



KubeCon



CloudNativeCon

Europe 2026

#KubeCon #CloudNativeCon

Intelligent Routing for optimized inference

Antonio Berbén (Solo.io) Felipe Vicens (Telefónica)



The Hidden Waste in AI Inference



KubeCon



CloudNativeCon

Europe 2026



AI Growth Is Driving Uncontrolled Costs

- AI adoption is exploding
- Inference cost scales with usage
- Many teams lose cost control

Static Model Selection Doesn't Work

- Most systems use a single default model
- Routing decisions are not context-aware
- Simple prompts waste expensive compute

The Hidden Waste in AI Inference

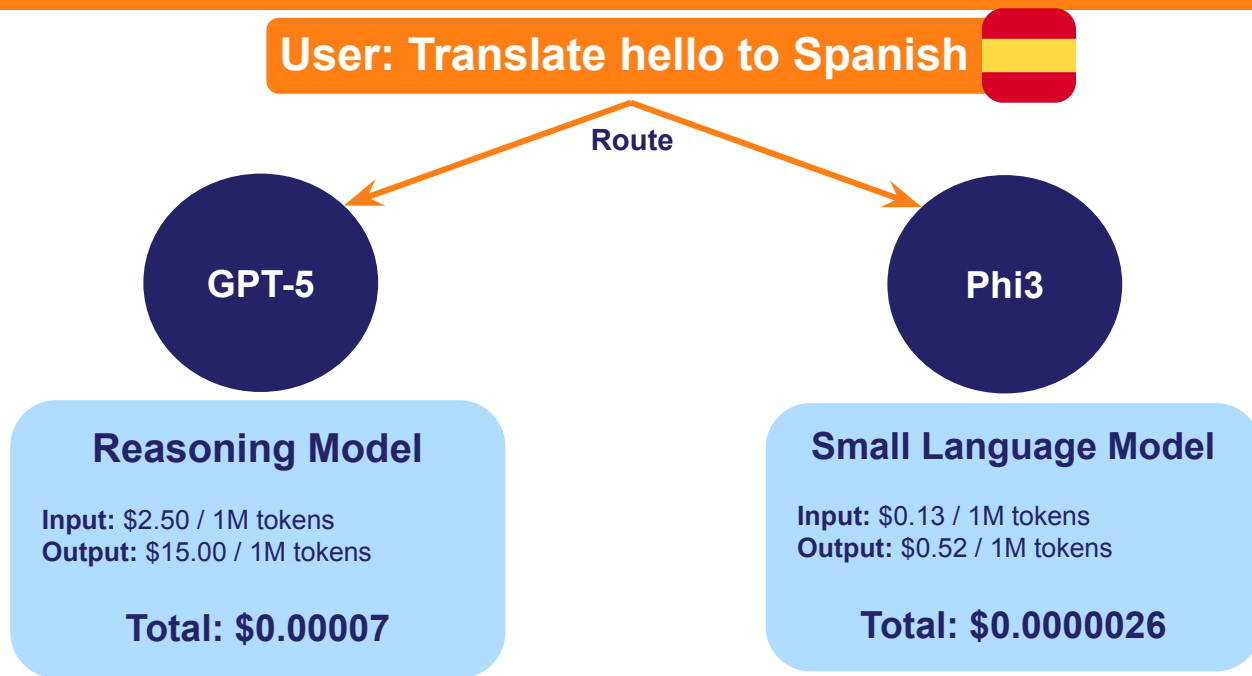


KubeCon



CloudNativeCon

Europe 2026



Cost difference: up to 27x

Owning the AI Value Chain



KubeCon



CloudNativeCon

Europe 2026



Telcos operate highly distributed infrastructure across edge, core, and cloud environments.



Regulation and data sovereignty requirements limit where AI workloads can run.



Token-based pricing creates a mismatch between revenue growth and inference cost.

