**Journey to Master AI/ML with Docker**

**A Practical Handbook for Docker with 3 Hands-On Projects**

**Vietnamese Full Name: Nguyen Cong Phat**

**My portfolio website:** [**https://my-portfolio-website-gamma-lemon.vercel.app/**](https://my-portfolio-website-gamma-lemon.vercel.app/)

**My github profile:** [**https://github.com/paht2005**](https://github.com/paht2005)

**A person holding a tablet in front of a container

AI-generated content may be incorrect.**

Image 1.

Table of Contents

[A. 📖 Introduction: Why Docker for AI/ML? 3](#_Toc199611719)

[1. 🐳 What is Docker and Why It Matters in AI/ML 3](#_Toc199611720)

[2. Docker vs Virtual Machines in ML Workflows 3](#_Toc199611721)

[3. 👨‍🔬 Who Benefits from Using Docker in AI/ML? 3](#_Toc199611722)

[4. 🏢 Real-World Adoption: How Industry Leaders Use Docker for AI 3](#_Toc199611723)

[a. Netflix: Orchestrating ML Workflows Using Docker 4](#_Toc199611724)

[b. Uber: Streamlining ML Pipelines with Docker 4](#_Toc199611725)

[c. Walmart: Scaling AI Solutions with Docker 4](#_Toc199611726)

[d. NASA: Accelerating Data Analysis with Docker 4](#_Toc199611727)

[e. Ingka Group (IKEA): Scalable MLOps with Docker and Kubernetes 4](#_Toc199611728)

[f. ZEISS Microscopy: Cross-Platform AI Model Deployment 5](#_Toc199611729)

[B. 🔧 Docker Essentials for AI/ML 5](#_Toc199611730)

[1. Images vs Containers 5](#_Toc199611731)

[2. Key Docker Commands for ML Practitioners 5](#_Toc199611732)

[3. Docker Networking, Volumes & Detached Mode 5](#_Toc199611733)

[4. Container Lifecycle Management 6](#_Toc199611734)

[C. 📦 ML Workflow & Toolchain with Docker 6](#_Toc199611735)

[1. Where Docker Fits into the ML Workflow? 6](#_Toc199611736)

[2. Development Environments 6](#_Toc199611737)

[3. Experiment Tracking 7](#_Toc199611738)

[4. Model Serving & APIs 7](#_Toc199611739)

[5. Deployment Options 7](#_Toc199611740)

[D. Docker in the world of LLMs / Agentic AI 8](#_Toc199611741)

[1. Running Models with Docker Model Runner 8](#_Toc199611742)

[2. Docker + MCP Tooling 9](#_Toc199611743)

[3. Deploying NVIDIA NIM with Docker 9](#_Toc199611744)

[4. Agentic AI + Docker 10](#_Toc199611745)

[5. Sample Agentic DevOps Setup 11](#_Toc199611746)

[E. Project 1: Dockerize Your AI App 11](#_Toc199611747)

[1. Goal: Take any Python-based AI app and: 11](#_Toc199611748)

[2. Example App: FastAPI Sentiment Classifier 11](#_Toc199611749)

[a. Step 1: Create Your App Files 11](#_Toc199611750)

[b. Step 2: Create the Dockerfile 13](#_Toc199611751)

[c. Step 3: Build and Run the Docker Image 13](#_Toc199611752)

# 📖 Introduction: Why Docker for AI/ML?

## 🐳 What is Docker and Why It Matters in AI/ML

* **Docker** is a lightweight containerization platform that enables you to package code, dependencies, and system configurations into a single, portable unit — a container. In AI/ML, where experiments rely heavily on specific Python versions, library dependencies, and GPU support, Docker provides a reproducible and isolated environment for development and deployment.
* In contrast to traditional "it works on my machine" scenarios, Docker ensures consistent behavior across laptops, development servers, CI/CD pipelines, and cloud environments. This is especially critical in machine learning, where reproducibility is essential for debugging, collaboration, and compliance.
* Docker simplifies the ML lifecycle by enabling:
* Faster environment setup for Jupyter, MLFlow, FastAPI, etc.
* Portable packaging of trained models for inference
* Seamless transition from local development to cloud deployment
* Scalability across GPU clusters or Kubernetes-based infra

