THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
INDIA

"Hey AI, Train Llama"

Making Kubeflow Agent-Native

Abhijeet Dhumal
Software Engineer, Red Hat
Kubeflow Maintainer and Contributor
CNCF, Kubeflow, Ray, Feast Feature Store

Akash Jaiswal
Software Engineer, Oracle
2× GSoC (Kubeflow, CCExtractor)
Speaker at 5+ events

It's 2026. AI is everywhere.

AI agent tooling has become production-ready, but standard environments still block direct agent control.

U

"Hey AI, build me the next Facebook, so I can be a billionaire."

A

"I've generated the React frontend, configured user authentication, set up the Postgres database, and deployed the stack. That will be $0.08 in tokens. In 24 days, you'll be a billionaire."

1. Coding & IDE Assistants

Developers writing, refactoring, and debugging code in natural language. **Example:** Asking Cursor/Claude to write unit tests, refactor legacy code, or explain an unfamiliar library.

2. AI Workflows

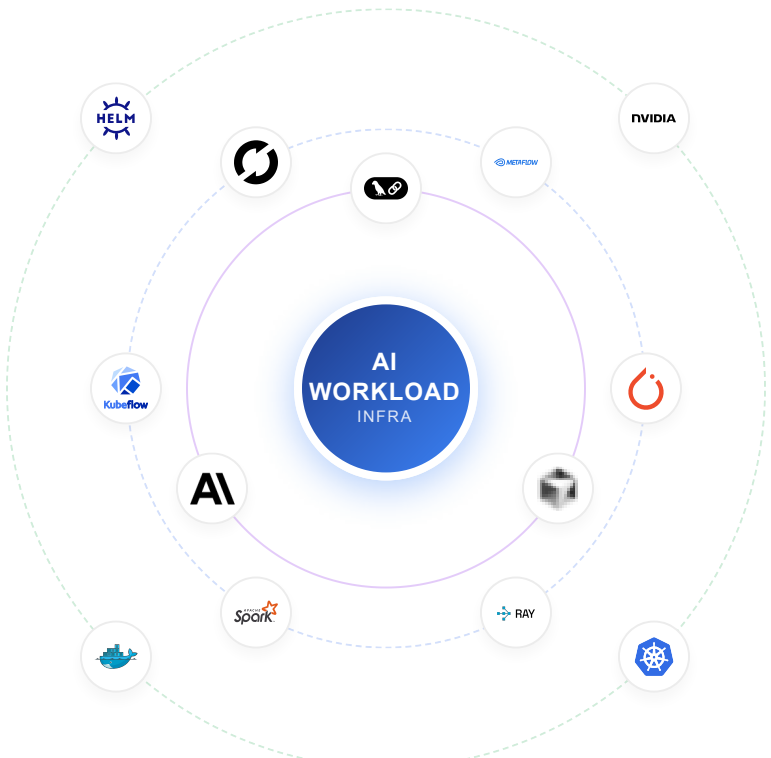
Autonomous agents orchestrating multi-step development loops. **Example:** An agent pulling a Jira ticket, analyzing the repo, making a fix, running local tests, and opening a PR.

3. MCP for Cloud

Standardizing secure infrastructure and API access for models. **Example:** Asking an agent to restart a failing Kubernetes pod, check database schemas, or query cluster resource usage.

AI Workloads on Infrastructure Today

Modern AI workloads span multiple complex layers, demanding specialized platform expertise to bridge high-level model code with raw compute.



