Open Source in Finance, New York, November 4-5
#OSinFinance | #OSFF2026

From Memory Budget to Cost Per Token: Serving Agentic Traffic with Large Open Weight Models



Yuchen Fama, Ph.D. Senior Principal
Product Manager, Red Hat
yfama@redhat.com



Ashish Kamra, Ph.D. Senior
Manager, Red Hat
akamra@redhat.com

Red Hat

Outline

01

Pareto Chart of Tokenomics

Cost from throughput per GPU vs. interactivity.

02

Open-Weight Economics

Cost comparison of self-hosted open models vs. proprietary APIs.

03

Challenges of Agentic Workloads

Serving and optimizing at enterprise production scale is hard

04

Platform Solution & Takeaways

How to improve self-hosted token cost

Red Hat Webinar:

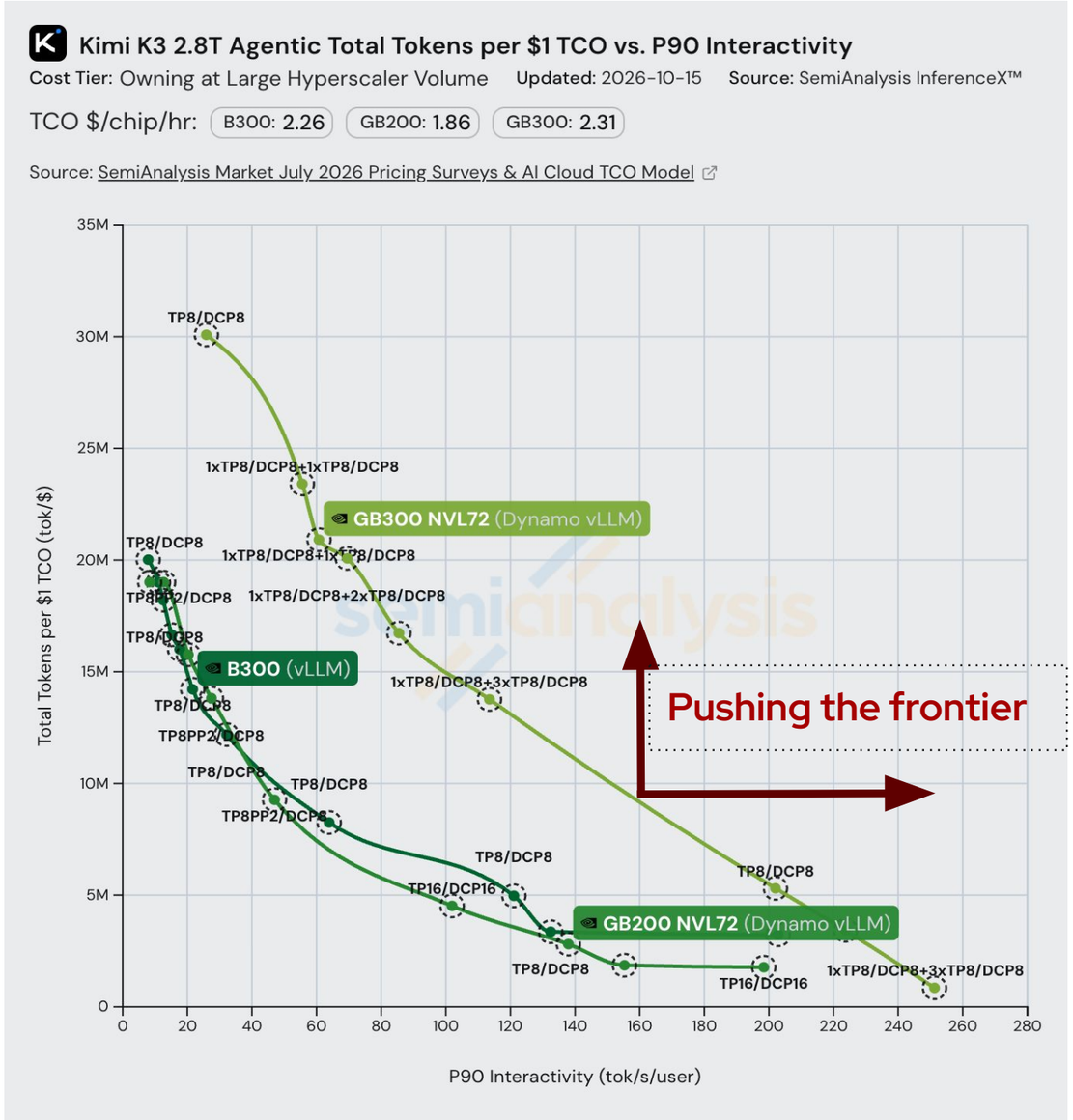
Mastering AI inference performance at scale

