Podman AI Lab

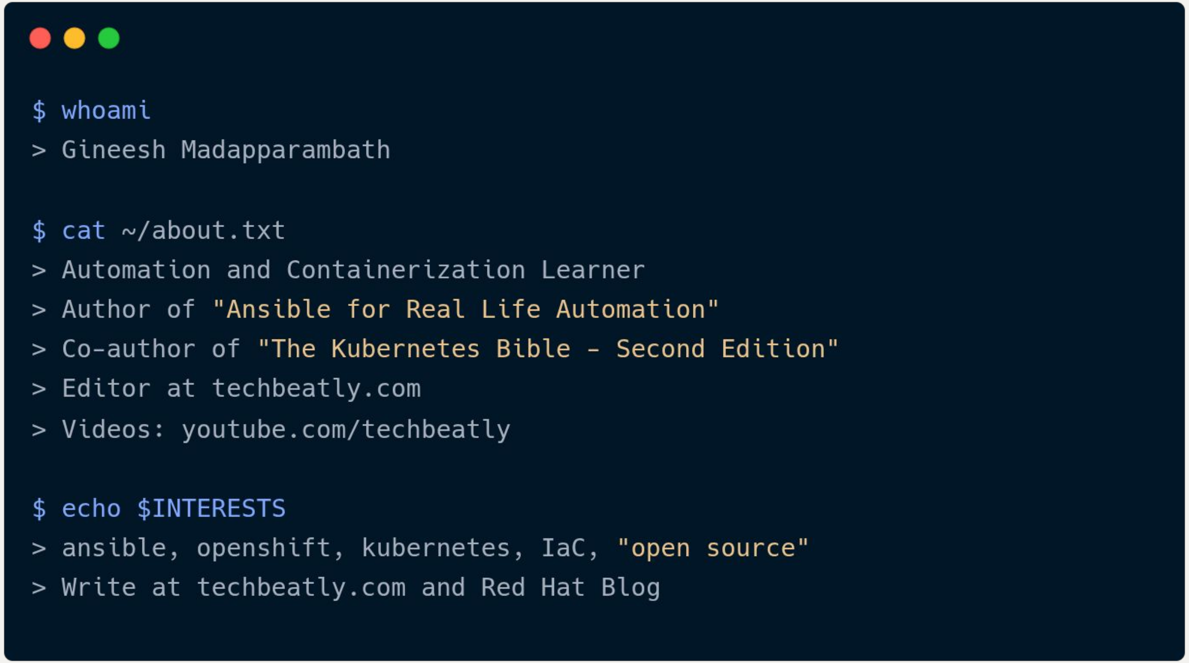
Local AI Development Made Simple

18 Mar 2025

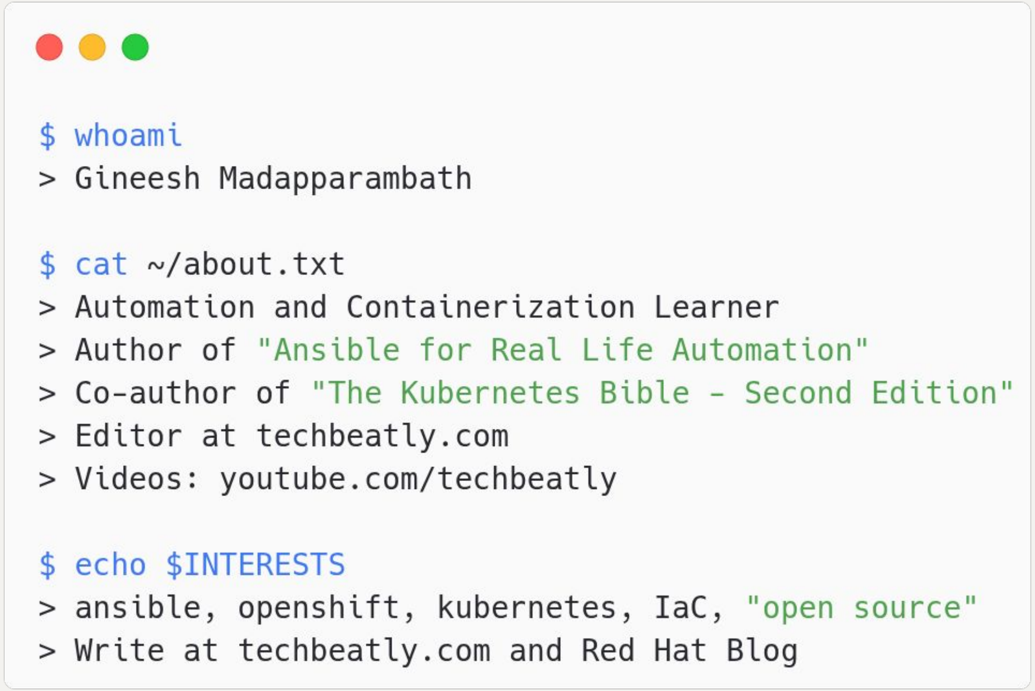
meetup.com/k8s-sg

We'll be discussing…

- Podman Introduction
- Podman Desktop Overview
- Podman AI Lab: Features
- Podman AI Lab: Deep Dive
- Local AI Benefits
- Demo
- Q&A



```
$ whoami
> Gineesh Madapparambath

$ cat ~/about.txt
> Automation and Containerization Learner
> Author of "Ansible for Real Life Automation"
> Co-author of "The Kubernetes Bible - Second Edition"
> Editor at techbeatly.com
> Videos: youtube.com/techbeatly

$ echo $INTERESTS
> ansible, openshift, kubernetes, IaC, "open source"
> Write at techbeatly.com and Red Hat Blog
```



```
$ whoami
> Gineesh Madapparambath

$ cat ~/about.txt
> Automation and Containerization Learner
> Author of "Ansible for Real Life Automation"
> Co-author of "The Kubernetes Bible - Second Edition"
> Editor at techbeatly.com
> Videos: youtube.com/techbeatly

$ echo $INTERESTS
> ansible, openshift, kubernetes, IaC, "open source"
> Write at techbeatly.com and Red Hat Blog
```