PLATFORM REALITY

Fragmented Control Plane
Inconsistent APIs and SDKs across the toolchain

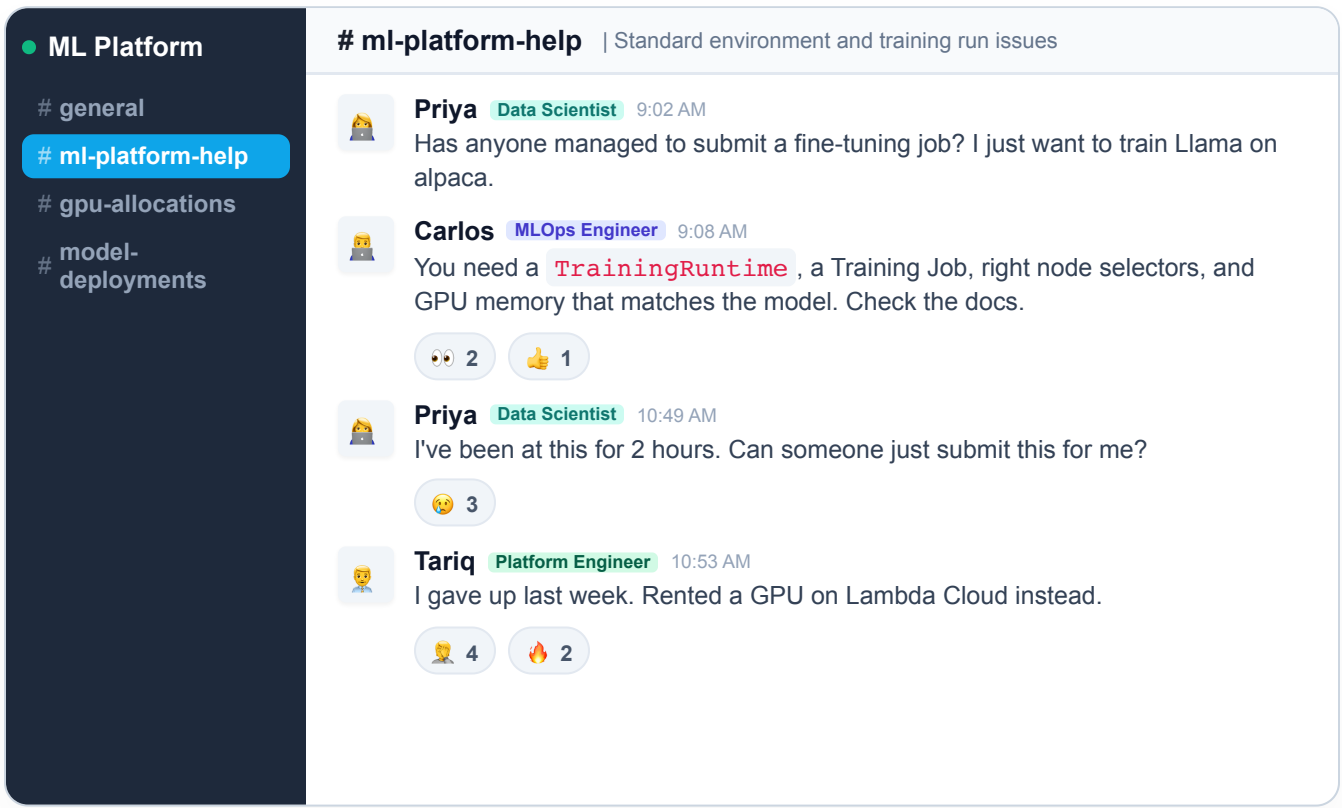
DevOps Bottlenecks
High human-coordination cost for GPU allocations

Disconnected Context
Infra has zero awareness of agentic intentions

AI Agents ML Platform Infra & Silicon

Every ML team has this conversation. Every week.

Real cost: hours of human coordination for a 5-minute job.



THE REAL PROBLEM

No Discoverability
Priya is blocked because she doesn't know which TrainingRuntime which matches Llama.

No Validation
Carlos tells her to read the docs, but fails to prevent OOM at epoch 1 after 2 hours of setup.

No Self-Service Guidance
Priya is begging others to write and submit YAMLs for her since the SDK isn't interactive.

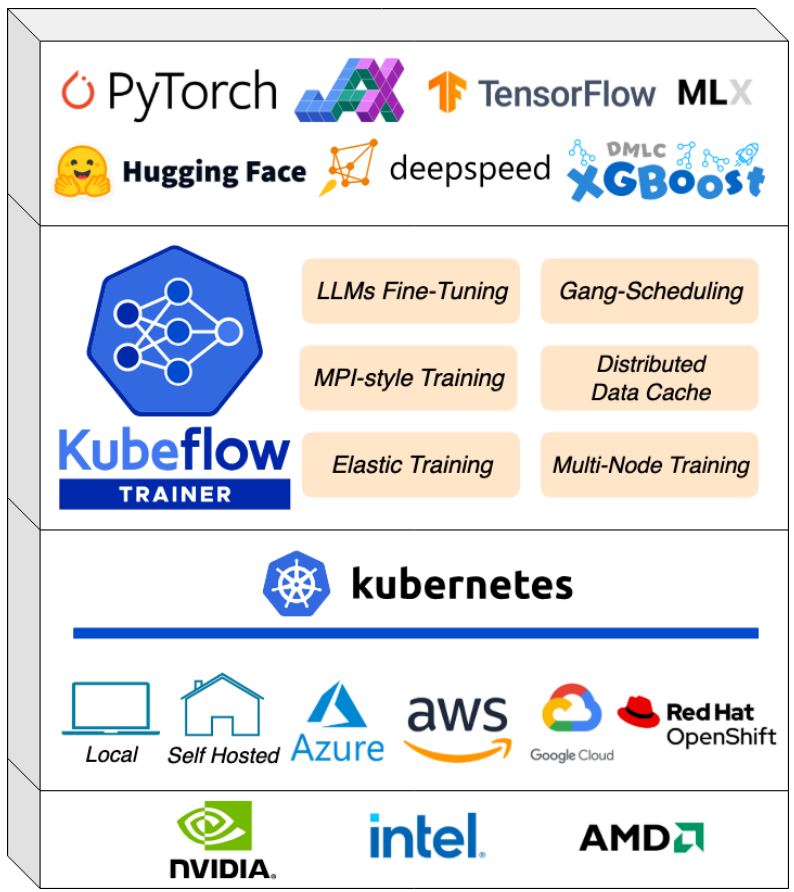
Shadow IT & Bypass
Platform Engineer Tariq bypassed the platform entirely to rent external GPUs on Lambda Cloud.

