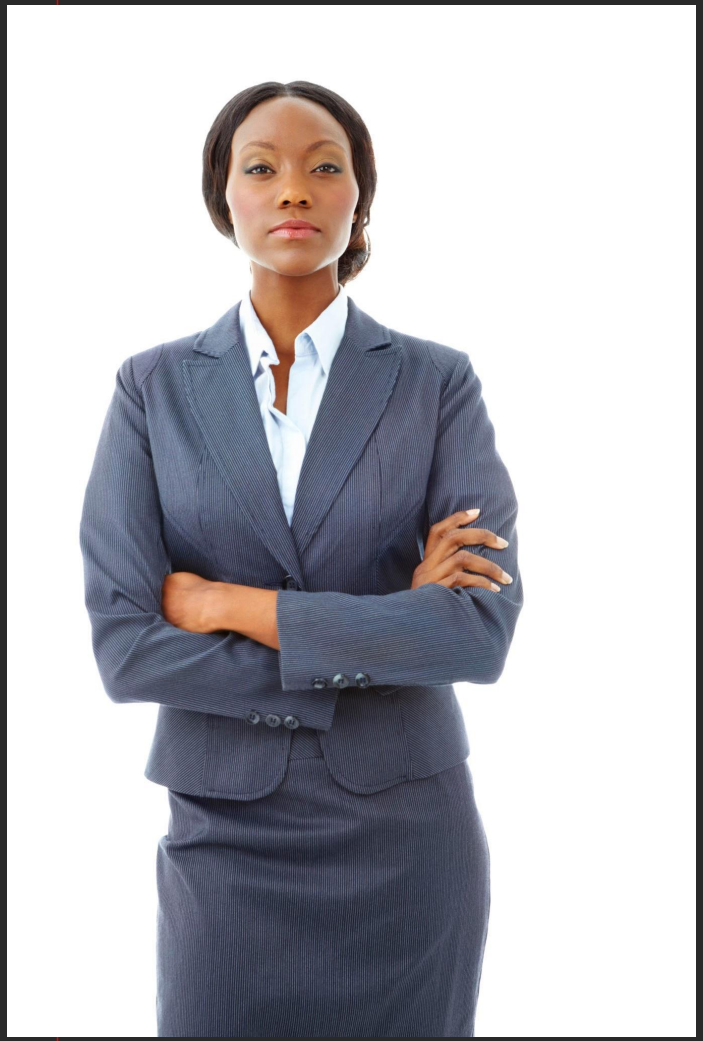
Owning the Inference Layer

When and How to Run your Own Models

Taylor Jordan Smith

Developer Advocate, Red Hat AI

Red Hat



> Why would I need to host my own models when I can just pay a couple bucks for a million tokens on OpenAI?

Jane Smith
CTO, Big Enterprise Inc

“Why open your own business, when Walmart will have cheaper prices? Why buy a car, when you can take a taxi?”

Taylor Smith (Not *that* Taylor)
Chief Cat Officer
Crimson Fedora, Inc.

Yet Another Car Analogy (YACA ™)

Yet Another Car Analogy (YACA): Take a taxi

- Model is hosted by a provider
- Send a request, get an answer
- Get charged based on usage (e.g. per 1M tokens)
- You're using a shared managed service
  - E.g. GPT, Claude, Gemini
- Easy to get started and low overhead, but:
  - Very little control
  - May not meet data security requirements.

Yet Another Car Analogy (YACA): Lease or Rent

Cloud commit (Lease / Rent)

- Dedicated model capacity or endpoints
- Provisioned throughput (predictable latency and scale)
- Infrastructure is managed, but hardware is not fully isolated

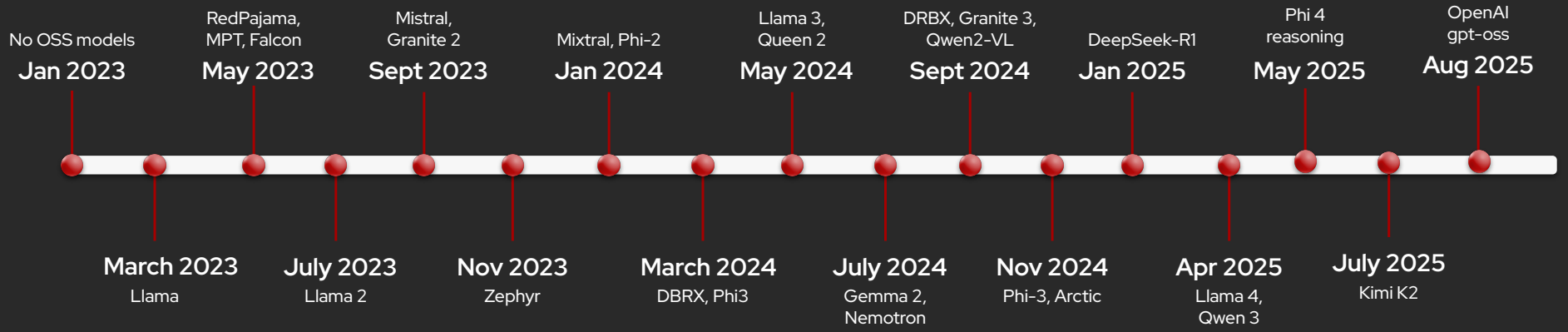
(E.g. Google Vertex AI, Azure AI Foundry, Amazon SageMaker)