November 18, 2026, 2:00 p.m. EST (UTC -5)

- Tyler Michael Smith, vLLM Core Maintainer, Chief Architect, Red Hat

-Yuchen Fama
vLLM and GuideLLM Contributor,
Senior Principal Product Manager, Red Hat

Pareto Chart of Tokenomics



inferencex.semianalysis.com/inference/kimi-k3

How to Read the Pareto Chart?

- **Y-axis:** Cost or throughput per GPU
- **X-axis:** Interactivity per request (how fast is an agent?)
- **Comparison:** Evaluates different configurations and amounts of load on a unified baseline

Open-Weight Economics

- **Dramatic Cost Cut:** Self-hosted open models (e.g. Kimi K3) can be up to 40× cheaper at 95% performance after optimizations on your own GPUs
- **75–90% Tokens Will Be Processed by Open Models** Driven by data privacy, IP leakage concerns, geopolitical risks, and massive budget savings.

– Arvind Krishna, IBM CEO (from All In with Nikesh Arora)

How we get >3× throughput gain at matched interactivity

https://vllm.ai/blog/2026-09-15-kimi-k3-dspark

https://vllm.ai/blog/2026-09-13-kimi-k3-performance-optimization

Red Hat

Fact or Fiction?

The Industry Premise

“Self hosted Open weight models are cheaper than comparable proprietary API access based models”

Is this statement universally true, or does economics depend heavily on GPU utilization, batch size, and request interactivity?

Red Hat

Self-Hosted vs Frontier API Cost Comparison

Print / Save PDF | Close

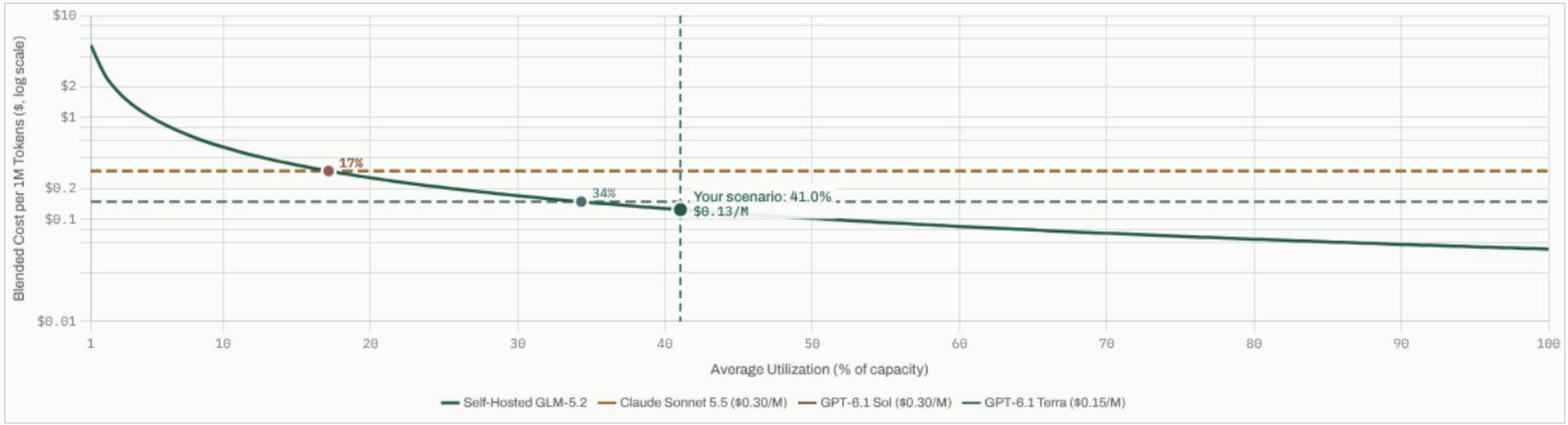
GLM-5.2, B200, 155k:1k ISL:OSL, 92.8% cache

