Spectro Cloud

PaletteAI

PaletteAI Inference Launchpad

Cut token costs by running inference locally, with frontier fallback as needed. PaletteAI Inference Launchpad gives you a validated solution to lower token spend, without having to build and operate the full AI stack yourself.

The token cost boom

AI adoption is scaling faster than token cost controls. Coding assistants, agents, and AI-powered workflows have moved from pilots to broad adoption, and token and API costs are rising with usage.

Many organizations can't see where their spend is going, which teams are driving it, or which requests actually need a frontier model.

You can't slow the innovation that AI enables — but you need control over cost, model usage, data exposure, and governance.

Enterprises trying to control token costs have three options:

1. Build a DIY private stack: Maximum control, but slow to deliver and operationally heavy.

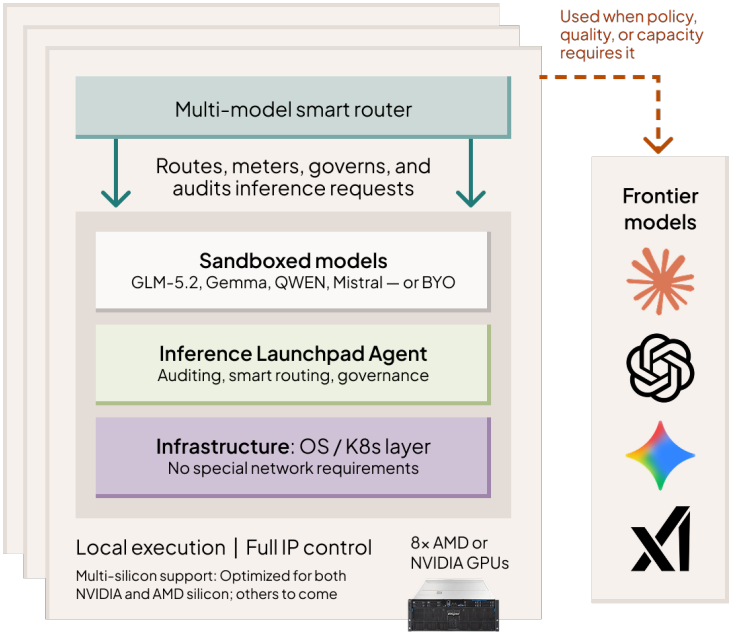
2. Use AI cloud / inference-as-a-service: Quick to start, but costly and lacking full-stack control.

3. Deploy Inference Launchpad: A fast, private path to local-first inference with KV cache efficiency, governance, frontier vs local model choice and built-in routing.

Meet Inference Launchpad

PaletteAI Inference Launchpad gives you fast time-to-value with full control over your data and token usage.

Validated for AMD's Instinct family of GPUs and NVIDIA GPUs, the Inference Launchpad enables data and token spend control while giving you choice to deploy it on-premises or as a hosted solution.



Inference Launchpad puts you in charge of your token spend

Intelligent model routing

- Route by user, team, app, workload, endpoint, or policy
- Fall back to frontier models based on quality, capacity, or policy

Governed consumption

- Set quotas, rate limits, and time-bound controls
- Audit usage, routing decisions, and policy enforcement

On-prem or hosted

- Get started quickly with a pre-validated solution
- Choose the right deployment option for your organization

Inference Launchpad: the best of both worlds

Packaged private AI: Faster than DIY, more control than AI clouds. Local models run in a sandbox without internet access to guarantee data sovereignty

Intelligent multi-model routing: Use local or frontier where they make sense. Choose open models out of the box, or bring your own.