Opportunity: *Dynamic routing across models and infrastructure*

Take what you know... a Gateway

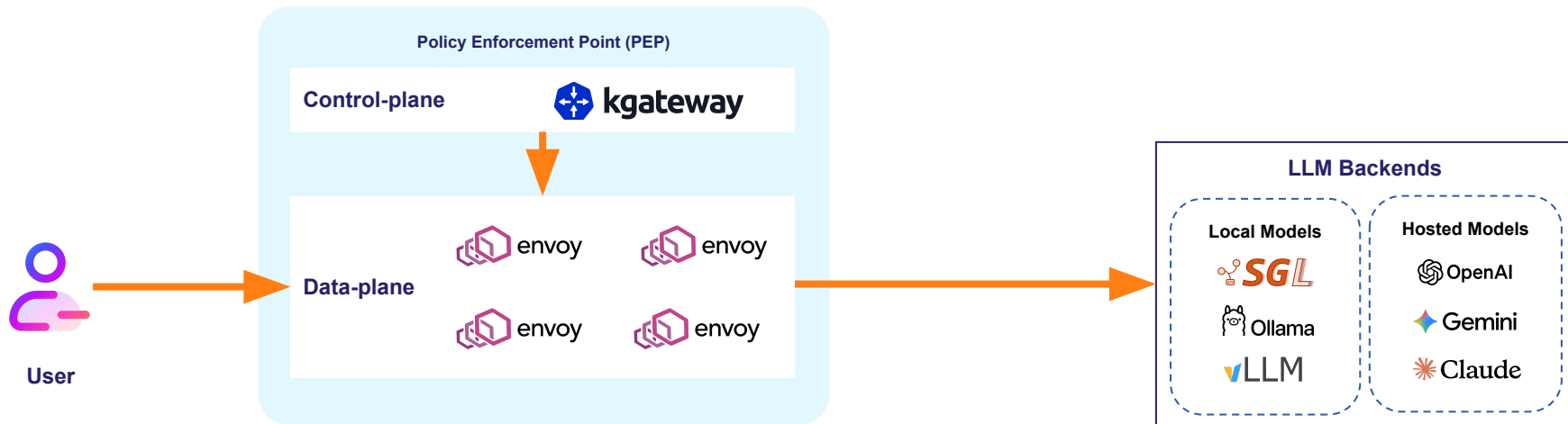


KubeCon



CloudNativeCon

Europe 2026



Envoy

- Open Source (CNCF)
- Best API Gateway
- C++

... make it better ...

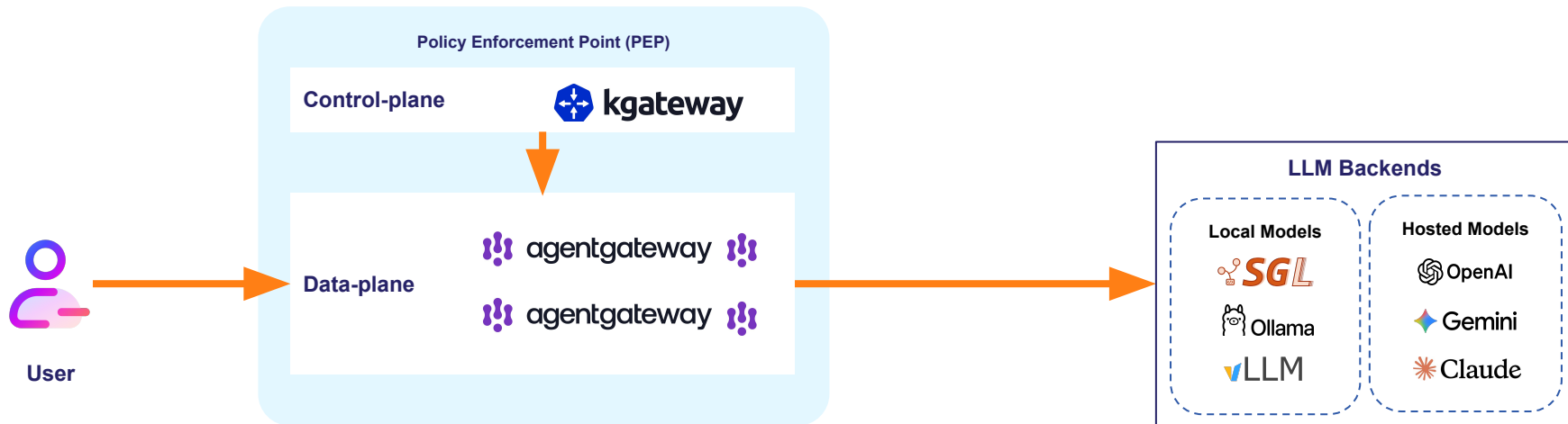


KubeCon



CloudNativeCon

Europe 2026



AgentGateway

- AgentGateway
- Open Source (Linux Foundation)
- Best Agentic Gateway
- Rust (extremely performant)

... extensible ...