11/2/2026

At 41.0% average utilization, self-hosting costs $0.13/M, below every enabled frontier API. The nearest is GPT-6.1 Terra at $0.15/M.

Parameters

Self-Hosted Model	GLM-5.2
GPU Type	B200
GPUs / Node	8
Nodes	1
Total GPUs	8× B200
$ / Node / Month	$21,608.00
Concurrency (sessions)	32
Cache Hit Rate	92.8%
Avg ISL (tokens)	155,205
Avg OSL (tokens)	997
Total Goodput (tok/s)	160,111
Output tok/s per User	98.30
Number of Users	200
Sessions / User / Day	1.00
Avg Model Calls / Session	251.47
Model Calls / User / Day	251.47
Working Days / Month	22
Annual TCO	$259,296
ISL:OSL Ratio	155.7:1
Self-Hosted Cost/M (100%)	$0.05/M tokens
Monthly Token Demand	172.8B tokens/month
Derived Avg Utilization	41.0%
Self-Hosted Cost at Util.	$0.13/M tokens



Breakeven summary

Frontier model	Blended $/M	Breakeven util.	Monthly API bill at estimated demand
Claude Sonnet 5.5	$0.30	17.16%	$51,697
GPT-6.1 Sol	$0.30	17.16%	$51,697
GPT-6.1 Terra	$0.15	34.31%	$25,848

Generated by Self-Hosted vs Frontier API Cost Comparison. Prices from OpenRouter.

Source: https://openshift-psap.github.io/self-hosted-llm-cost-calculator/

Red Hat

Scenario – GLM 5.2 on 8xB200

Infrastructure

GPU type	GPUs per node
B200	8
Number of nodes	**$/node/month**
1	21608

Workload shape

Concurrent sessions	Cache hit rate (%)
32	92.8
Avg ISL (tokens)	**Avg OSL (tokens)**
155205	997

Serving performance

Total goodput (tok/s)	Output tok/s per user
160111	98.3

Demand estimate

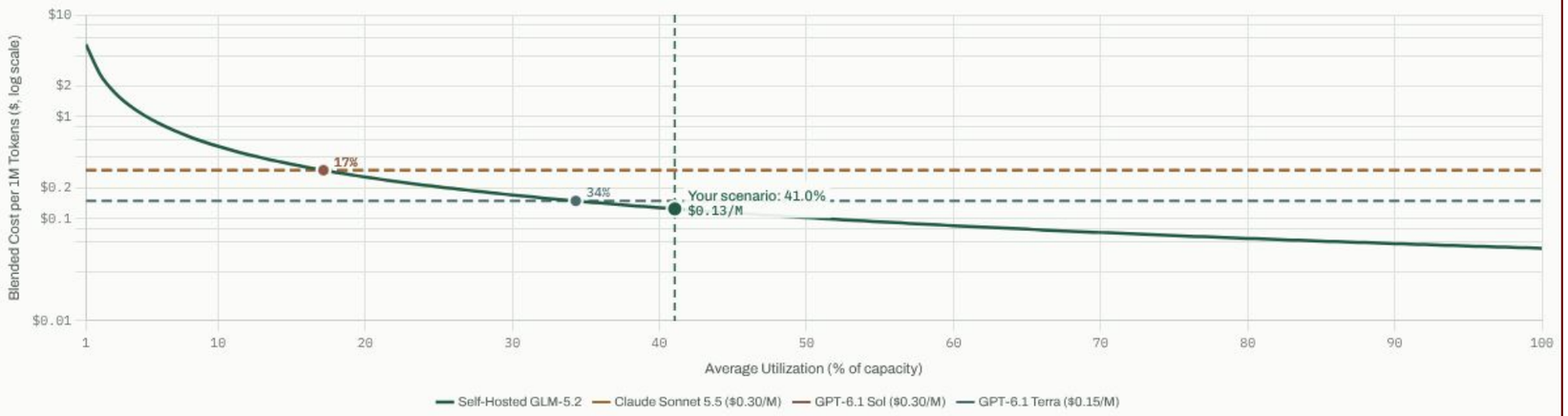
Number of users	Sessions / user / day
200	1
Avg model calls / session	**Working days / month**
251.47	22

