



# ICT Training Center



Il tuo partner per la Formazione e la Trasformazione digitale della tua azienda





# SPRING AI

## GENERATIVE ARTIFICIAL INTELLIGENCE CON JAVA

---

Simone Scannapieco

Corso avanzato per Venis S.p.A, Venezia, Italia

Novembre 2025

 **DOCKER MODEL RUNNER**

---

- ➔ Risposta di Docker ad Ollama
  - ➔ LLM in Docker container locali
  - ➔ Modelli AI generici dockerizzabili (WIP)
  - ➔ Docker mette a disposizione una serie di modelli *open source* scaricabili tramite Engine o Desktop
- ➔ Requisiti: <https://www.ajeetraina.com/docker-model-runner-tutorial-and-cheatsheet-mac-windows-and-linux-support/>
  - ↗ <https://docs.docker.com/ai/model-runner/>
  - ↗ <https://www.docker.com/blog/run-llms-locally/>
  - ↗ <https://www.docker.com/blog/introducing-docker-model-runner/>

# DOCKER MODEL RUNNER

## COME SI USA?

- ➔ Tramite Docker Engine
- ➔ Tramite Docker Desktop

### Caricamento LLM in locale

```
docker model pull ai/gemma3
```

### Esecuzione LLM in locale

```
docker model run ai/gemma3
```

### Elenco LLM in locale

```
docker model list
```

### Eliminazione LLM in locale

```
docker model rm ai/gemma3
```

### 1 Verificare le impostazioni relative ad AI

The screenshot shows the Docker Desktop settings interface. The left sidebar has a 'Settings' header and a search bar. The 'AI' section is highlighted with a grey background. The main area shows the 'Docker Model Runner' section with two checkboxes: 'Enable Docker Model Runner' (checked) and 'Enable host-side TCP support' (unchecked). Below it is 'Enable GPU-backed inference' (checked), with a note that additional components will be downloaded to `~/.docker/bin/inference`. At the bottom right are 'Close' and 'Apply' buttons. The status bar at the bottom shows 'Engine running', 'RAM 0.91 GB CPU 0.05%', 'Disk: 1.24 GB used (limit 1006.85 GB)', 'Terminal v4.48.0', and a user icon.

Settings [Give feedback](#)

Search settings

AI

General

Resources

Docker Engine

Builders

**AI**

Kubernetes

Software updates

Extensions

Beta features

Notifications

Enable Docker Model Runner [Give feedback](#)

Enable GPU-accelerated inference engines on `/var/run/docker.sock` and `model-runner.docker.internal:80`. Note: Inference engine support may take a few minutes to initialize.

Enable host-side TCP support

Enable GPU-backed inference

Additional components will be downloaded to `~/.docker/bin/inference`.

Close Apply

Engine running RAM 0.91 GB CPU 0.05% Disk: 1.24 GB used (limit 1006.85 GB)

> Terminal v4.48.0

### 2 Verificare le impostazioni relative alle Beta features

The screenshot shows the Docker Desktop settings interface. The left sidebar has a 'Beta features' option highlighted with a gray background. The main area is titled 'Beta features' and contains the following text: 'Beta features are work in progress and subject to change. Use them to explore and influence upcoming functionality.' Below this, there are three checkboxes:

- Enable Docker AI [Learn more](#) (Enables the "Ask Gordon" feature in Docker Desktop and CLI.)
- Enable Docker MCP Toolkit (Enables the "MCP Toolkit" feature in Docker Desktop and CLI.)
- Enable Wasm, requires the [containerd image store](#) (Installs runtimes that lets you run Wasm workloads.)

At the bottom, it says 'Check back for more features soon, or sign up for our [Developer Preview Program](#)'.

At the bottom right of the window are 'Close' and 'Apply' buttons. The status bar at the bottom shows 'Engine running', 'RAM 0.97 GB CPU 0.00%', 'Disk: 1.24 GB used (limit 1006.85 GB)', 'Terminal v4.48.0', and a user icon with 'Simone Scannapieco'.

# DOCKER MODEL RUNNER

## UTILIZZO DI BASE - DOCKER DESKTOP

### 3 Accedere al pannello Docker Hub della sezione Models

The screenshot shows the Docker Desktop interface. On the left, a sidebar menu includes options like Ask Gordon (BETA), Containers, Images, Volumes, Kubernetes, Builds, Models (which is selected and highlighted in grey), MCP Toolkit (BETA), Docker Hub, Docker Scout, and Extensions. The main area is titled "Models" with a "Docker Hub" tab selected. A search bar at the top says "Search tools". Below it, there's a list of available models:

- granite-docling** [Pull](#)  
Granite Docling is a multimodal model for efficient document conversion.  
Downloads: 3.6K | Stars: 1
- moondream2** [Pull](#)  
An open-source visual language model that interprets images via text prompts, fast and powerful.  
Downloads: 2.0K | Stars: 0
- smolvlm** [Pull](#)  
SmolVLM: lightweight multimodal model for video, image, and text analysis, optimized for devices.  
Downloads: 3.0K | Stars: 0
- granite-4.0-h-small** [Pull](#)  
32B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.  
Downloads: 2.6K | Stars: 0
- granite-4.0-h-tiny** [Pull](#)  
7B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.  
Downloads: 2.6K | Stars: 2
- granite-4.0-h-micro** [Pull](#)  
3B long-context instruct model with RL alignment, IF, tool calling, and enterprise readiness.  
Downloads: 1.6K | Stars: 2
- granite-4.0-micro** [Pull](#)  
3B long-context instruct model with RL alignment, IF, tool use, and enterprise optimization.  
Downloads: 1.2K | Stars: 0
- devstral-small** [Pull](#)  
Agentic coding LLM (24B) fine-tuned from Mistral-Small-3.1 with a 128K context window.  
Downloads: 3.0K | Stars: 1
- magistral-small-** [Pull](#)
- ...** [Pull](#)
- ...** [Pull](#)
- ...** [Pull](#)
- ...** [Pull](#)

Pagination controls at the bottom show pages 1, 2, and >.

At the bottom of the screen, status information includes: Engine running, RAM 0.97 GB, CPU 0.05%, Disk: 1.24 GB used (limit 1006.85 GB), Terminal v4.48.0, and a note that Spring AI - Corso avanzato is running.

# DOCKER MODEL RUNNER

## UTILIZZO DI BASE - DOCKER DESKTOP

### 4 Selezionare il modello ed eventuale versionamento quantizzato

The screenshot shows the Docker Desktop interface with the 'Models' tab selected in the sidebar. A single model entry is visible in the main pane:

Name	ID	Quantization	Parameters	Size	Created	Actions
ai/gemma3:1B-Q4_K_M				0.15 of 0.75 GiB		<a href="#">Edit</a> <a href="#">Delete</a>

Below the table, it says "No rows". At the bottom, status information includes "Engine running", "RAM 0.97 GB CPU 0.00% Disk: 1.24 GB used (limit 1006.85 GB)", "Terminal v4.48.0", and "Showing 1 item".

### 5 Utilizzare il modello da linea di comando integrata

The screenshot shows the Docker Desktop application window. On the left, a sidebar menu includes options like Ask Gordon (BETA), Containers, Images, Volumes, Kubernetes, Builds, Models (which is selected and highlighted in grey), MCP Toolkit (BETA), Docker Hub, Docker Scout, and Extensions. The main pane displays a container named "Gemma3:1B-Q4\_K\_M" with the identifier "dbe29fa0169". Below the container name, there are tabs for Chat, Inspect, and Requests, with "Chat" being the active tab. A sub-header "Chat using ai/gemma3:1B-Q4\_K\_M" is followed by a message from the AI: "Ciao! Sono un modello linguistico di grandi dimensioni, addestrato da Google. In poche parole, sono un programma informatico progettato per comprendere e generare testo." To the right of this message is a blue button labeled "Parlami di te". Below the message, another text block says "Ecco alcuni dettagli su cosa posso fare:" followed by a bulleted list of capabilities. At the bottom of the main pane is a large input field with the placeholder "Ask a question..." and a send button. The status bar at the bottom of the window shows "Engine running", "RAM 0.97 GB CPU 0.05% Disk: 1.24 GB used (limit 1006.85 GB)", "Terminal v4.48.0", and a "Spring AI" icon.

Ciao! Sono un modello linguistico di grandi dimensioni, addestrato da Google. In poche parole, sono un programma informatico progettato per comprendere e generare testo.

Ecco alcuni dettagli su cosa posso fare:

- Genero testo: Posso scrivere diversi tipi di testo, come poesie, codice, script, pezzi musicali, e-mail, lettere, ecc.
- Rispondo a domande: Posso rispondere a domande anche se sono aperte, impegnative o strane. Cerco di fornire risposte informative e complete, anche se le mie risposte sono basate su dati che ho appreso durante l'addestramento.
- Traduco lingue: Posso tradurre tra diverse lingue.
- Seguo istruzioni: Posso seguire le istruzioni che mi vengono date in modo preciso.
- Sono un assistente: Posso aiutarti con compiti di scrittura, brainstorming e apprendimento.

Ask a question... ➤

Engine running || RAM 0.97 GB CPU 0.05% Disk: 1.24 GB used (limit 1006.85 GB) Terminal v4.48.0

### → Come fosse un servizio OpenAI!

#### File pom.xml

```
...
<dependencies>
    <dependency>
        <groupId>org.springframework.boot</groupId>
        <artifactId>spring-boot-starter-web</artifactId>
    </dependency>
    <dependency>
        <groupId>org.springframework.ai</groupId>
        <artifactId>spring-ai-starter-model-openai</artifactId>
    </dependency>
...

```

#### File application.yml

```
spring:
  application:
    name: demo
  ai:
    openai:
      api-key: pippoplutopaperino
      base-url: http://localhost:12434/engines
      chat:
        options:
          model: ai/gemma3
```