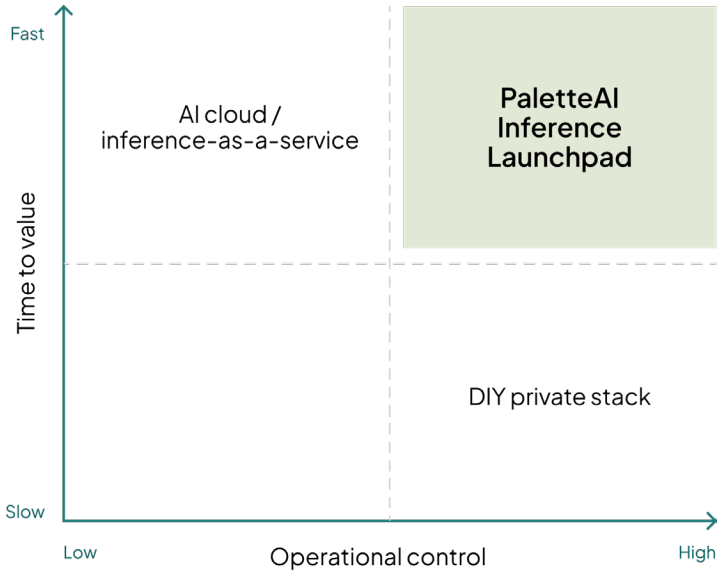
Token governance: Meter, quota, rate-limit, and audit usage before costs get out of control.

Hardware choice: Launchpad supports both AMD and NVIDIA GPUs.

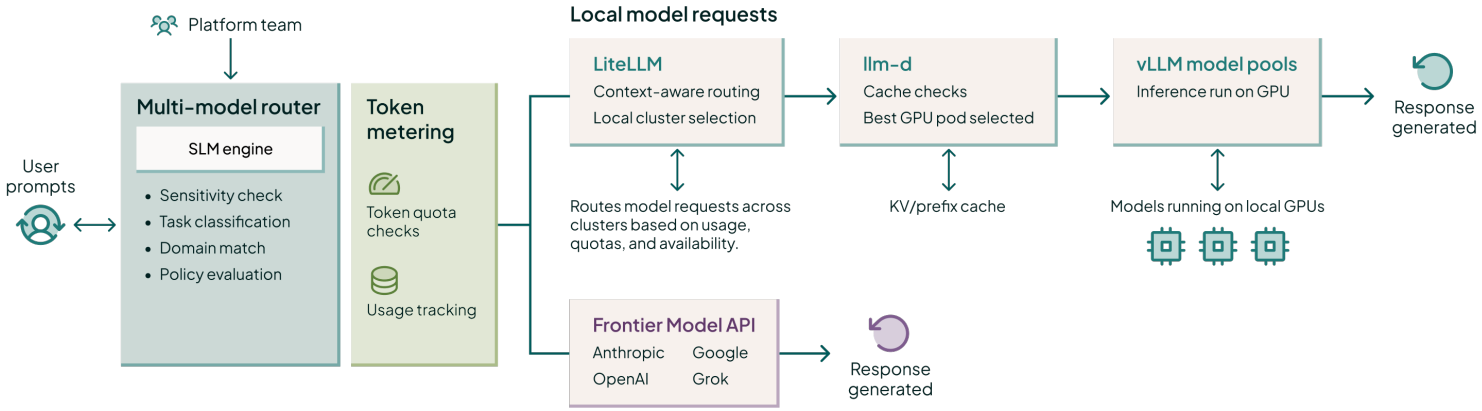
Scale-out architecture: Start self-contained on a single server and add servers as capacity needs grow.

Local-to-central management: Manage locally and seamlessly transition to central management for fleet-wide consistency.

Deployment options: Use as a managed service or deploy on-premises.



How it works, step by step



Illustrative TCO example

Map out your own savings with our calculator: **spectrocloud.com/ai-tco**

$1.2M
Current annual token spend

$370k
Inference Launchpad annual cost inc hardware capex

$830k
Annual savings: 70% reduction

Assumptions: 50 developers at $2K monthly spend per developer, using 1× Supermicro server with 8× AMD or NVIDIA GPUs. Calculation assumes annualized infrastructure cost and Inference Launchpad subscription, with 10% frontier fallback.

Your next steps

Connect with us to quantify your current spend, estimate local-routing potential, and define your first Inference Launchpad deployment path. Book a meeting at **spectrocloud.com/get-started**.

Spectro Cloud