white duck

Kubernetes Community Days
MUNICH 2024

## OUR SPONSORS

metalstack cloud
cisco
consol Enterprising IT
APE FACTORY
paloalto NETWORKS
veeam
dynatrace
Pulumi
StackState
camp to camp INNOVATIVE SOLUTIONS BY OPEN SOURCE EXPERTS
splunk> a cisco company
mindcurv Part of Accenture Song
spectro cloud
Akamai
ISOVALENT now part of cisco
EXOSCALE
TIGERA
STEADFORCE
ADN
CSP² powered by T
DGi
MAIBORNWOLFF
REPLY LIQUID
white duck The Azure Cloud Experts
QA|WARE SOFTWARE ENGINEERING

Speaker: Martin Brandl, Philip Welz

Company: white duck

# Build and run your own intelligent application based on open source using Semantic Kernel and Kaito

# Who we are

**Philip Welz**
Senior Platform & Kubernetes Engineer, Azure MVP

☎ +49 8031 230159-0

✉ philip.welz@whiteduck.de

@philip_welz

in www.linkedin.com/in/philip-welz

**Martin Brandl**
CTO, Cloud Solution Architect, Azure MVP

☎ +49 8031 230159-0

✉ martin.brandl@whiteduck.de

@martin_jib

in www.linkedin.com/in/mbrandl

# Agenda

- Intelligent applications
- Kaito
- Semantic Kernel
- Demo

# Intelligent applications

Kubernetes Community Days Munich 2024

# Intelligent applications

- ....leverage artificial intelligence (AI) services to perform tasks autonomously, making decisions and taking actions
- Model-as-a-Service is the common approach
  - OpenAI, ChatGPT, Gemini, ...
- But what if you need to run the model closer to the data?

# Why run models in Kubernetes?

- Leverage existing infra and knowledge
- Network and data security (sovereignty)
- Support for popular open-source LLMs or BYO
- Reduce time to inference
- MaaS approach can be a black-box