At **41.0%** average utilization, self-hosting costs **$0.13/M**, below every enabled frontier API. The nearest is GPT-6.1 Terra at $0.15/M.

Self-hosted at demand	Cheapest API	Fleet bill	Utilization
$0.13	$0.15/M	$21.6k	41.0%
per million tokens	GPT-6.1 Terra	vs $25.8k cheapest API	of measured capacity

Cost per million tokens by utilization

Log scale



Breakeven summary

Frontier model	Blended $/M	Breakeven	Monthly API bill
Claude Sonnet 5.5	$0.30	17.16%	$51,697
GPT-6.1 Sol	$0.30	17.16%	$51,697
GPT-6.1 Terra	$0.15	34.31%	$25,848

Red Hat

Let’s do the numbers

100% GPU Utilization (Maximum Capacity)

Infrastructure Cost

Infra TCO = 1 × $21,608/mo × 12 = $259,296/yr

Serving Throughput

Total Goodput = 160,111 tok/s (measured)

Annual Token Volume

Annual Tokens = 160,111 × 31,557,600s = 5,052.7B

Theoretical Cost Floor (100% Utilized)

$0.05 / 1M tokens

($259,296 ÷ 5,052.7B) × 10^6

Demand Estimate & Real World Cost

1. User Workload Profile

- **Tokens / Call:** 156,202 (155.2k ISL + 997 OSL, cached)
- **Calls / User / Day:** 251.47 (1.00 session × 251.47 calls)
- **Monthly / User: 864.2M tokens** (251.47 × 22 days × 156.2k)

2. Fleet Balance (200 Active Users)

- **Monthly Demand:** 200 × 864.2M = **172.8B tokens**
- **Monthly Supply:** 160,111 tok/s × 2,629,800s = **421.1B tokens**
- **Average Utilization: 41.0%**

Actual Self-Hosted Cost (at 41% Utilization)

$0.13 / 1M tokens

($259,296 ÷ (172.8B × 12)) × 10^6

Red Hat

Play with the calculator

openshift-psap.github.io/self-hosted-llm-cost-calculator

Calculates and plots blended cost per million tokens for self-hosted GPUs against closed frontier APIs to identify breakeven points.

Live Frontier Pricing

Real-time pricing data fetched directly via OpenRouter API.

Scenario Presets

Built-in presets from InferenceX benchmarks, defaulting to GLM-5.2 on B200.

At-a-Glance Verdict

Instant summaries of total costs, monthly bills, utilization, and breakeven conclusions.

Workload-Aware Costs

Calculates blended costs by factoring in custom token mixes and cache hit rates.

Measured Capacity

Accurately models capacity requirements based on direct fleet goodput metrics.

Fixed Infrastructure Cost

Establishes a single fixed self-hosted baseline cost regardless of token type.

Red Hat

Tokenomics Takeaways

Goodput Metric

Goodput (throughput under a latency constraint) is crucial to use for calculations.

