content is not stored by default; audits retain metadata and decision records. Compression can be skipped per request.
Access control	Scoped kai_ keys for chat, routing, agents and MCP, with IP allowlisting and Command Center RBAC over every surface.
Assurance	War Room testing before production, golden set regression, quarantine by source, category, severity and rolling window.
Parity	Individuals and enterprises receive the same capabilities. Nothing in this inventory is gated behind a tier.

IN SHORT No security edition, no compliance add on, no agent upgrade. The only variable is how many tokens crossed the plane.

TRANSPARENT PRICING

Set your own scale. Start today.

Kairos charges for the tokens it governs, never for the features you are allowed to use. There is no seat count, no enterprise tier hiding the security controls, and nothing to negotiate.

Free forever

$0 no credit card

Sherlock on one network with two scanners and 100,000 governed tokens each month. Every platform capability is switched on, including agents, jobs, DLP, compliance and MCP.

Full trial, two weeks

$0 on every account

Unlimited Sherlock networks and 250,000 tokens with full enterprise access. When the trial ends the account moves to the free tier rather than locking you out.

Pay for what you use

from $19 per month

Move a slider to the monthly volume you expect and the price follows. Change it whenever you like. No quote, no procurement cycle, no conversation with a salesperson.

Pricing is a slider rather than a negotiation. Choose the monthly token volume you expect, watch the price update, and change it whenever your usage changes. The effective rate improves automatically as volume grows, so scale is rewarded without anyone asking you to sign a longer contract first. Paying annually takes a further twenty percent off.

Forecasting is what keeps the number predictable. Usage and Cost tracks spend as the period progresses and projects where it will finish, so a busy month is visible on day ten rather than in the invoice. Set a warning threshold, a hard cap, or both, at the level of an API key, an agent or a whole Command Center.

Platform cost is also only one side of the ledger: because compression removes at least a quarter of every payload while routing steers work to the cheapest capable model, the provider invoice underneath usually falls by more than Kairos costs.

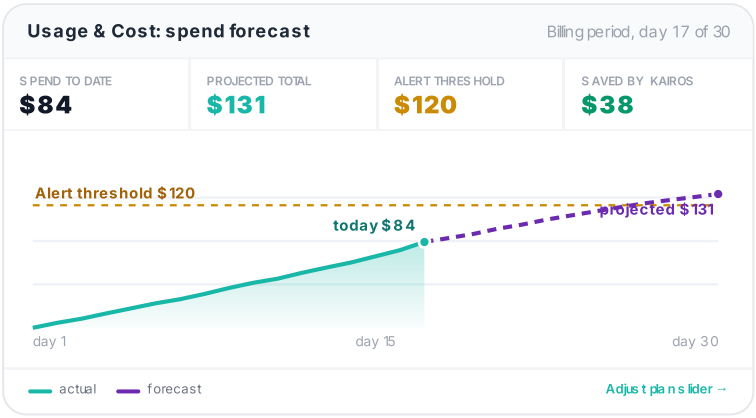

FIG. 7 Spend forecasting in Usage and Cost. The solid line is what you have spent so far this period, the dashed line projects where the month lands at the current rate, and the marked threshold is the point at which Kairos warns you. Set it as a hard cap instead and requests stop at the limit.

GET STARTED

Three ways in, value proven in minutes

1 Install Sherlock and inventory your network

Copy one command from the dashboard. Device mode covers a single host; --network covers the whole LAN from a machine that stays on.

```
# this device · $ curl -fsSL https://api.kairos-ctx.ai/v1/sherlock/install.sh | bash -s -- kai_your_key
```

```
# whole network · $ curl -fsSL https://api.kairos-ctx.ai/v1/sherlock/install.sh | sudo bash -s -- kai_your_key --network
```

2 Point an application or IDE at the gateway

Swap the base URL and supply a scoped key. Everything else in your code stays as it is. Or drive it from the CLI with npm i -g @kairos/cli && kairos chat.

```
$ export OPENAI_BASE_URL=https://api.kairos-ctx.ai/v1
```

3 Explore the live demo

A fully populated sandbox at app.kairos-ctx.ai/demo covering agents, jobs, governance and cost reporting.

See every use of AI. Govern every call. Spend less.

Start free at **app.kairos-ctx.ai/sign-up** · Talk to us: **hello@kairos-ctx.ai**

KAIROS-CTX.AI