## Docker vs Virtual Machines in ML Workflows

|  |  |  |
| --- | --- | --- |
| **Feature** | **Docker** | **Virtual Machines** |
| **Startup Time** | **⚡️Fast: Seconds (near instant)** | **🐢Slow:** Minutes (b**oot OS)** |
| **Resource Use** | **Efficient:** Lightweight, shares kernel | **Heavy:** Heavy, full OS per V**M** |
| **Portability** | **High** | **Medium** |
| **Isolation** | High (via namespaces) | Very High (hardware **emulation)** |
| **Dev Workflow** | Smooth, scriptable | Clunky, manual provisioning |

Docker eliminates the overhead of full VMs while maintaining strong isolation — making it ideal for iterative ML workflows where fast provisioning and reproducibility are critical.

## 👨‍🔬 Who Benefits from Using Docker in AI/ML?

* **ML Engineers**: Reproducible training pipelines, scalable model serving.
* **Data Scientists**: Rapid experimentation without dependency hell.
* **DevOps Teams**: Simplified deployment across staging and production.
* **AI Hobbyists**: Run state-of-the-art models locally with minimal setup.

By using containers, teams become 10x more productive. They spend less time setting up environments and more time building models and delivering results.

## 🏢 Real-World Adoption: How Industry Leaders Use Docker for AI

Big names rely on containers to make AI reliable

### Netflix: Orchestrating ML Workflows Using Docker

Netflix utilizes Docker containers to **manage and scale its machine learning workflows** efficiently. By containerizing their ML tasks, Netflix ensures consistent environments across development and production, aiding in tasks like **content recommendation and streaming optimization.**

**Source :** [**Maestro: Data/ML Workflow Orchestrator at Netflix | by Netflix Technology Blog | Netflix TechBlog**](https://netflixtechblog.com/maestro-netflixs-workflow-orchestrator-ee13a06f9c78)

### Uber: Streamlining ML Pipelines with Docker

Uber employs Docker to **containerize its machine learning workflows, facilitating consistent environments** for development and deployment. This approach enhances the scalability and reproducibility of their ML models, which are integral to services like ETA predictions and dynamic pricing.

**Source :** [**From Predictive to Generative - How Michelangelo Accelerates Uber’s AI Journey | Uber Blog**](https://www.uber.com/en-IN/blog/from-predictive-to-generative-ai/?utm_source=chatgpt.com)

### Walmart: Scaling AI Solutions with Docker

Walmart leverages Docker to **containerize its AI applications**, facilitating **scalable and efficient** **deployment** across its vast retail infrastructure. This strategy supports various use cases, including **inventory management and customer experience** enhancements.

**Source :** [**Machine Learning Platform at Walmart | by Thomas Vengal | Walmart Global Tech Blog | Medium**](https://medium.com/walmartglobaltech/machine-learning-platform-at-walmart-b06819825ef7)

### NASA: Accelerating Data Analysis with Docker

NASA employs Docker containers to **standardize and expedite** its machine learning workflows, particularly in processing vast amounts of satellite data. Containerization aids in maintaining consistent environments, crucial for the reproducibility of scientific analyses.

**Source :** [**How NASA JPL Processes 70 TB of Satellite Data Products a Day Using Amazon EC2 Auto Scaling with Spot Instances | Case Study | AWS**](https://aws.amazon.com/solutions/case-studies/nasa-jpl-spot-case-study/?utm_source=chatgpt.com)

### Ingka Group (IKEA): Scalable MLOps with Docker and Kubernetes

Ingka Group, the parent company of IKEA, adopted Docker and Kubernetes to build a robust **MLOps platform**. This setup allows for **dynamic scaling** of AI/ML applications, **improved collaboration** through **uniform development environments**, and **enhanced security**. The containerized approach accelerates prototyping and deployment of new models, aligning with IKEA's commitment to innovation

**Source:** [**Case Study: Ingka and Docker | Docker**](https://www.docker.com/customer-stories/ingka/?utm_source=chatgpt.com)

### ZEISS Microscopy: Cross-Platform AI Model Deployment

ZEISS, a leader in optics and optoelectronics, utilizes **Docker to deploy AI models** across various platforms, including **cloud and local Windows-based clients**. By containerizing their AI solutions, ZEISS ensures **consistent performance** and **simplifies the distribution** of complex models, enhancing their microscopy software's capabilities.