Gineesh Madapparambath (He/Him)

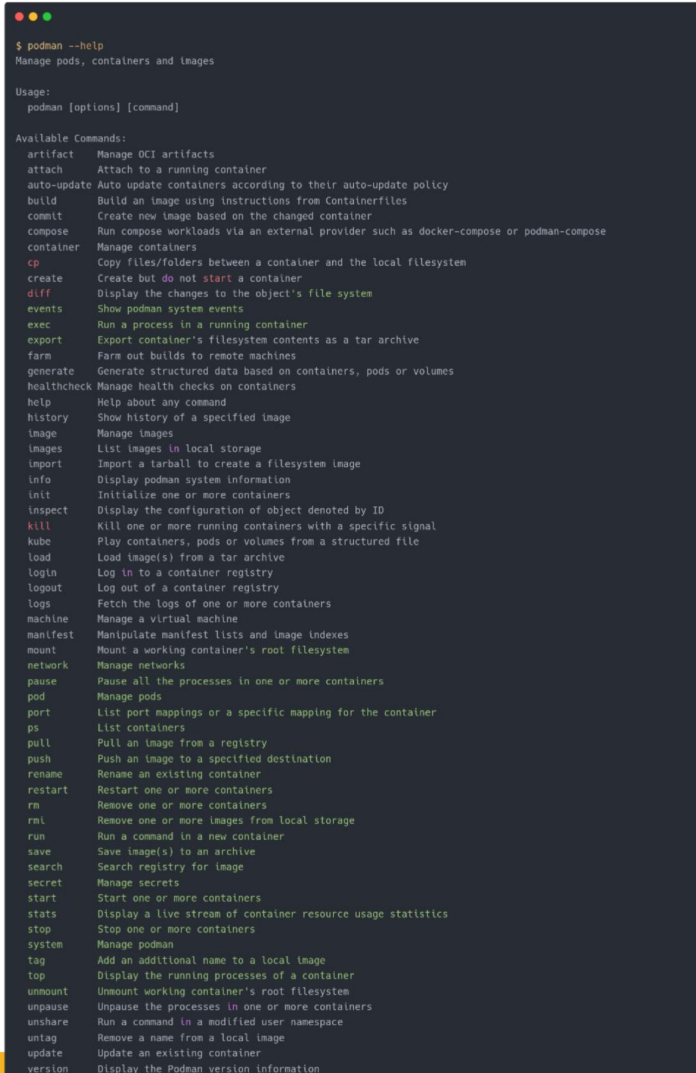
Red Hat

Author of Kubernetes and Ansible books, Automation and Containerization Explorer, techbeatly.com/youtube

iamgini.com

Podman: Rootless Container Engine

- Podman: Open-source container engine.
- Rootless Architecture: Runs containers without root privileges.
- Enhanced Security: Safer container development.
- Docker Compatibility: Works with Docker images and commands.
- Podman is designed to work with OCI-compatible containers and container images.
- Flexible Foundation: For containerizing applications.

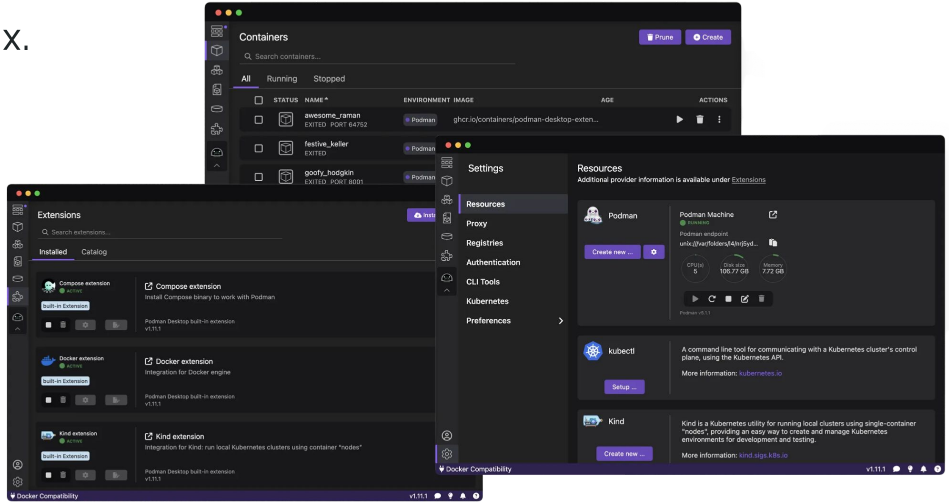


```
$ podman --help
Manage pods, containers and images

Usage:
  podman [options] [command]

Available Commands:
  artifact    Manage OCI artifacts
  attach      Attach to a running container
  auto-update Auto update containers according to their auto-update policy
  build       Build an image using instructions from Containerfiles
  commit      Create new image based on the changed container
  compose     Run compose workloads via an external provider such as docker-compose or podman-compose
  container   Manage containers
  cp          Copy files/folders between a container and the local filesystem
  create      Create but do not start a container
  diff        Display the changes to the object's file system
  events      Show podman system events
  exec        Run a process in a running container
  export      Export container's filesystem contents as a tar archive
  farm        Farm out builds to remote machines
  generate    Generate structured data based on containers, pods or volumes
  healthcheck Manage health checks on containers
  help        Help about any command
  history     Show history of a specified image
  image       Manage images
  images      List images in local storage
  import      Import a tarball to create a filesystem image
  info        Display podman system information
  init        Initialize one or more containers
  inspect     Display the configuration of object denoted by ID
  kill        Kill one or more running containers with a specific signal
  kube        Play containers, pods or volumes from a structured file
  load        Load image(s) from a tar archive
  login       Log in to a container registry
  logout      Log out of a container registry
  logs        Fetch the logs of one or more containers
  machine     Manage a virtual machine
  manifest    Manipulate manifest lists and image indexes
  mount       Mount a working container's root filesystem
  network     Manage networks
  pause       Pause all the processes in one or more containers
  pod         Manage pods
  port        List port mappings or a specific mapping for the container
  ps          List containers
  pull        Pull an image from a registry
  push        Push an image to a specified destination
  rename      Rename an existing container
  restart     Restart one or more containers
  rm          Remove one or more containers
  rmi         Remove one or more images from local storage
  run         Run a command in a new container
  save        Save image(s) to an archive
  search      Search registry for image
  secret      Manage secrets
  start       Start one or more containers
  stats       Display a live stream of container resource usage statistics
  stop        Stop one or more containers
  system      Manage podman
  tag         Add an additional name to a local image
  top         Display the running processes of a container
  unmount     Unmount working container's root filesystem
  unpause     Unpause the processes in one or more containers
  unshare     Run a command in a modified user namespace
  untag       Remove a name from a local image
  update      Update an existing container
  version     Display the Podman version information
```

