





Pocker is a platform that packages applications into containers.
 Containers include everything needed to run your app − code, libraries, and dependencies − so it works anywhere.

Think of it like:

"A lightweight, portable box for your software."

- Key Benefits:
 - 🕨 Consistency across machines 🖳 🔂 📣
 - Faster deployments #
 - Easy isolation between apps 🔐
 - Smaller than virtual machines
- \bigcirc Example: Run a full Al model in Docker and move it from laptop \rightarrow cloud with no changes!

Why Docker Matters in AI?

Al apps depend on many frameworks — PyTorch, TensorFlow, Transformers, etc. Managing them manually is messy.

Docker simplifies this:

- One container = One environment
- Run Al models, vector DBs, and APIs together
- Easily share AI setups with others

Example:

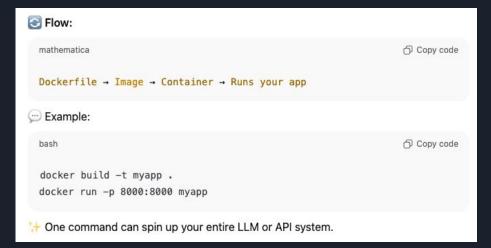
Your entire RAG system (FastAPI + Ollama + Chroma) can live in one Docker Compose file.

Consistency = reliability when training or deploying ML models.

How Docker Works 🃣

E Core Components:

- 1. Dockerfile instructions for building an image 🙊
- 2. Image blueprint of your app 📸
- 3. Container running instance of the image 🔅
- 4. Docker Hub / Registry where images are stored ...



Key Docker Commands You Must Know

- **\sqrt{}** Common commands every engineer uses:
- Pro Tip: Add --rm to auto-remove containers after they stop.
- In Al projects: use these commands to debug your LLM API containers easily.

Command	Purpose	0
docker ps	List running containers	
docker images	List images on system	
docker build -t app .	Build an image	
docker run -d -p 8000:8000 app	Run a container	
docker logs <id></id>	View logs	
docker exec -it <id> bash</id>	Access inside container	

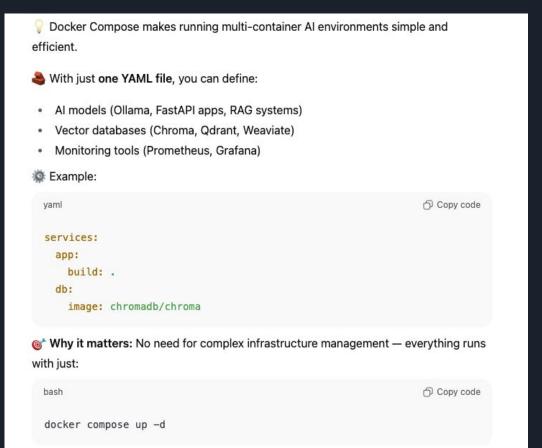
Docker for the Real World

♥ Docker isn't just for demos — it powers production systems across industries.

Real-world uses:

- Run local LLMs (Ollama, Mistral, Gemma)
- Q Host Vector Databases (Chroma, Qdrant)
- Serve APIs (FastAPI, Flask, Django)
- Connect to Observability (Grafana, Prometheus)
- Deploy anywhere AWS, GCP, Azure, or your laptop
- "Docker is the backbone of modern AI development it makes building, testing, and scaling intelligent systems faster and more reliable."

Why Docker Compose for AI Workloads?



Perfect for Rapid Prototyping

- When working with Al projects, speed matters.
- Compose allows you to:
 - Spin up multiple services instantly
 - Test LLMs, APIs, and databases together
 - Rebuild and reset your environment quickly
- Example use cases:
 - Testing new embeddings in Chroma
 - Switching between models like Mistral or Llama
 - Integrating RAG pipelines rapidly
- ✓ In Al labs, time-to-iteration is everything and Docker Compose delivers that.

Lightweight Alternative to Kubernetes

Kubernetes is powerful, but heavy for small AI projects.

Docker Compose gives 80% of the functionality with 20% of the complexity.

Perfect for students, startups, and prototype labs before scaling to Kubernetes.

Kubernetes	Docker Compose
Complex	2 mins
High	Low
Production clusters	Local & small-scale AI workloads
Steep	Easy
Large	Minimal
	Complex High Production clusters Steep

Easy Integration with AI Tools

- **compose integrates easily with AI components like:**

 - Chroma / Qdrant → vector databases for retrieval
 - FastAPI → lightweight AI agent server
 - Grafana + Prometheus → system observability
- Each service is isolated, reproducible, and restartable.
- You can rebuild specific containers without touching others.
- Example: Update your LLM version in seconds while keeping Chroma's data safe.

Real Benefits for AI Engineers

- Why Docker Compose belongs in every Al Engineer's toolkit:
 - 1. Modular Architecture isolate Al agents, databases, and APIs.
- 3. Grant Security Controls can add .env and network isolation.
- 4. Observability pair with Prometheus + Grafana easily.
- 5. Focus on AI, not Infrastructure spend time training and deploying, not debugging Kubernetes YAMLs.
- "Docker Compose bridges the gap between AI experimentation and production deployment."
- At Remoder, we use it to teach engineers how to design real-world AI systems fast, secure, and scalable.

QUESTIONS???

Contact us & come learn with us $9(9_{\circ})^{\circ}$

- https://www.linkedin.com/company/remoder
- https://www.linkedin.com/in/sanjars/
- https://www.youtube.com/@remoder-inc
- remoder.com
- Full walkthrough of this project is available on our "Master Al Deployment " course