**Source:** [**Case Study: ZEISS and Docker | Docker**](https://www.docker.com/customer-stories/zeiss/?utm_source=chatgpt.com)

# 🔧 Docker Essentials for AI/ML

## Images vs Containers

* **Docker Image**: A static blueprint that defines what is inside the container, including the base OS, Python environment, ML libraries (like scikit-learn, pandas, torch), and your code.
* **Docker Container**: A running instance of an image — like a live, isolated process with its own filesystem, ports, and environment variables.

In ML projects, Docker images capture the exact environment needed to run notebooks, train models, or serve APIs, while containers execute those tasks consistently across platforms.

## Key Docker Commands for ML Practitioners

 docker build -t <image-name> .  
→ Build an image from a Dockerfile.

 docker run -p <host-port>:<container-port> <image-name>  
→ Run a container and map ports for web apps (e.g., Jupyter or FastAPI).

 docker ps  
→ List running containers.

 docker logs <container-name>  
→ View logs of a running or stopped container.

 docker exec -it <container-name> bash  
→ Open an interactive shell inside a container.

These commands are essential for tracking, debugging, and managing containerized ML workloads.

## Docker Networking, Volumes & Detached Mode

* **Networking (-p)**: Exposes container ports to the host system.
  + E.g., MLFlow at localhost:5555, Jupyter at localhost:8888.
* **Volumes (-v)**: Maps host directories to the container for persistent storage — useful for saving notebooks or models.
* **Detached Mode (-d)**: Runs the container in the background, ideal for long-running services like APIs or experiment trackers.

These options enable full control and flexibility during ML experimentation and deployment.

## Container Lifecycle Management

* docker stop <id> / docker start <id>: Pause/resume containers.
* docker rm <id>: Remove containers.
* docker images / docker rmi <image>: Manage image versions.
* docker logs -f <name>: Stream logs live (useful during model training or API serving).
* docker exec -it <name> sh: Debug inside the container.

Mastering lifecycle commands is crucial for iterating quickly and maintaining a clean dev environment.

# 📦 ML Workflow & Toolchain with Docker

## Where Docker Fits into the ML Workflow?

A diagram of a model

AI-generated content may be incorrect.

Image 2. Typical ML Workflow

## Development Environments

* **Jupyter Notebooks**: Easily run notebooks inside prebuilt images like jupyter/scipy-notebook, preserving your work with volume mounts.
* **VS Code (Remote - Containers)**: Attach your editor directly to the container environment for seamless ML development.

## Experiment Tracking

* **MLFlow**: Lightweight experiment manager and model registry — easily run via container on port 5555.
* **DVC (Data Version Control)**: Manages datasets and model versions alongside code.
* **Weights & Biases**: Industry-grade experiment tracking with rich visualizations and hyperparameter tuning support.

Docker enables these tools to be quickly spun up without manual setup.

## Model Serving & APIs

 **FastAPI / Flask**: Pythonic web frameworks to serve ML models via REST APIs — containerized for portability and deployment ease.

 **TensorFlow Serving / TorchServe**: Production-grade model servers optimized for TensorFlow or PyTorch.

 **Gradio / Streamlit**: For rapid prototyping and interactive demos, often deployed via Docker for reproducibility.

## Deployment Options

* **Docker Hub**: Hosts and distributes public/private images for reuse.
* **Kubernetes**: Container orchestration system for scalable, distributed ML workloads.
* **Hugging Face Spaces (Docker SDK)**: Fastest way to deploy ML apps with a Dockerfile and GitHub repo — free for most use cases.