Podman Desktop: Your Local Container Development Hub

- Free and open source!
- Control center for container workflows
- UI for managing containers, images, storage, local Kubernetes, OpenShift… and now AI
- It works across macOS, Windows, and Linux.

podman
desktop



STATUS	NAME	ENVIRONMENT	IMAGE	AGE	ACTIONS
	awesome_raman EXITED PORT 64752	Podman	ghcr.io/containers/podman-desktop-exten...		
	festive_keller EXITED	Podman			
	goofy_hodgkin EXITED PORT 8001	Podman			

Compose extension ACTIVE built-in Extension

Compose extension

Install Compose binary to work with Podman

Podman Desktop built-in extension v1.11.1

Docker extension ACTIVE built-in Extension

Docker extension

Integration for Docker engine

Podman Desktop built-in extension v1.11.1

Kind extension ACTIVE built-in Extension

Kind extension

Integration for Kind: run local Kubernetes clusters using container "nodes"

Podman Desktop built-in extension v1.11.1

Resources

Additional provider information is available under Extensions

Podman

Create new ...

Podman Machine RUNNING

Podman endpoint unix:///var/folders/l4/nrj5yd...

CPU(s) 5 Disk size 106.77 GB Memory 7.72 GB

Podman v5.1.1

kubectl

Setup ...

A command line tool for communicating with a Kubernetes cluster's control plane, using the Kubernetes API.

More information: kubernetes.io

Kind

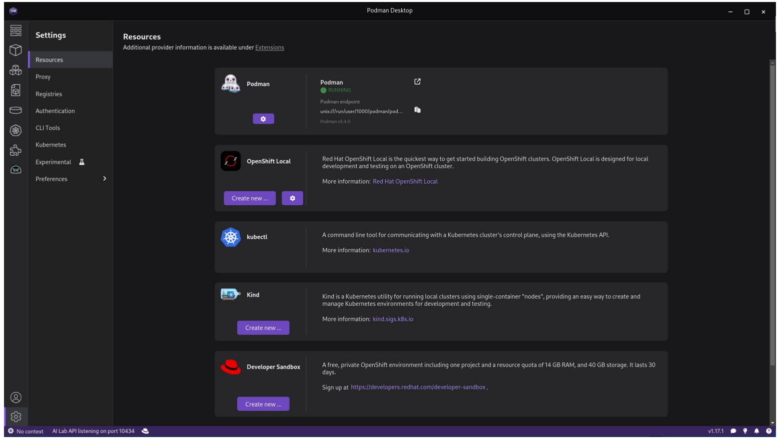
Create new ...

Kind is a Kubernetes utility for running local clusters using single-container "nodes", providing an easy way to create and manage Kubernetes environments for development and testing.

More information: kind.sigs.k8s.io

Podman Desktop: Everything from localhost

- Spin up Kubernetes clusters
- Connect to OpenShift local or Developer Sandbox



Resources

Additional provider information is available under Extensions

Podman

Podman
RUNNING
Podman endpoint
unix:///run/user/1000/podman/pod...
Podman v5.4.0

OpenShift Local

Red Hat OpenShift Local is the quickest way to get started building OpenShift clusters. OpenShift Local is designed for local development and testing on an OpenShift cluster.

More information: Red Hat OpenShift Local

kubectl

A command line tool for communicating with a Kubernetes cluster's control plane, using the Kubernetes API.

More information: kubernetes.io

Kind

Kind is a Kubernetes utility for running local clusters using single-container "nodes", providing an easy way to create and manage Kubernetes environments for development and testing.

More information: kind.sigs.k8s.io

Developer Sandbox

A free, private OpenShift environment including one project and a resource quota of 14 GB RAM, and 40 GB storage. It lasts 30 days.