Kubeflow Ecosystem - Working with Kubeflow Trainer

Kubernetes-native distributed training

- **Unified API:** `TrainJob`
- **Supported Frameworks:** PyTorch, JAX, DeepSpeed, MPI, MLX
- **Orchestration:** Handles pod orchestration, networking, and gang-scheduling
- **Bridge:** Bridges high-level ML toolkits ⟺ raw Kubernetes infrastructure

> KFT makes it easy to train state-of-the-art ML at scale.



Kubeflow Ecosystem - Working with Kubeflow SDK

A single, unified Python SDK to build, train, and deploy ML workflows locally or on Kubernetes.

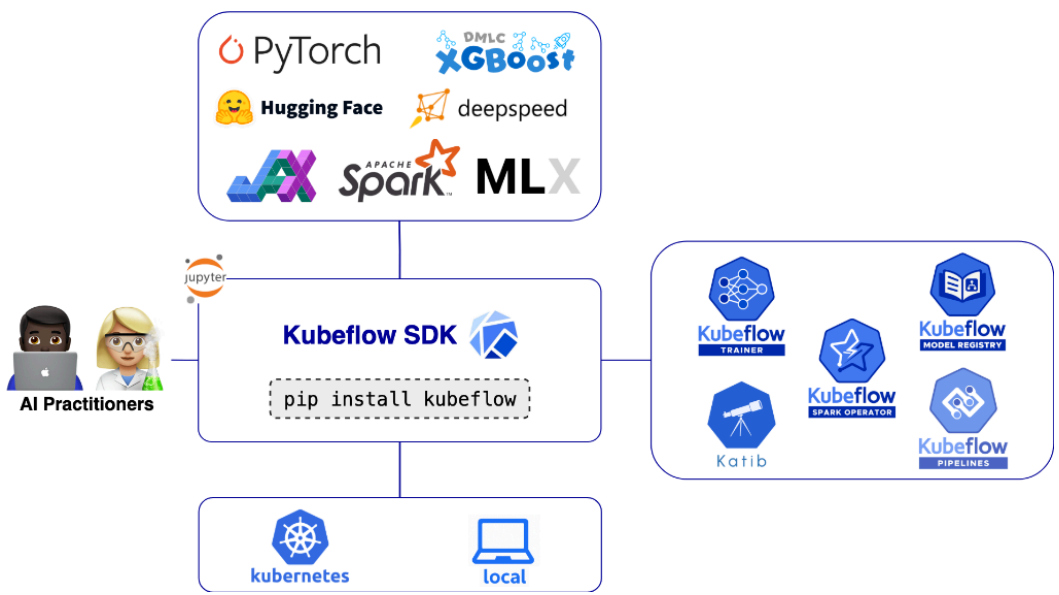
Unified Pythonic API
One consistent programming model for writing local experiments or production deployments.

Infra Abstraction
Translates Python functions automatically into Kubernetes TrainJob resources without YAML.

Flexible Backends
Seamlessly switch execution from Local processes, isolated Containers, to Kubernetes.

Ecosystem Integration
One interface supporting Trainer (models), Katib (tuning), Registry, and Pipelines.

Demo Video
Streamlining ML Workflows With the Unified Kubeflow SDK



What if you could just...

talk to your cluster?