Using Docker as the foundation makes these deployment targets interchangeable and production-ready.

# Docker in the world of LLMs / Agentic AI

A container with a brain on it

AI-generated content may be incorrect.

Image 3.

## Running Models with Docker Model Runner

* **Dockerized abstraction of any trained model:** Package your model with all dependencies
* **Launch a ready-to-serve REST API in seconds:** No need to write custom API code
* **Framework agnostic (TF, Torch, XGBoost):** Works with all major ML frameworks
* **Example:** docker run -p 8080:8080 ghcr.io/mlc-ai/model-runner:latest

## Docker + MCP Tooling

A diagram of a diagram of a diagram

AI-generated content may be incorrect.

Image 4. Source: <https://hoangndst.com/blog/model-context-protocol>

* **What's MCP?** Model Context Protocol lets AI models access real-world tools
* **Docker + MCP =**
* Self-hosted MCP toolkits (Terraform, Kubernetes, CLI agents)
* Tool-aware autonomous agents
* **Example:** docker run -p 3000:3000 realops/kubernetes-mcp server:latest

## Deploying NVIDIA NIM with Docker

A diagram of a container

AI-generated content may be incorrect.

Image 5. Source: <https://developer.nvidia.com/blog/nvidia-docker-gpu-server-application-deployment-made-easy/>

* **Containerized inference microservices** (Mixtral, LLaMA2, etc.)
* **REST interface for easy integration**
* **GPU-accelerated, secure, scalable**
* **Example:** docker run --gpus all -p 8000:8000 nvcr.io/nim/mixtral:latest

A computer with green squares and white text

AI-generated content may be incorrect.

Image 6. Souce: <https://pub.towardsai.net/an-introduction-to-using-nvidias-nim-api-da653041b212>

## Agentic AI + Docker

* **Agentic AI** = LLMs + Tools + Memory + Goals

A screenshot of a computer

AI-generated content may be incorrect.

Image 7.

Build and deploy autonomous AI agents with ease

## Sample Agentic DevOps Setup

A diagram of a software structure

AI-generated content may be incorrect.

Image 8. Sample Agentic DevOps Setup

Reproducible. Scalable. Composable

# Project 1: Dockerize Your AI App

## Goal: Take any Python-based AI app and:

* Package it into a Docker container
* Include models, dependencies, and web interface
* Run it anywhere with a single command

## Example App: FastAPI Sentiment Classifier

### Step 1: Create Your App Files

* **File app.py**

# file app.py  
  
from fastapi import FastAPI

from pydantic import BaseModel

from transformers import pipeline

app = FastAPI()

classifier = pipeline("sentiment-analysis")

class InputText(BaseModel):

    text: str

@app.post("/predict")

def predict(input: InputText):

    result = classifier(input.text)[0]

    return {

        "label": result["label"],

        "confidence": round(float(result["score"]) \* 100, 2)

    }

* **File requirements.txt**

fastapi

uvicorn

transformers

torch

### Step 2: Create the Dockerfile

# Use Python base image

FROM python:3.10-slim

# Set working directory

WORKDIR /app

# Copy files

COPY requirements.txt requirements.txt

COPY app.py app.py

# Install dependencies

RUN pip install --no-cache-dir -r requirements.txt

# Expose API port

EXPOSE 8000

# Run the app

CMD ["uvicorn", "app:app", "--host", "0.0.0.0", "--port", "8000"]

### Step 3: Build and Run the Docker Image

# Build the Docker image

docker build -t ai-sentiment-app .

# Run the container

docker run -p 8000:8000 ai-sentiment-app

* Then open: [*http://localhost:8000/docs*](http://localhost:8000/docs)
* To test the /predict endpoint with JSON:

{

“text": "This product is amazing!"

}

## Project Summary

I hope these projects provide you with a clearer understanding of Object-Oriented Programming in Python and how to apply it in real-world scenarios. Feel free to run each project, examine the source code, and even try to extend them to further hone your skills!