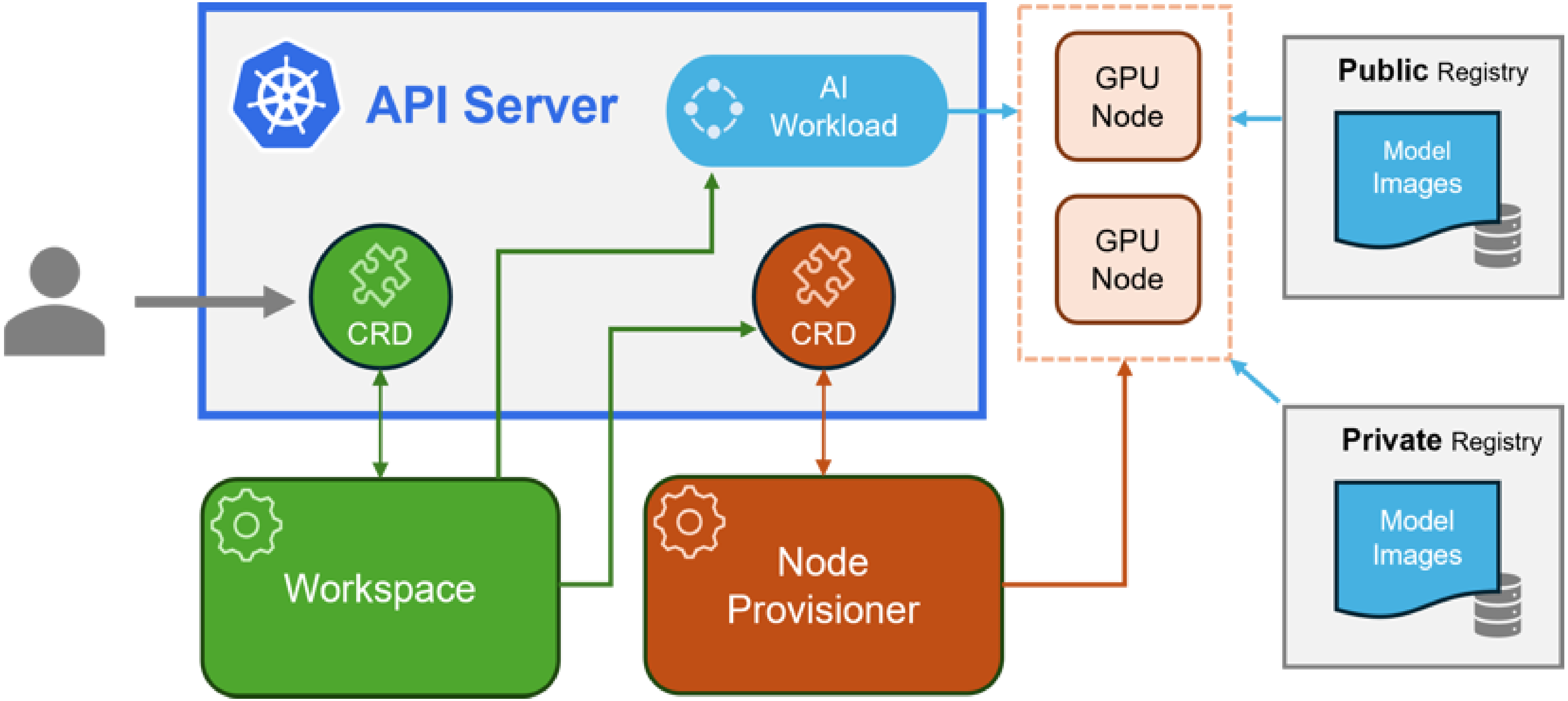
# Kaito

Kubernetes Community Days Munich 2024

# Kaito

- Kubernetes AI toolchain operator
- Open-source
- Managed or standalone
- Manage LLMs using container images
- Hardware
  - Presets eliminating the need for manual tuning of deployment parameters to fit GPU hardware
  - automatically provisions GPU nodes based on the requirements