Theoretical Cost Floor

Cost per million tokens at 100% GPU utilization gives you the floor on self-hosted costs.

Demand Modelling

Demand modelling gives you an idea of your actual utilization which drives up self-hosted costs.

OPTIMIZATION STRATEGY A

If Demand > Capacity

Invest in improving goodput to serve more tokens efficiently and maximize throughput under latency constraints.

OPTIMIZATION STRATEGY B

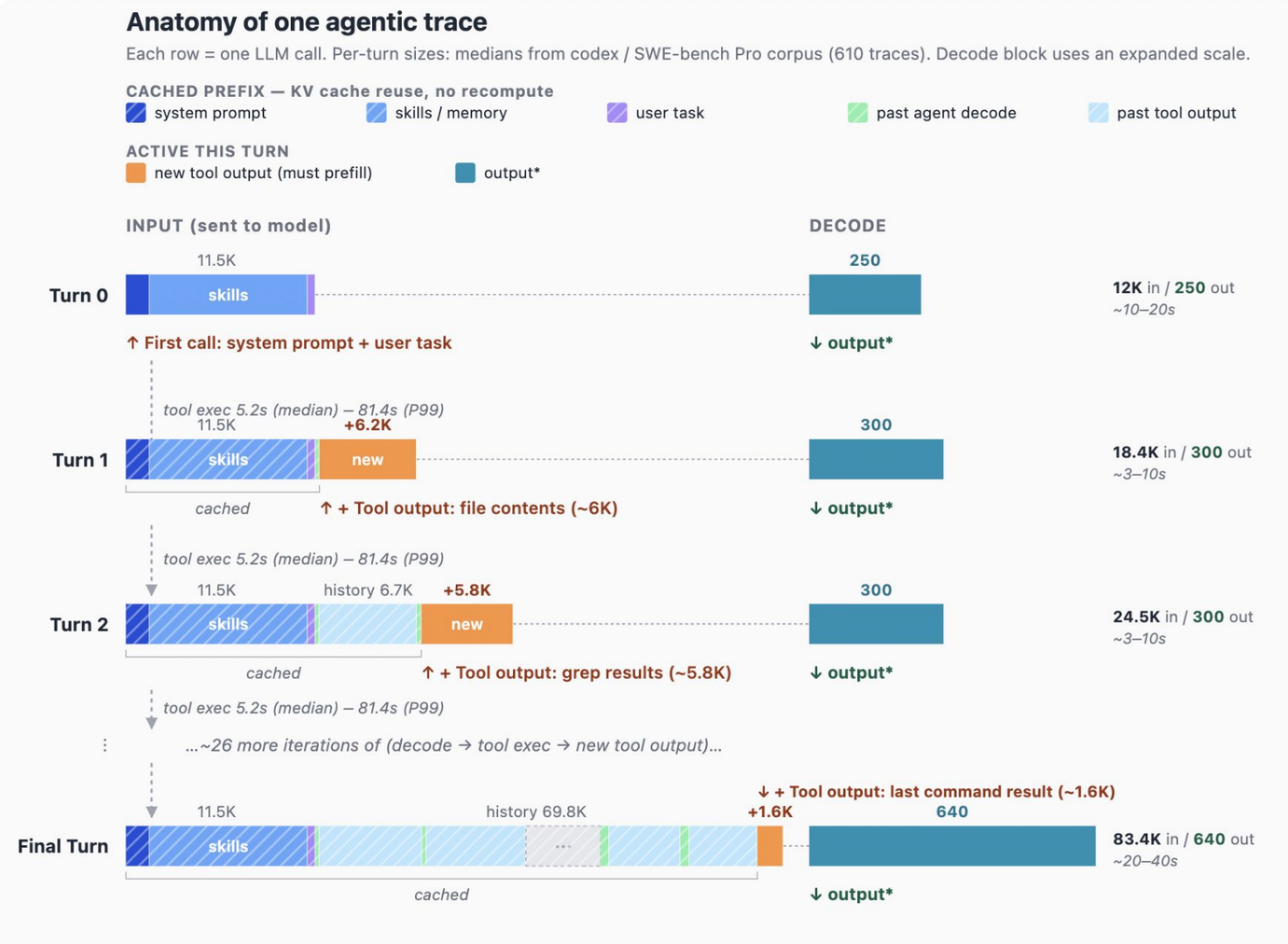
If Demand < Capacity

Invest in consolidation and centralization of your inference services to significantly improve GPU utilization.