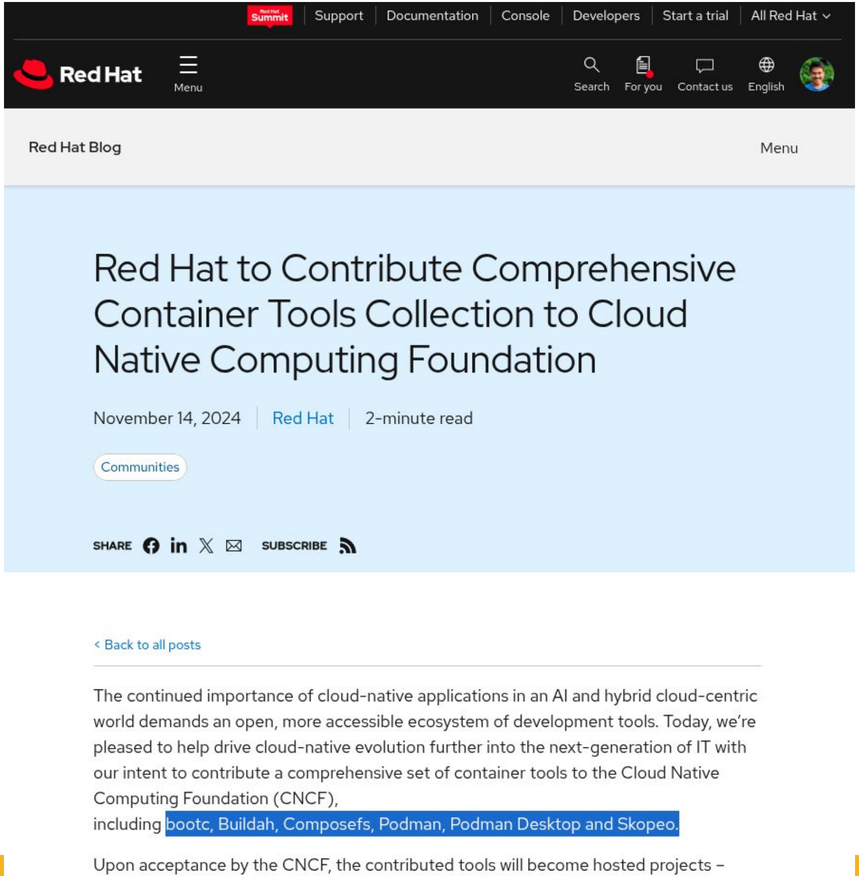
Sign up at https://developers.redhat.com/developer-sandbox .

No context   AI Lab API listening on port 10434

Podman Desktop -> CNCF

***Red Hat Blog**: The continued importance of cloud-native applications in an AI and hybrid cloud-centric world demands an open, more accessible ecosystem of development tools. Today, we're pleased to help drive cloud-native evolution further into the next-generation of IT with our intent to contribute a comprehensive set of container tools to the Cloud Native Computing Foundation (CNCF), including **bootc**, **Buildah**, **Composefs**, **Podman**, **Podman Desktop** and **Skopeo**.*

https://www.redhat.com/en/blog/red-hat-contribute-comprehensive-container-tools-collection-cloud-native-computing-foundation



Red Hat to Contribute Comprehensive Container Tools Collection to Cloud Native Computing Foundation

November 14, 2024 | Red Hat | 2-minute read

The continued importance of cloud-native applications in an AI and hybrid cloud-centric world demands an open, more accessible ecosystem of development tools. Today, we're pleased to help drive cloud-native evolution further into the next-generation of IT with our intent to contribute a comprehensive set of container tools to the Cloud Native Computing Foundation (CNCF), including bootc, Buildah, Composefs, Podman, Podman Desktop and Skopeo.

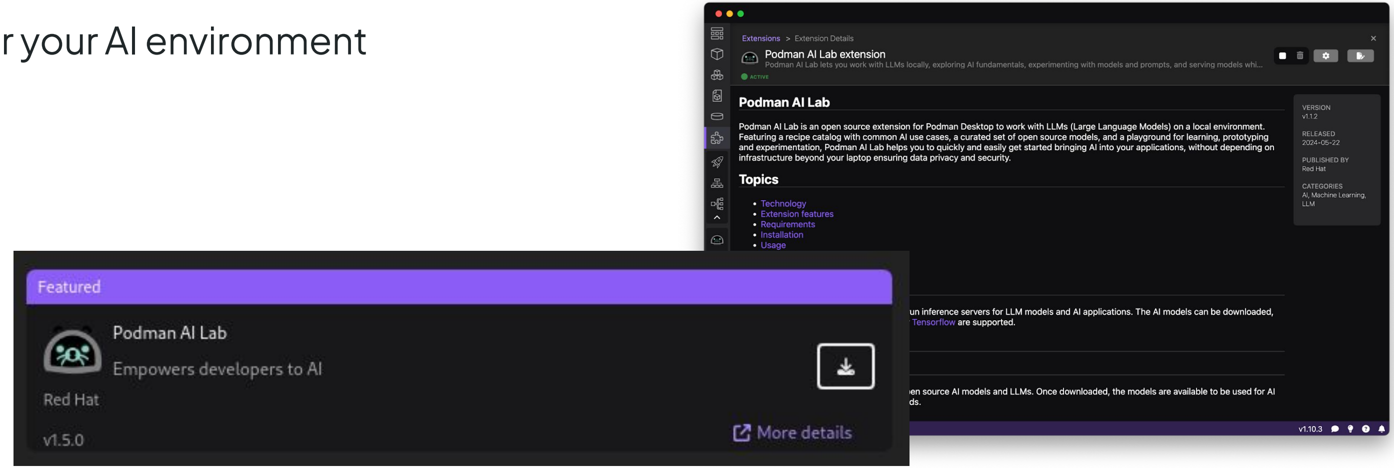
Upon acceptance by the CNCF, the contributed tools will become hosted projects –

Using CNCF projects with Podman Desktop

- minikube
- Backstage
- Dapr
- and many others

Podman AI Lab: Features & Benefits

- Read to install extension
- Experiment, develop, and deploy AI applications without relying solely on cloud resources
- Privacy, reduced latency, offline development, and control over your AI environment



Podman AI Lab lets you work with LLMs locally, exploring AI fundamentals, experimenting with models and prompts, and serving models whi...