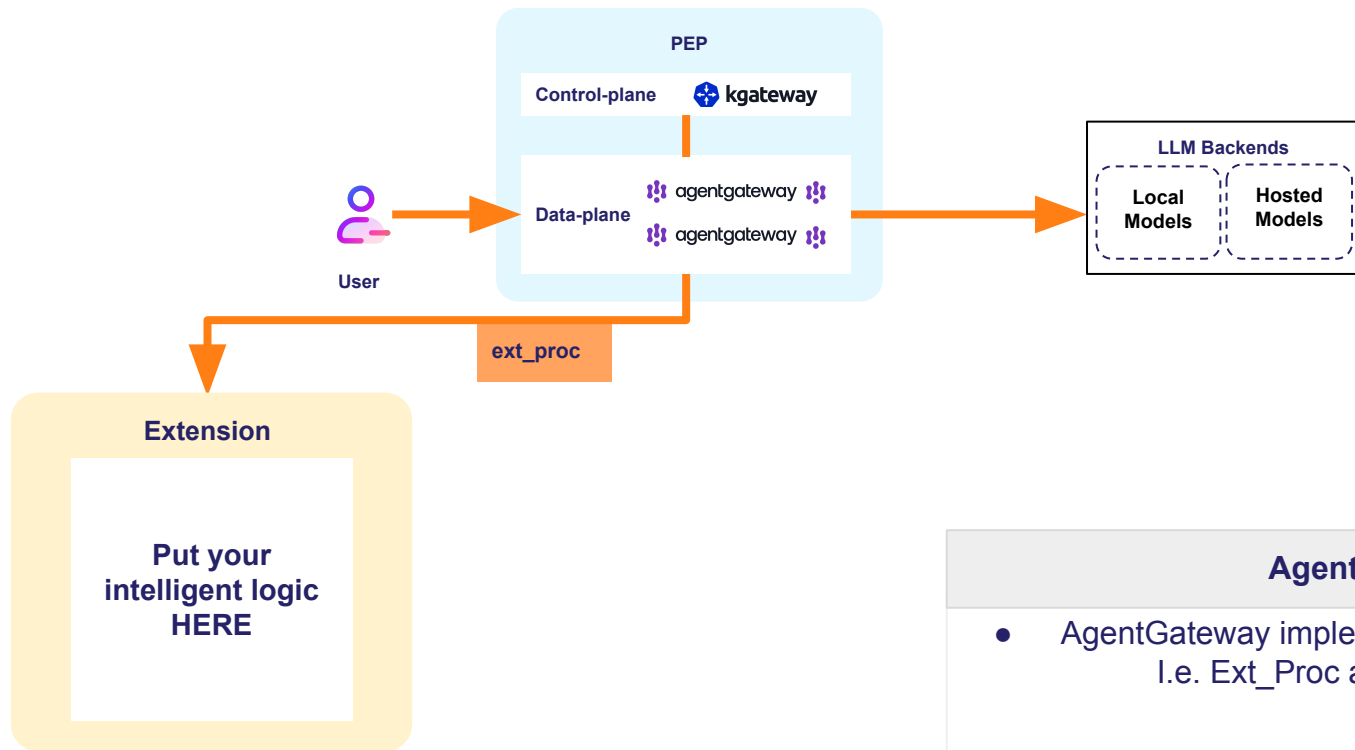


KubeCon



CloudNativeCon

Europe 2026



AgentGateway

- AgentGateway implements Envoy's API
I.e. `Ext_Proc` and `Ext_Authz`

And... Agentic!