Adopting a distributed inference solution such as llm-d or dynamo can help.

Red Hat

Challenges of Serving Agentic Workloads



Source: vLLM x Mooncake's joint blog on Codex/SWE-bench Pro corpus data [here]

Long context, high ISO/OSL

- Expensive prefills, high variance
- Needs a lot of KV Cache Space

Multi-turn/Tool-calling

Need a high cache hit rate

- Prefix-aware routing
- KV offloading

Large MoE Models

- Multinode with a lot of data movement, bursty traffic

Observability

- Agentic traces

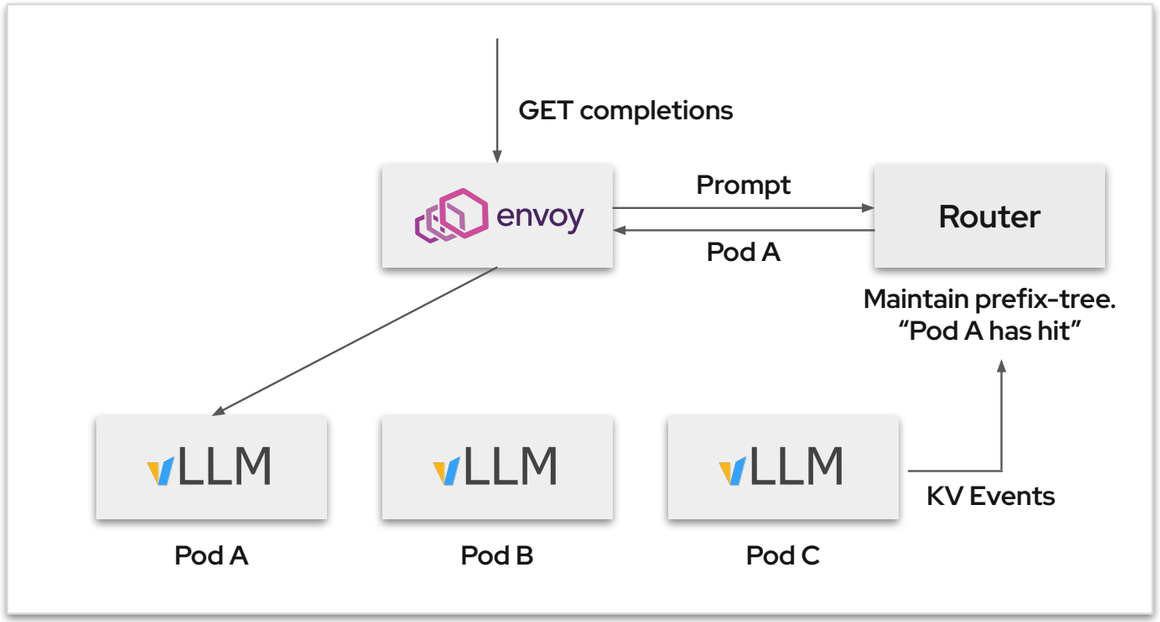
AI Gateway and Guardrails

Red Hat

Cache is King: 10x Cost of Cached vs Uncached Token

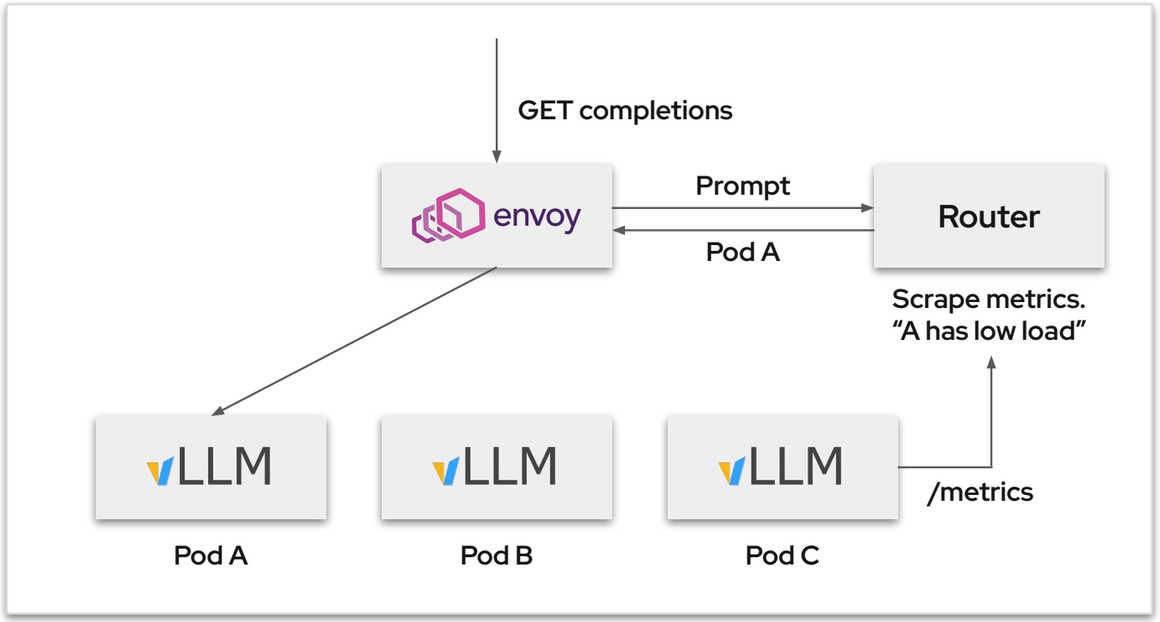
llm-d Intelligent routing: reducing costs and improving user experience