Podman AI Lab

Podman AI Lab is an open source extension for Podman Desktop to work with LLMs (Large Language Models) on a local environment. Featuring a recipe catalog with common AI use cases, a curated set of open source models, and a playground for learning, prototyping and experimentation, Podman AI Lab helps you to quickly and easily get started bringing AI into your applications, without depending on infrastructure beyond your laptop ensuring data privacy and security.

Topics

- Technology
- Extension features
- Requirements
- Installation
- Usage

un inference servers for LLM models and AI applications. The AI models can be downloaded,
Tensorflow are supported.

en source AI models and LLMs. Once downloaded, the models are available to be used for AI
ds.

VERSION
v1.1.2

RELEASED
2024-05-22

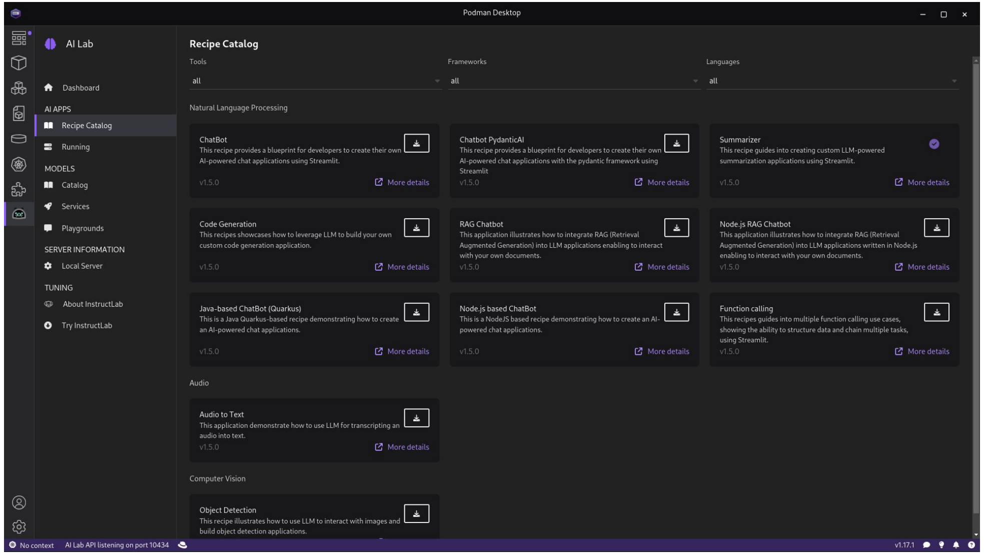
PUBLISHED BY
Red Hat

CATEGORIES
AI, Machine Learning, LLM

Podman AI Lab
Empowers developers to AI
Red Hat
v1.5.0

Curated catalog of AI recipes,

Explore pre-built AI use cases with ready-to-use models and code examples. Easily swap models for comparison.



Recipe Catalog

Tools: all
Frameworks: all
Languages: all

Natural Language Processing

ChatBot
This recipe provides a blueprint for developers to create their own AI-powered chat applications using Streamlit.
v1.5.0 More details

Chatbot PydanticAI
This recipe provides a blueprint for developers to create their own AI-powered chat applications with the pydantic framework using Streamlit
v1.5.0 More details

Summarizer
This recipe guides into creating custom LLM-powered summarization applications using Streamlit.
v1.5.0 More details

Code Generation
This recipes showcases how to leverage LLM to build your own custom code generation application.
v1.5.0 More details

RAG Chatbot
This application illustrates how to integrate RAG (Retrieval Augmented Generation) into LLM applications enabling to interact with your own documents.
v1.5.0 More details

Node.js RAG Chatbot
This application illustrates how to integrate RAG (Retrieval Augmented Generation) into LLM applications written in Node.js enabling to interact with your own documents.
v1.5.0 More details

Java-based ChatBot (Quarkus)
This is a Java Quarkus-based recipe demonstrating how to create an AI-powered chat applications.
v1.5.0 More details

Node.js based ChatBot
This is a NodeJS based recipe demonstrating how to create an AI-powered chat applications.
v1.5.0 More details

Function calling
This recipes guides into multiple function calling use cases, showing the ability to structure data and chain multiple tasks, using Streamlit.
v1.5.0 More details

Audio

Audio to Text
This application demonstrate how to use LLM for transcripting an audio into text.
v1.5.0 More details

Computer Vision

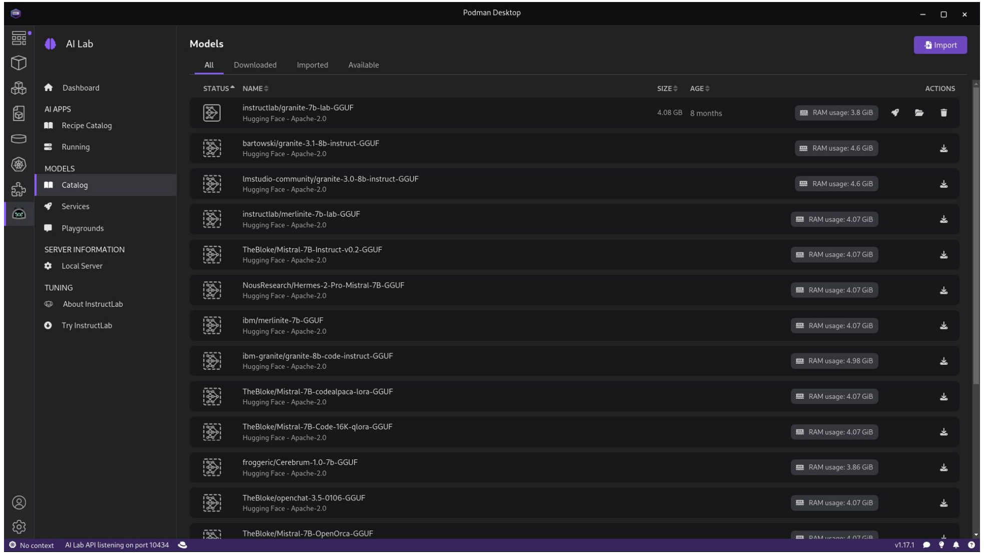
Object Detection
This recipe illustrates how to use LLM to interact with images and build object detection applications.