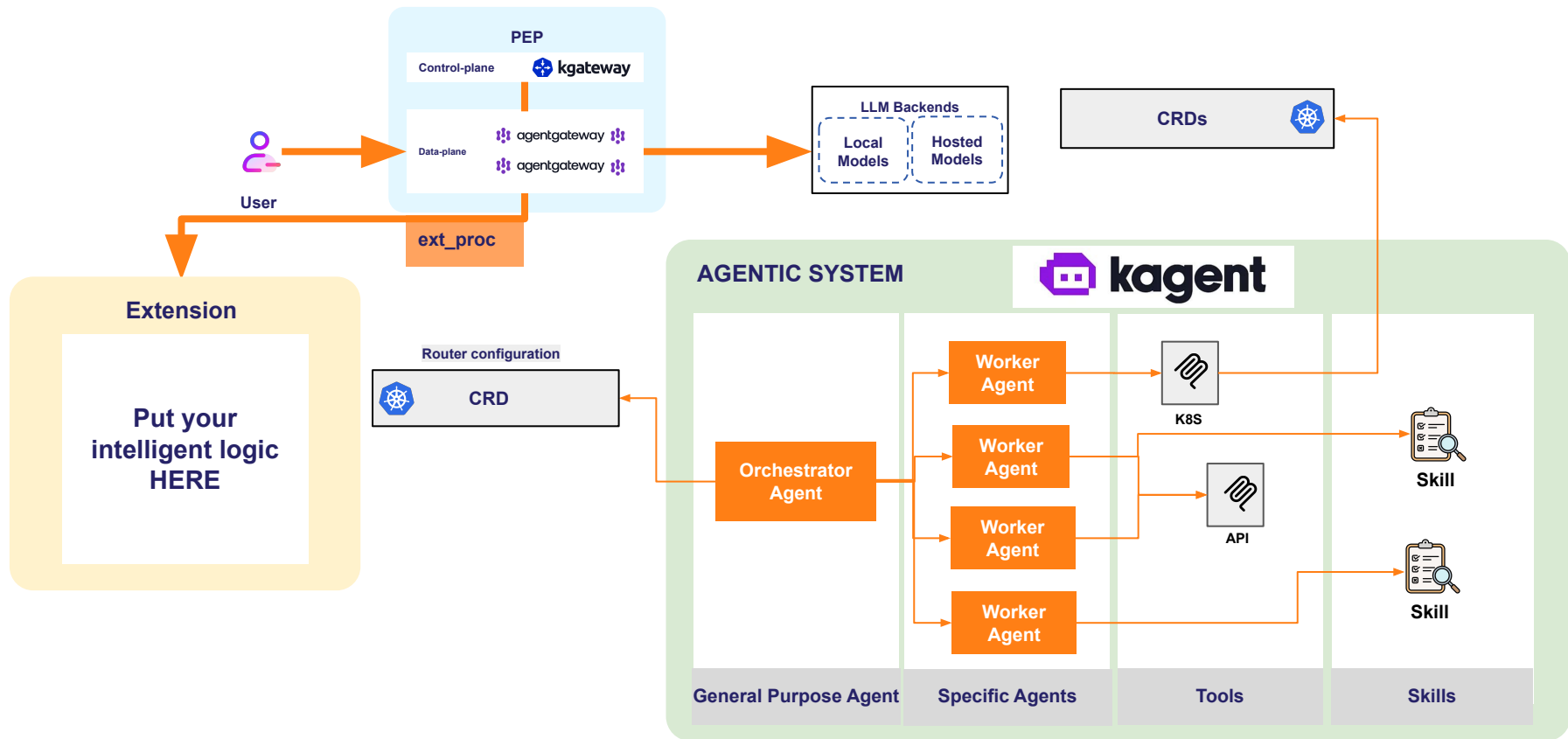


KubeCon



CloudNativeCon

Europe 2026



But, How?

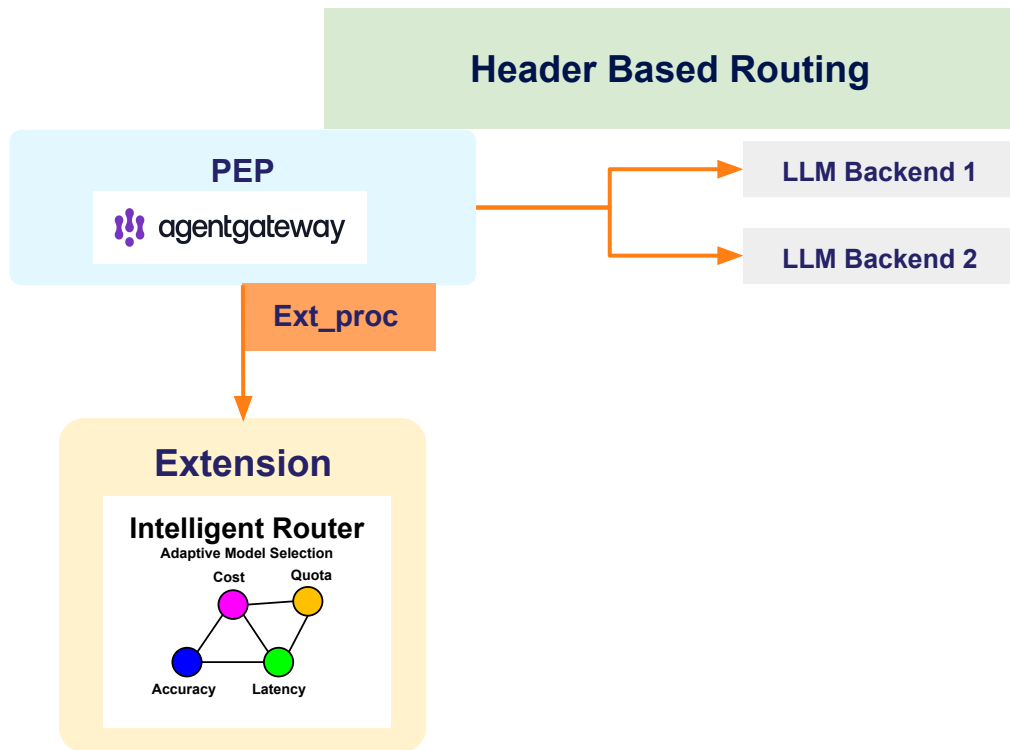


KubeCon



CloudNativeCon

Europe 2026



Intelligent Router:



KubeCon



CloudNativeCon

Europe 2026

LLM-D

Strengths:

- Production Ready Architecture
 - Inference Gateway
 - Body-based Routing
- Backend selection optimized by load and KV cache

