

AI Cost Observability.



Control your AI spend as part of holistic governance.

Why this matters now.

AI is now one of the fastest-growing line items in the enterprise budget, and one of the least understood. Each AI provider sends a bill with a total token count. But the pricing varies by provider, model family, version, input tokens, output tokens, with almost no understanding on how to predict any of that. You can't forecast what you can't break down, cut waste you can't see, or defend a number that has no detail behind it. Cost observability is a key missing component of current AI governance, and it starts with the token, the unit you're actually billed on.

What observability means for AI cost.

The ability to answer, at any moment: how much, on what, by whom, and is this normal. FireTail builds that from the token up.

THE QUESTION	WITHOUT OBSERVABILITY	WITH FIRETAIL
How much are we spending?	One total per provider	Consumption across every provider in one view
On which tools?	Unknown	Broken down across all tools in use
Driven by whom?	No owner	Attributed to the individual
Is this spike normal?	Found at invoice time	Visible in the trend as it happens
What's it worth vs its risk?	Two separate tools	Cost and risk on the same record

Cost as a security indicator.

One often overlooked aspect of tracking costs is looking for spikes in consumption. Sometimes, these are indicators that an employee's account, or an AI agent, has been compromised or has compromised credentials (API keys, tokens, secrets). And because this runs on the same platform that discovers and scores your AI, that spike doesn't sit in isolation: you see it next to the model's risk and inventory, so a cost anomaly and a security signal become the same alert instead of two teams noticing two different things.

From invoice to insight.

See the spend, break it down, and give it an owner.

01 SEE IT

A flat monthly total hides the truth: your spend surges. Each spike is money already gone, and if the invoice is where you see it, you're finding out too late to act. Watching the trend as it forms is the line between managing spend and reconciling it.

02 BREAK IT DOWN

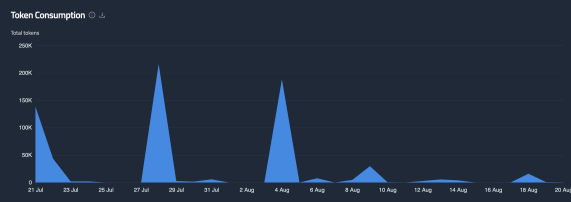
One number gives you nothing to pull on. Split into the individual tools actually running, spend becomes a set of decisions: kill the tool nobody uses, question the one that spiked, consolidate the ones that overlap. Attribution is what turns a bill into an action.

03 OWN IT

Spend with no owner never gets managed. When consumption maps to real people and teams, AI usage finally behaves like every other budget item: someone is accountable, patterns get questioned, and outliers get a conversation instead of a shrug. That is how cost actually comes in line with budget.

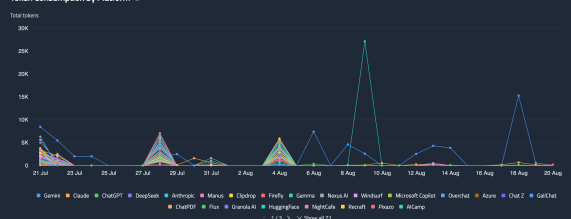
TOKEN CONSUMPTION

Token Usage



TOKEN CONSUMPTION BY PLATFORM

Token Consumption by Platform



TOP CONSUMERS

Top Consumers



THE TAKEAWAY

You can't manage what you can't see, and you can't defend what you can't explain. FireTail makes AI consumption visible, attributed, and accountable, so AI spend gets run like a business line, not absorbed as a mystery.

Schedule a Technical Deep Dive.

See FireTail in action across your own codebase and get a complete AI model inventory in just 15 minutes.

🌐 www.firetail.ai
 @ contact@firetail.ai



ABOUT FIRETAIL

FireTail is a recognized leader in AI security, with multiple awards, analyst citations and Top 4 finalist finishes at both Black Hat USA & Asia in 2025. The company is headquartered in Northern Virginia, USA with additional offices in Dublin, Ireland and Helsinki, Finland. FireTail is backed by leading investors Paladin Capital, SecureOctane, General Advance and Zscaler. For more information, please visit www.firetail.ai