- Scales to high throughput with planning
- Better data isolation than shared APIs
- Requires some infrastructure and ML knowledge
- More control over performance and deployment

Yet Another Car Analogy (YACA): Own the car

- Run the model on infrastructure you control
- Cloud, on-prem, air-gapped
- You manage the scaling, serving, and hardware utilization
- Highest engineering and operational investment, but:
  - Strongest data privacy and deployment flexibility
  - Maximum control overall
  - Best optimization potential

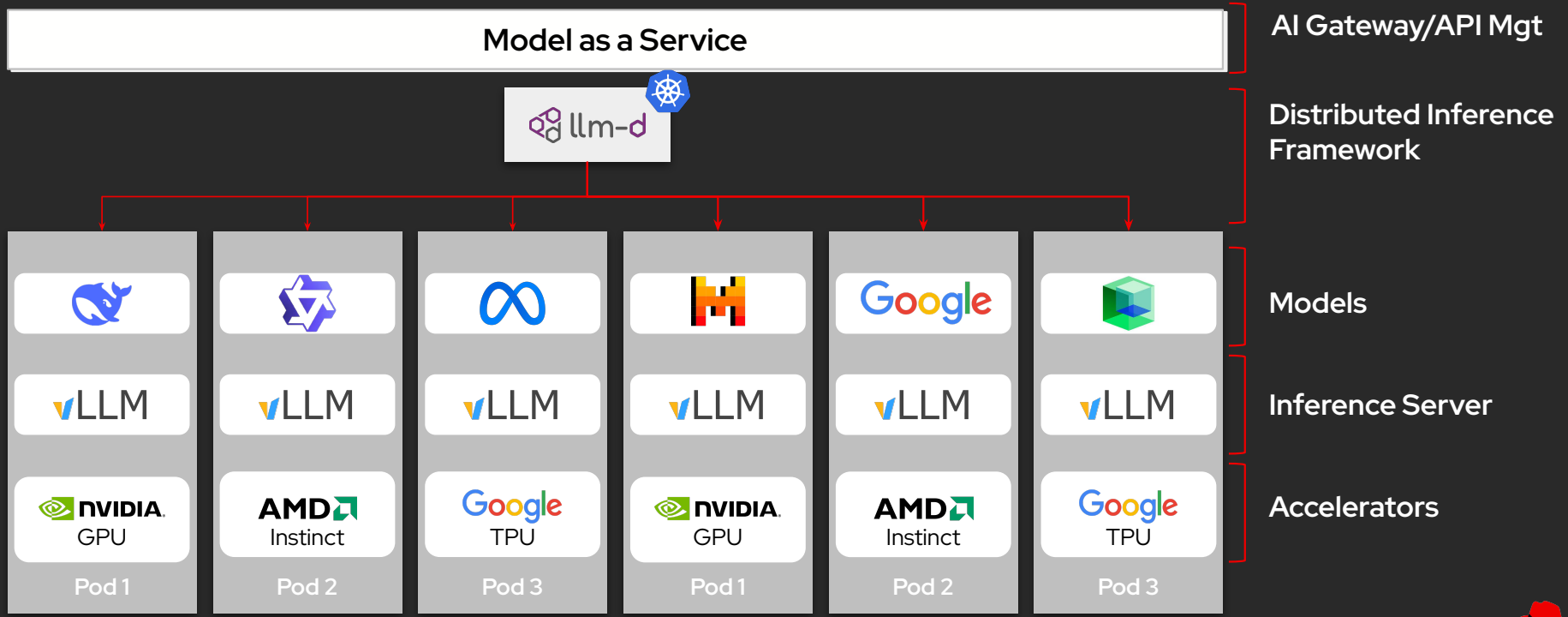
Why this is possible now: model capability exploded



Meta · Ai2 · NVIDIA · MISTRAL AI_ · databricks · Hugging Face · snowflake · IBM · Google · Microsoft · deepseek

Red Hat

And the infrastructure caught up



People don’t host models for just one reason...

You need to customize the models for higher accuracy

You can’t send your data to an external API

Your token bill has stopped making sense

You’re not hitting your latency targets

How to decide what to do next

Stay with APIs (taxi) if:

- You're experimenting
- Traffic is low or unpredictable
- Speed > optimization

Move to lease/rent if:

- You need reliability (latency, throughput)
- You're scaling but not fully committed
- You want isolation without full ownership

Own inference if:

- You have sustained high throughput
- Cost per token is becoming a problem
- Latency or UX is critical
- You need deep control (privacy, tuning)

This isn’t about **hosted** vs **self-hosted**.

It’s about understanding when your architecture needs to evolve.

Owning inference isn’t step one. It’s step ***next***.

Thank you

Taylor Jordan Smith on Linkedin
Developer Advocate @Red Hat

Red Hat AI
Developers Newsletter

Red Hat