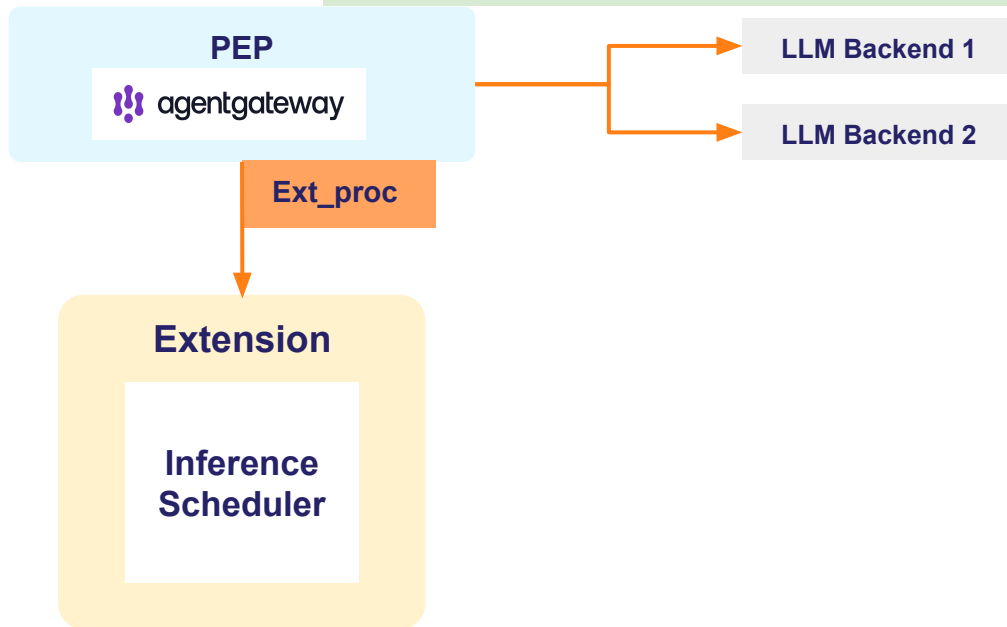
Considerations:

- Prioritizes hardware-level optimization, without support for offloading to public endpoints
- Limited flexibility for distributed SLMs on CPU

Header Based Routing:

x-gateway-destination-endpoint = IP:port (LLM Backend 1)

x-gateway-destination-endpoint = IP:port (LLM Backend 2)



Intelligent Router:



Open Policy Agent



KubeCon



CloudNativeCon

Europe 2026

Open Policy Agent

Strengths:

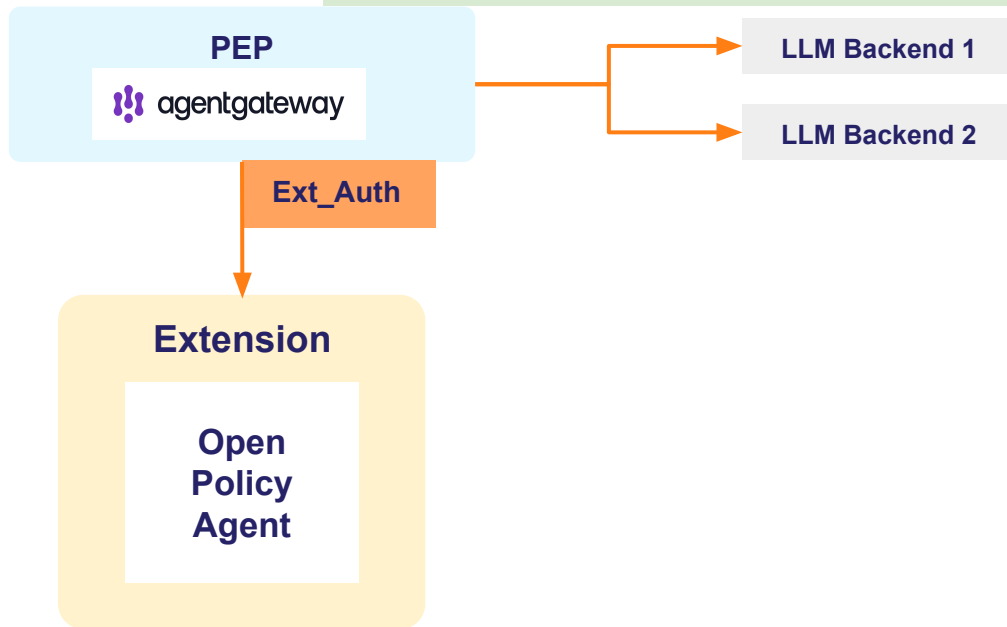
- Easy to implement
- Fast decision engine
- Supports public and local endpoints

Considerations:

- No support for modifying request body parameters
- Limited ability to incorporate real-time signals (e.g., latency, cost, energy)

Header Based Routing:

x-model = LLM Backend 1
x-model = LLM Backend 2



Intelligent Router: vLLM Semantic Router



KubeCon



CloudNativeCon

Europe 2026

vLLM Semantic Router

Strengths:

- Production-ready model router
- Signal-driven, extensible design
- Supports public & private endpoints

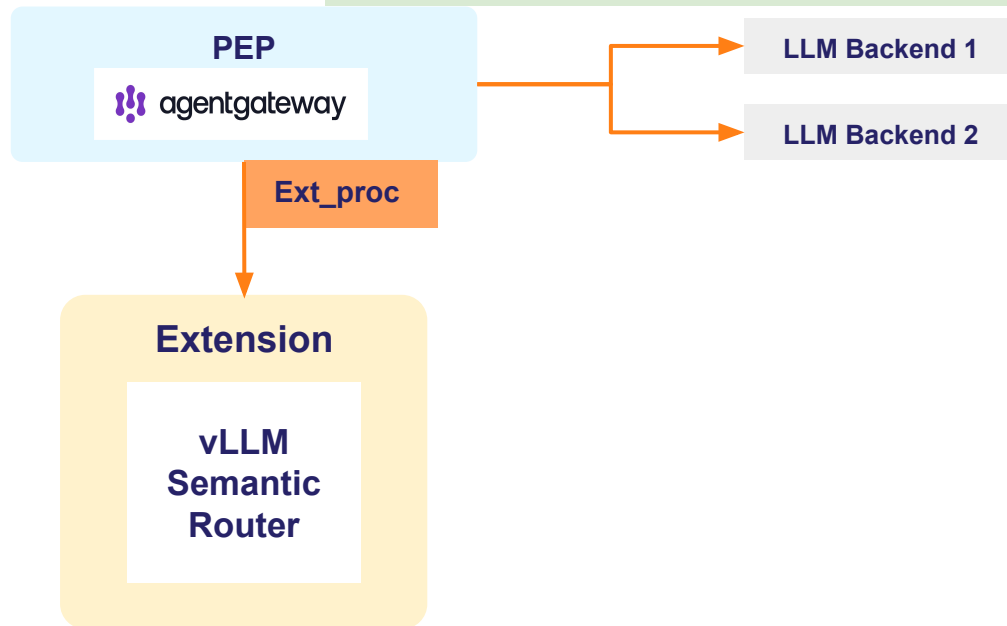
Considerations:

- CRD configuration does not yet expose all available options
- Possible model/config misalignment
- Added complexity at scale

Header Based Routing:

x-selected-model = LLM Backend 1

x-selected-model = LLM Backend 2



Intelligent Router: Bring Your Own

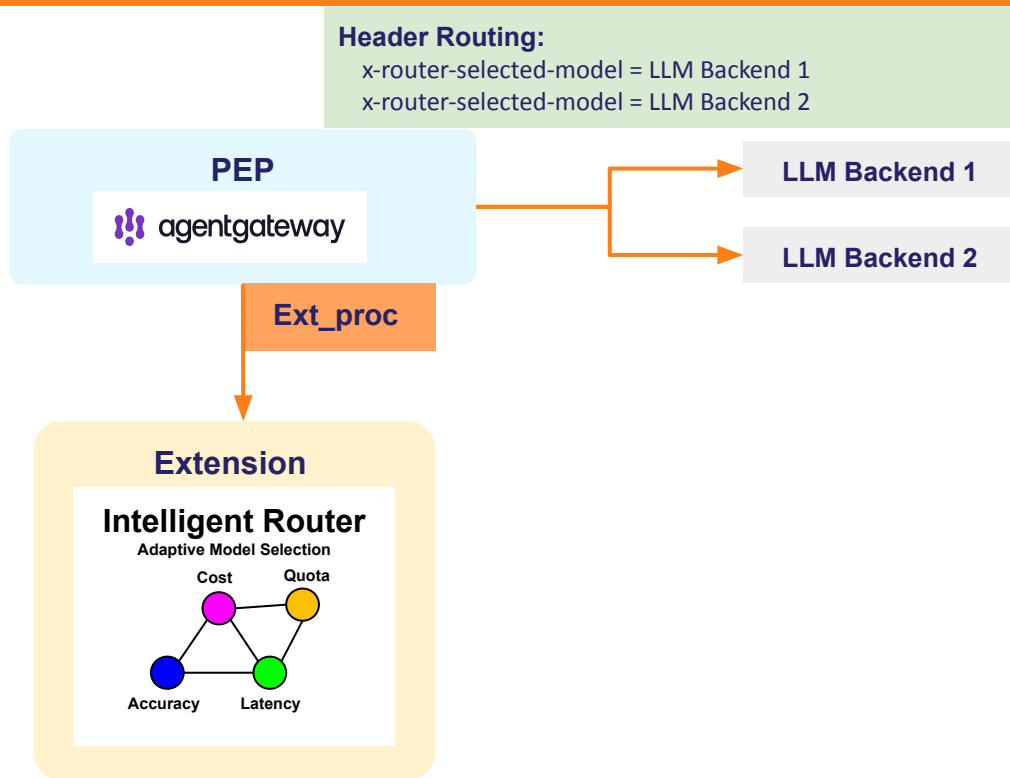


KubeCon



CloudNativeCon

Europe 2026



Agentic Control Plane for Intelligent LLM Routing

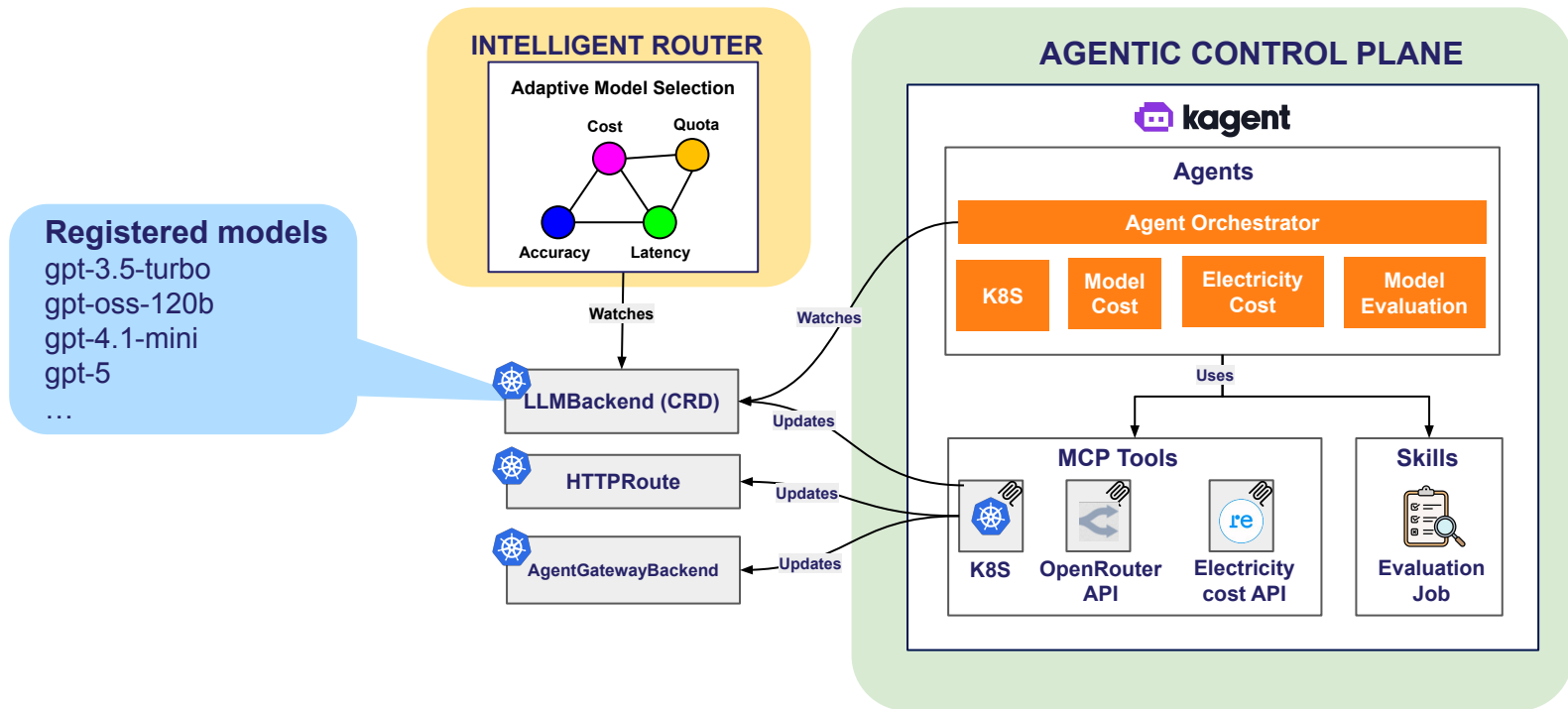


KubeCon



CloudNativeCon

Europe 2026



Agentic Control Plane for Intelligent LLM Routing



KubeCon

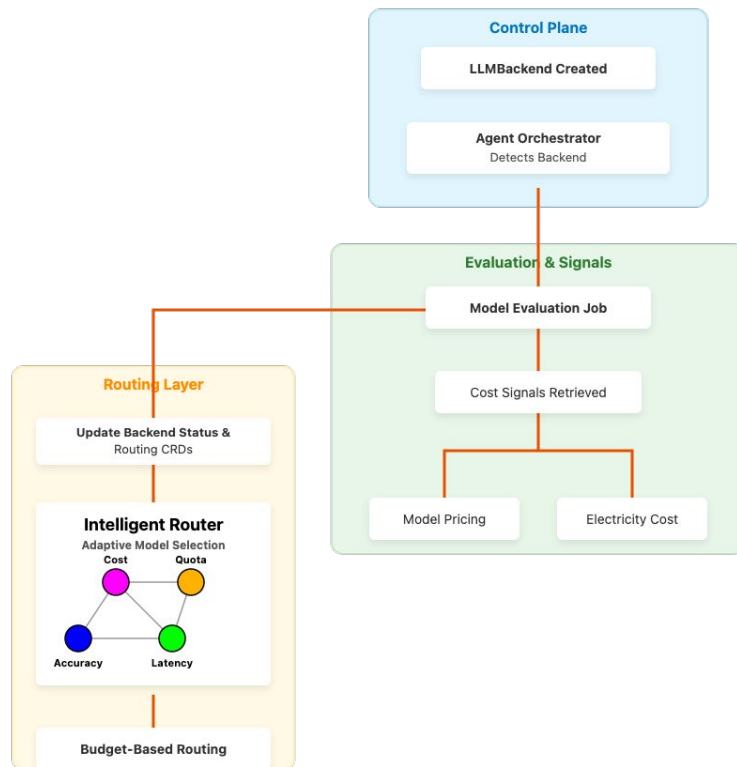


CloudNativeCon

Europe 2026

LLM registration flow

1. A new LLMBackend resource is created
2. Agent Orchestrator detects the backend
3. Model Evaluation is triggered
4. Cost signals are retrieved
(model pricing and electricity cost)
5. LLMBackend status and routing resources are updated
6. Intelligent Router includes the backend in routing



Missing features



KubeCon



CloudNativeCon

Europe 2026



Telemetry Signals

Live metrics are required to autonomously take routing decisions.

If we want to autonomously take routing decisions based on metrics, those should be as live as possible



Geo-Location

Models might not be deployed in the same locations across regions.

Models might not be deployed in the same locations



Zero Trust

Data protection easily addressed with Istio (Ambient mode).

Data Protection / Sovereignty
These can easily be addressed with Istio (Ambient mode)



Other use cases to consider



**Geo-Location
Data
Residency
Constraints**



**Thermal and
Power-Aware
Scheduling**



**Energy per
Token
Ecological
Optimization**



**PII Vaulting
and Format
Preservation**



**Exponential
Backoff
Resilience
Routing**

... And 50 more here in github

[Github -> Hybrid LLM Routing Strategies](#)



Lessons Learned



KubeCon



CloudNativeCon