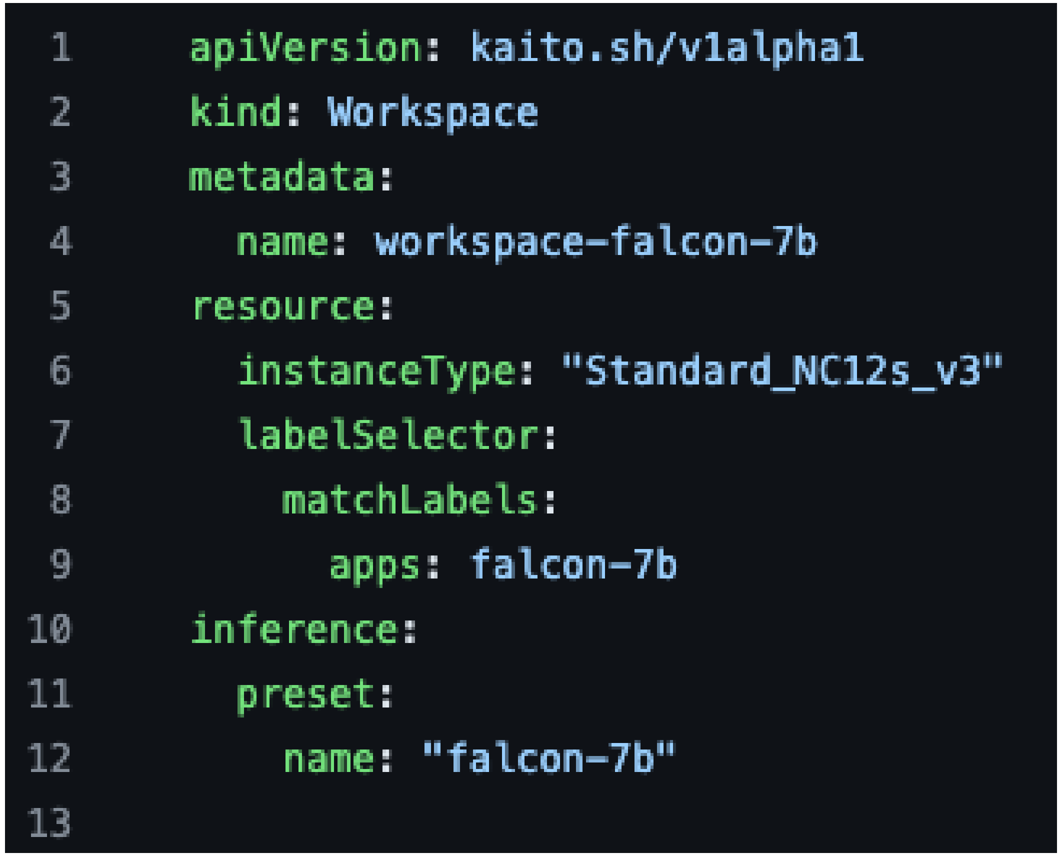
# Kaito



Picture source: https://github.com/Azure/kaito

# Kaito

- Model presets
  - falcon
  - llama2
  - llama2chat
  - mistral
  - phi-2
  - phi-3 (coming soon)



```yaml
apiVersion: kaito.sh/v1alpha1
kind: Workspace
metadata:
  name: workspace-falcon-7b
resource:
  instanceType: "Standard_NC12s_v3"
  labelSelector:
    matchLabels:
      apps: falcon-7b
inference:
  preset:
    name: "falcon-7b"

```

# Kaito

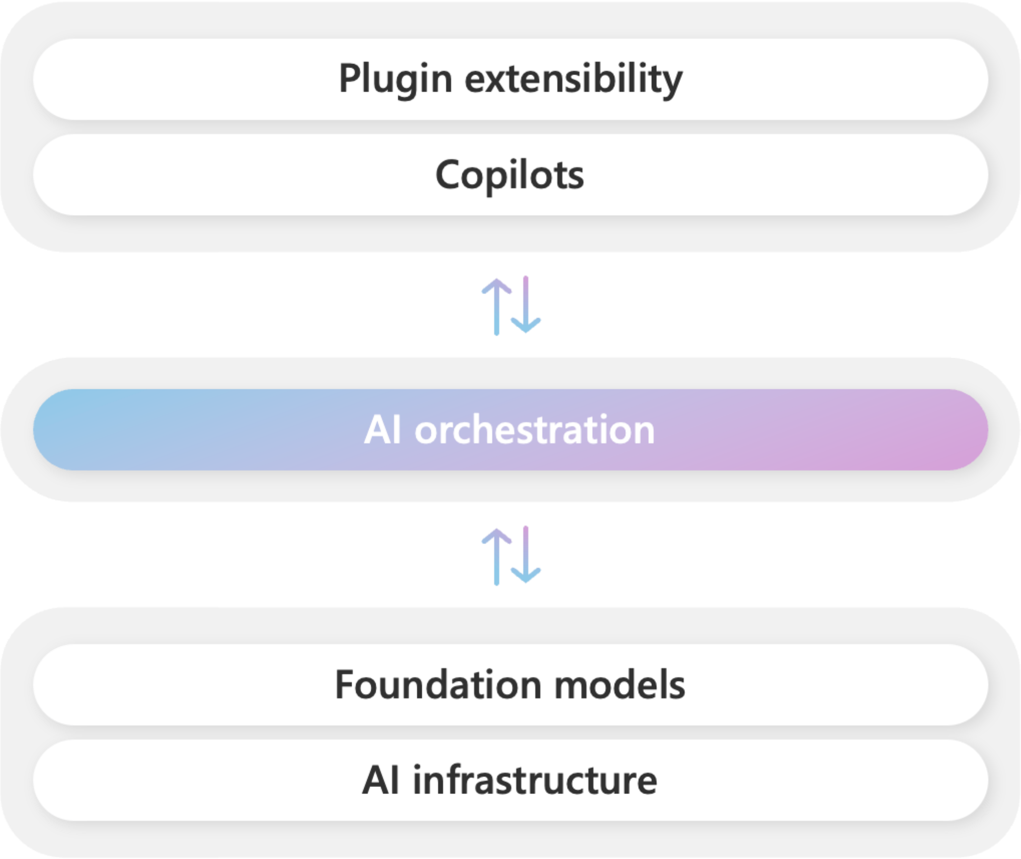
- Propose new OSS model
  - Kaito maintainers will setup the model presets configuration
- Deploy OSS model from your private registry
- Fine-tuning of the models
  - Leverage Workspace API for parameter-efficient fine-tuning (PEFT) of models