Model Catalog

Access a selection of open-source LLMs with clear license information.

Import your own models in formats like GGUF

You can also import your own AI models to run within Podman AI Lab.

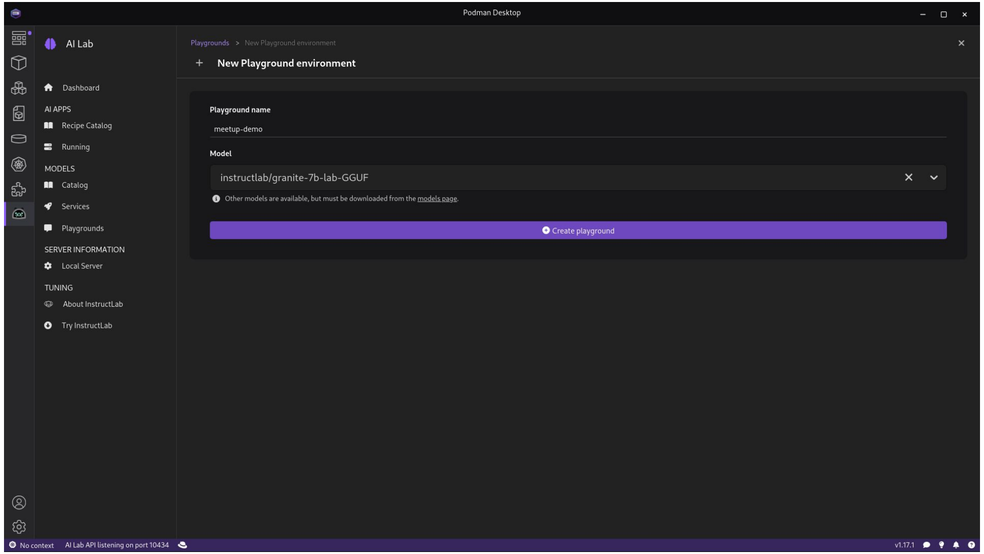


Models

STATUS	NAME	SIZE	AGE		ACTIONS
	instructlab/granite-7b-lab-GGUF Hugging Face - Apache-2.0	4.08 GB	8 months	RAM usage: 3.8 GiB	
	bartowski/granite-3.1-8b-instruct-GGUF Hugging Face - Apache-2.0			RAM usage: 4.6 GiB	
	lmstudio-community/granite-3.0-8b-instruct-GGUF Hugging Face - Apache-2.0			RAM usage: 4.6 GiB	
	instructlab/merlinite-7b-lab-GGUF Hugging Face - Apache-2.0			RAM usage: 4.07 GiB	
	TheBloke/Mistral-7B-Instruct-v0.2-GGUF Hugging Face - Apache-2.0			RAM usage: 4.07 GiB	
	NousResearch/Hermes-2-Pro-Mistral-7B-GGUF Hugging Face - Apache-2.0			RAM usage: 4.07 GiB	
	ibm/merlinite-7b-GGUF Hugging Face - Apache-2.0			RAM usage: 4.07 GiB	
	ibm-granite/granite-8b-code-instruct-GGUF Hugging Face - Apache-2.0			RAM usage: 4.98 GiB	
	TheBloke/Mistral-7B-codealpaca-lora-GGUF Hugging Face - Apache-2.0			RAM usage: 4.07 GiB	
	TheBloke/Mistral-7B-Code-16K-qlora-GGUF Hugging Face - Apache-2.0			RAM usage: 4.07 GiB	
	froggeric/Cerebrum-1.0-7b-GGUF Hugging Face - Apache-2.0			RAM usage: 3.86 GiB	
	TheBloke/openchat-3.5-0106-GGUF Hugging Face - Apache-2.0			RAM usage: 4.07 GiB	
	TheBloke/Mistral-7B-OpenOrca-GGUF			RAM usage: 4.07 GiB	

Playground

Experiment with models, adjust parameters like temperature and max tokens, and fine-tune your prompts.