Prefix-Aware Routing



Dramatically increase prefix-cache hit rate

Load-Aware Routing

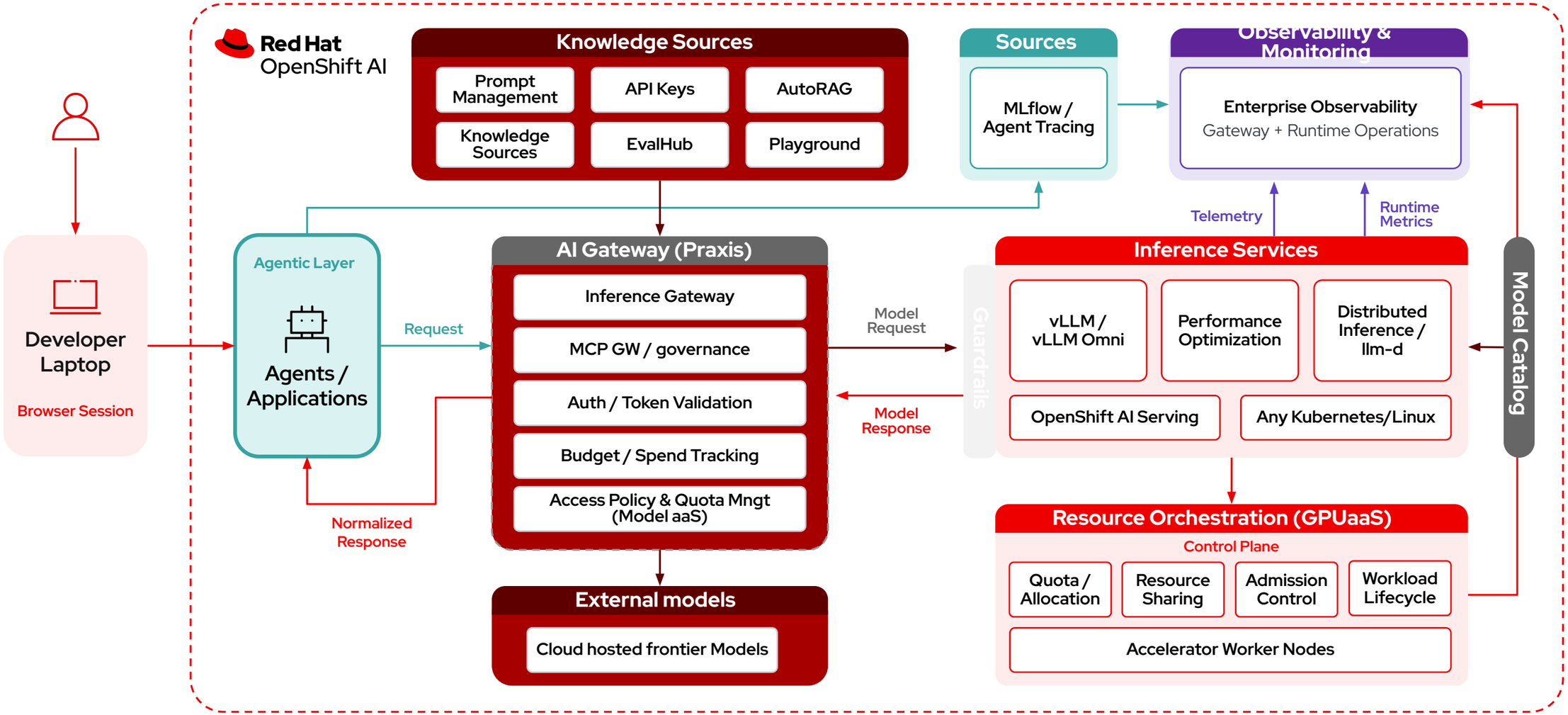


Load-balancing based on actual replica state

The question isn't whether caching matters. It's whether your infrastructure is smart enough to use it.

Red Hat

Enterprise Agentic Inference Platform on Openshift AI



Red Hat

Our Tokenomics Leadership

PUBLIC LEADERBOARD

InferenceX + MLPerf Leadership

Across text, audio, and images

vLLM + llm-d across major accelerators

[Red Hat MLPerf Inference v6.0 results (Apr 2026)]

https://inferencex.semianalysis.com/inference

COMMUNITY

>3×

Throughput gain at matched interactivity

Kimi K3 on B300 with ~4x faster decoding and acceptance holding to ~1M context

https://vllm.ai/blog/2026-09-15-kimi-k3-dspark

ENTERPRISE ADOPTION

60B+

Tokens served weekly

Via Red Hat AI across customers running inference at scale

Internal reporting. Two major US banks running Red Hat's inference stack in production

Red Hat

More sessions from Red Hat

WEDNESDAY, NOVEMBER 4

Operationalizing Security for AI Systems in Financial Services

Nov 4 • 4:30 PM - 4:45 PM | Hub 4

De-Risking Enterprise GenAI: Aligning Software Architecture and Compliance on AI Models Licensing

Nov 4 • 4:50 PM - 5:05 PM | Hub 4

Agents on a Leash: Deterministic Agentic AI for Financial Services

Nov 4 • 5:15 PM - 5:50 PM | Hall 1

THURSDAY, NOVEMBER 5

From Tokens to Tokenomics: Optimizing AI Costs with Open Source

Nov 5 • 2:45 PM - 3:20 PM | Hub 2

The AI Platform Engineer's Playbook: Operationalizing Agentic Workflows at Scale with Gateway

Nov 5 • 4:10 PM - 4:25 PM | Hub 1

Red Hat



Thank you

akamra@redhat.com

yfama@redhat.com

From Memory Budget to Cost Per Token: Serving Agentic Traffic with Large Open Weight Models

linkedin.com/company/red-hat

facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat

Red Hat