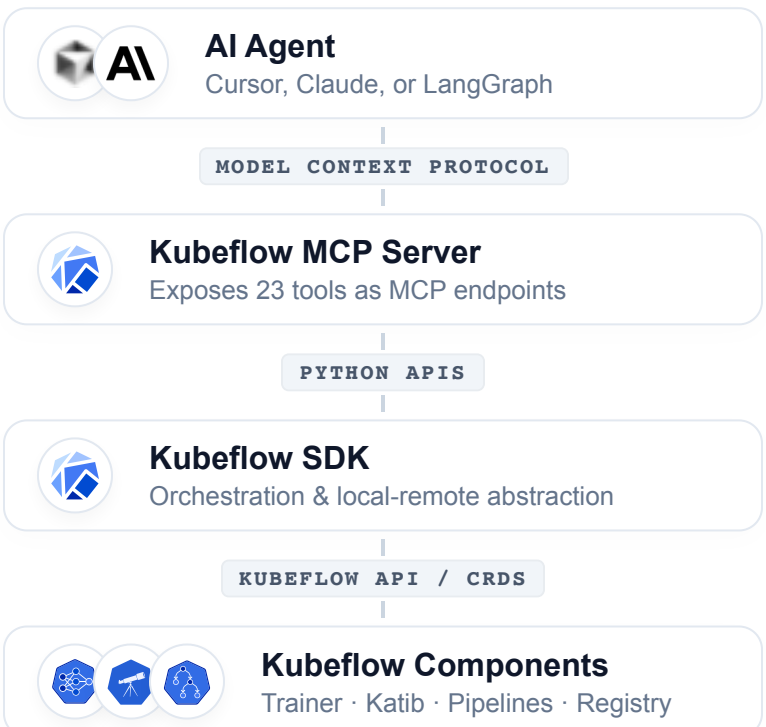
One sentence in natural language. Your model starts training. No YAML. No SDK boilerplate. No Kubernetes expertise required.

Inside Cursor, to an AI agent

```
"Hey AI, train Llama 3.2 on the alpaca dataset, 3 epochs"
```

Introducing – Kubeflow MCP

Connecting AI Agents directly to Kubeflow Platform.



AGENT TOOL CALLS 23 tools

PERSONA: `readonly` → `data-scientist` → `ml-engineer` → `platform-admin`

READ `list_training_jobs()`
Returns all active and completed TrainJobs in the namespace.
→ `TrainJob[]`

WRITE `fine_tune(model, dataset, gpus, …)`
Configures the ClusterTrainingRuntime and starts model training.
→ `{ job_name, status }`

READ `get_training_logs(name)`
Streams stdout/stderr of a running pod back to the agent in real-time.
→ `LogStream`

READ `pre_flight(model)`
Estimates GPU memory + validates cluster quotas before launch — catches OOM before a pod is scheduled.
→ `ValidationResult`

🔐 **Every mutating tool previews first (`confirmed=False`) — the agent shows exactly what it will do and waits for explicit approval before touching your cluster.**

Three demos

LiteLLM · Claude · Cursor IDE — same MCP server, different clients

DEMO 1 LANGCHAIN + LITELLM

LangChain + LiteLLM

- `pre_flight()` catches GPU memory — no more OOM at epoch 1
- `confirmed=False` preview → explicit approval before submit
- Progressive mode — 3 meta-tools for small local models

DEMO 2 CLAUDE

Custom training — Claude

- `run_custom_training()` with a PyTorch script
- Submitted to Kubeflow via MCP API
- Any LLM client on any machine

DEMO 3 CURSOR

Built-in trainer — Cursor IDE

- `fine_tune()` with torchtune runtime
- Persona filtering live in the IDE
- GPU scheduling, confirm gate, logs

All three demos use the same Kubeflow cluster — Demo 1 via stdio locally, Demo 2 via HTTP, Demo 3 via Cursor IDE. Same 23 tools, different client and inference backend.

Kubeflow MCP – Roadmap

THE LINUX FOUNDATION
OPEN SOURCE SUMMIT INDIA

From basic TrainJob control to fully autonomous ML Ops on Kubernetes.

PHASE 1: CORE ALPHA | PHASE 2: HARDENING | PHASE 3: ENTERPRISE | PHASE 4: ADVANCED

PHASE 2 · NOW

Production Hardening

- HTTP Authentication & Identity Wiring
- **Anthropic Direct Adapter** — native streaming without LiteLLM proxy
- **LiteLLM Provider** — unified model routing for local & cloud LLMs
- Dockerfile Optimizations & PyPI Release

PHASE 3 · NEXT

Enterprise & In-Cluster

- **Helm Chart** support for cluster-wide deployments
- **OAuth / OIDC Integration** for secure multi-tenancy
- **Multi-MCP Coordination** across platform nodes

PHASE 4 · FUTURE

Advanced Training & Ops

- Checkpoint management & GPU visibility in client IDEs
- Real-time metrics tracking and autoscaling
- **Self-healing:** Auto-resubmit on GPU OOM
- **Progress Tracking** for distributed training jobs.

Standardizing how agents command, optimize, and serve ML on Kubernetes.

Get Involved

Shape the future of developer experience and AI agents in the Kubeflow community.

Kubeflow SDK and ML Experience WG

- Join **the CNCF Slack**
  - `#kubeflow-ml-experience` channel
- Participate in **Kubeflow MCP development**
- Join Kubeflow SDK & ML Experience WG **bi-weekly meetings**:
  - Wednesdays at 8:30PM IST

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT INDIA