New Playground environment

Playground name: meetup-demo

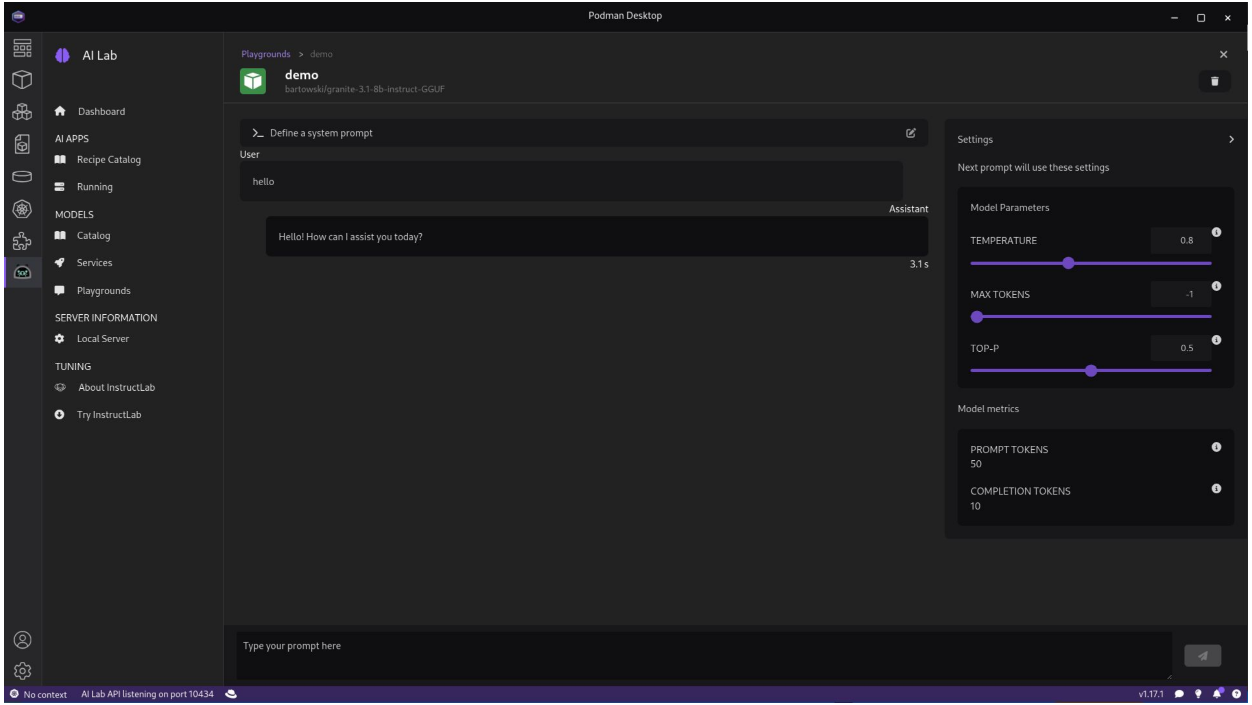
Model: instructlab/granite-7b-lab-GGUF

Other models are available, but must be downloaded from the models page.

No context AI Lab API listening on port 10434

v1.17.1

Interact with Model in Playground



demo

bartowski/granite-3.1-8b-instruct-GGUF

Define a system prompt

User

hello

Assistant

Hello! How can I assist you today?

3.1 s

Settings

Next prompt will use these settings

Model Parameters

TEMPERATURE 0.8

MAX TOKENS -1

TOP-P 0.5

Model metrics

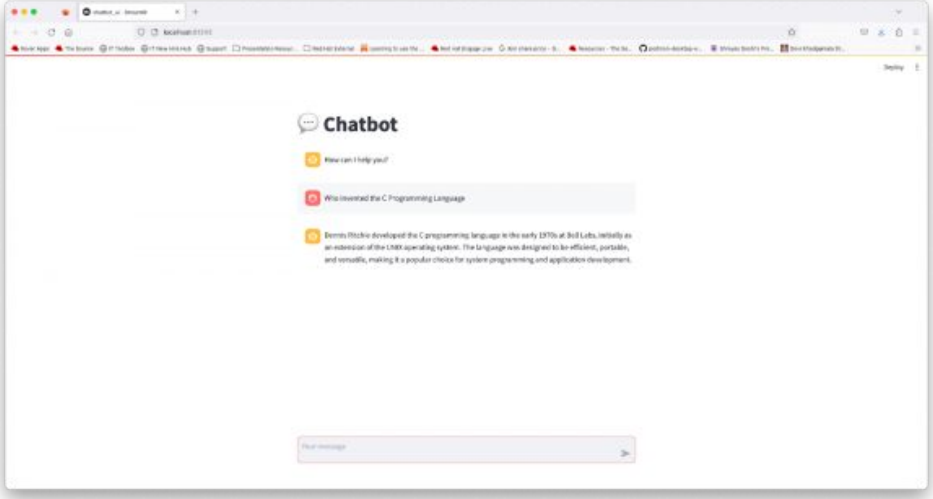
PROMPT TOKENS 50

COMPLETION TOKENS 10

Type your prompt here

Local Model Serving

Deploy model inference servers locally using containers, with OpenAI-compatible APIs for easy integration.



Chatbot

How can I help you?

Who invented the C Programming Language

Dennis Ritchie developed the C programming language in the early 1970s at Bell Labs, initially as an extension of the UNIX operating system. The language was designed to be efficient, portable, and versatile, making it a popular choice for system programming and application development.

Why Local AI Development?

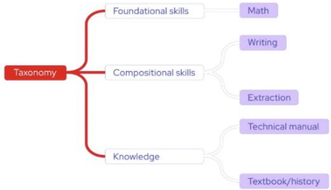
- Data Privacy: Keep sensitive data on your local machine.
- Reduced Latency: Faster response times for development and testing.
- Offline Development: Work on AI projects without an internet connection.
- Cost-Effective: Minimize cloud resource usage.
- Control: Full control over your AI environment and dependencies.

InstructLab: Collaborative LLM Customization

Leverage a taxonomy-guided synthetic data generation process and a multi-phase tuning framework to improve model performance.