Europe 2026

Agentic AI for everything

- Use it to configure your systems in an agentic way to make them autonomous

Skills are extremely powerful

- Agent Workflows feels old with Skills to control complex flows

Cost Based routers is a real thing

- We need to think how to reduce cost

Best model is context-dependent, not universal

- Context defines which model to use

Cost visibility must come before cost optimization

- You can't optimize what you can't measure



Demo

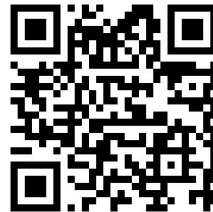


KubeCon



CloudNativeCon

Europe 2026



Youtube Video

[Link](#)



Github Repo

[Link](#)



KubeCon



CloudNativeCon

Europe 2026

THANK YOU!

Let's keep building together, connect with us on LinkedIn



Antonio Berben

<https://es.linkedin.com/in/antonio-berben>



Felipe Vicens

<https://www.linkedin.com/in/fjvicens/>



Q&A



The **future** present is Agentic



KubeCon



CloudNativeCon

Europe 2026

TM Forum Moonshot challenge:

Zero-trust agents: Vendor-neutral agent runtime for autonomous networks

<https://www.tmforum.org/catalysis/projects/C26.0.933/zero-trust-agents-vendor-neutral-agent-runtime-for-autonomous-networks>



Industry Missions

Open Digital Architecture (ODA)

Member projects

Learn

News and insight

Events

About



Home > Projects > Zero-trust agents: Vendor-neutral agent runtime for autonomous networks

[All projects](#)

Zero-trust agents: Vendor-neutral agent runtime for autonomous networks

URN C26.0.933

Topics AI (Artificial Intelligence), Autonomous networks, Governance

Vendor neutral zero-trust agent runtime for Autonomous Networks, with cryptographically signed identity chains and end-to-end traceability.

Following



Visit the project