# Semantic Kernel

Kubernetes Community Days Munich 2024

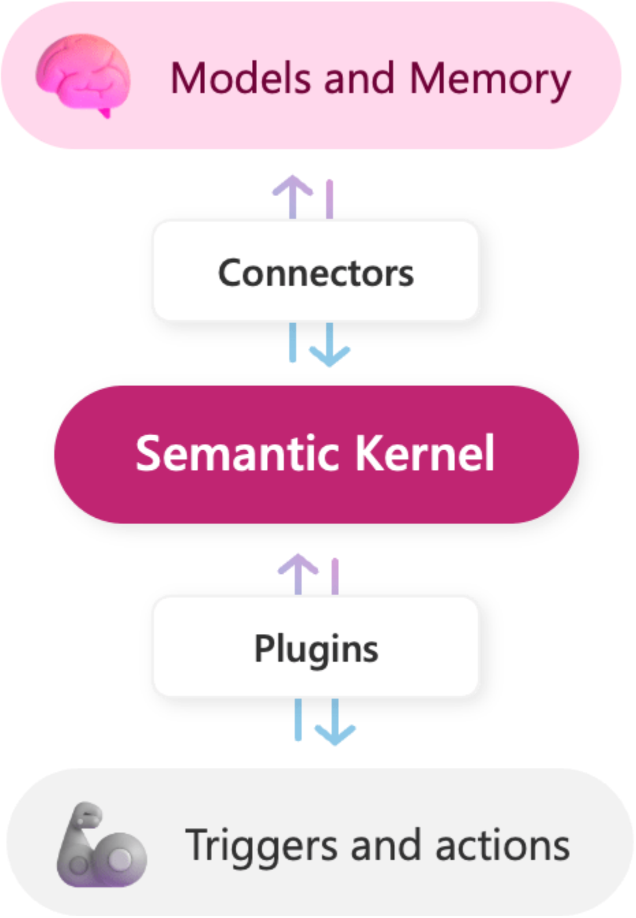
# Semantic Kernel

- Open Source SDK
- Seamlessly integrates AI services into your application
- Central component of an agent/copilot stack
- Supports multiple LLMs like Azure OpenAI and Hugging Face (no Lock-In to AI Providers)
- Available on multiple platforms like C#, Java and Python



# Semantic Kernel – Connectors & Plugins

- **Connectors:**
  - Integrate AI model
  - Add memories (Vector DB)
- **Plugins**
  - Easily Integrate Existing Code
  - Two Types
    - Native functions
    - Prompts
  - Out-of-the-box plugins (Math, HTTP, Wait, …)



# Semantic Kernel – Planners

- Automatically orchestrate plugins with AI
- Determines the optimal approach to fulfill the request
- May increase the costs but can also save money ☺

The planner
The plan

# Demo, demo, demo 🚀

- Demo
  - Falcon (Kaito) with Semantic Kernel
  - Azure OpenAI with Semantic Kernel
- Slides are also available within the repo
  - https://github.com/whiteducksoftware/demo-ai-kaito-semantic-kernel

# Questions?

**Philip Welz**

Senior Platform & Kubernetes Engineer,
Azure MVP

☎ +49 8031 230159-0

✉ philip.welz@whiteduck.de

@philip_welz

in www.linkedin.com/in/philip-welz

**Martin Brandl**

CTO, Cloud Solution Architect, Azure MVP

☎ +49 8031 230159-0

✉ martin.brandl@whiteduck.de

@martin_jib

in www.linkedin.com/in/mbrandl

white duck

Thank you!