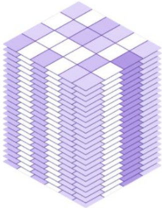
- Open Collaboration
- Community-Driven LLM Improvement
- Simple Contribution Process
- Focus on Teachability
- Scalable LLM Enhancement
- Open Model Weights
- Empowering LLM Customization

InstructLab: How it helps



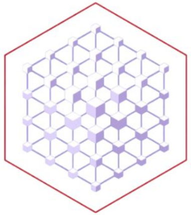
Taxonomy-driven Data Curation

Folder structure with Q&A pairs for topics to teach a model



Large-scale Synthetic Data Generation

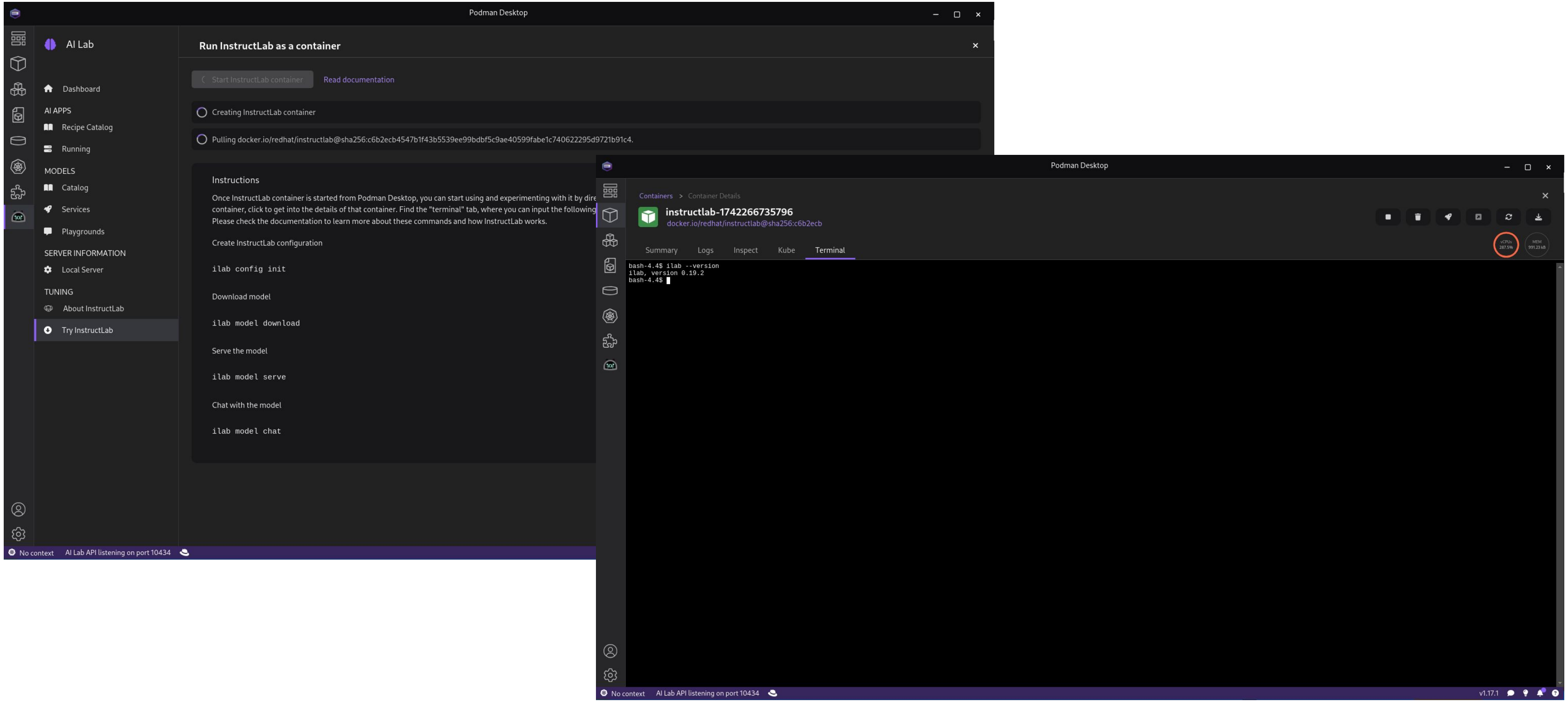
Generate additional training data to expand dataset automatically



Model Training with New Data

Phased, large-scale alignment tuning (with knowledge and skills)

Podman and InstructLab



Run InstructLab as a container

○ Creating InstructLab container

○ Pulling docker.io/redhat/instructlab@sha256:c6b2ecb4547b1f43b5539ee99bdbf5c9ae40599fabe1c740622295d9721b91c4.

Instructions

Once InstructLab container is started from Podman Desktop, you can start using and experimenting with it by dire
container, click to get into the details of that container. Find the "terminal" tab, where you can input the following
Please check the documentation to learn more about these commands and how InstructLab works.

Create InstructLab configuration

```
ilab config init
```

Download model

```
ilab model download
```

Serve the model

```
ilab model serve
```

Chat with the model

```
ilab model chat
```

instructlab-1742266735796

docker.io/redhat/instructlab@sha256:c6b2ecb

```
bash-4.4$ ilab --version
ilab, version 0.19.2
bash-4.4$ 
```

Photo by Christina Morillo from Pexels

Let us see them in Action

System and Software Prerequisites

- Software: Podman Desktop 1.10.3+ and Podman 5.0.1+.
- Hardware: At least 12GB RAM, 4+ CPUs (more is better for large models).

References

Podman: https://podman.io

Podman Desktop: https://podman-desktop.io

Have fun coloring and learn about Podman: https://developers.redhat.com/e-books/container-commandos-coloring-book

Build your AI application with AI Lab extension in Podman Desktop (Hands on Lab):

https://developers.redhat.com/learn/rhel/build-your-ai-application-ai-lab-extension-podman-desktop

Getting Started with AI Lab on Podman Desktop (Video): https://www.youtube.com/watch?v=xuyk39iO2LI

InstructLab: https://instructlab.ai

Questions & Feedback

Photo by Andrea Piacquadio from Pexels

Thank You

linkedin.com/in/gineesh

iamgini.com

Join us – K8SUG Singapore

meetup.com/k8s-sg