



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 has a 'Search settings' bar and a heading 'AI'. Below it, a note says 'Settings for beta AI features can be found [here](#)'. The 'Docker Model Runner' section contains two checkboxes: 'Enable Docker Model Runner' (checked) and 'Enable host-side TCP support' (unchecked). The 'Enable GPU-backed inference' checkbox is checked, with a note below stating '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.

2 Verificare le impostazioni relative alle Beta features

The screenshot shows the Docker Desktop settings interface. The left sidebar has a 'Beta features' option selected, indicated by a grey 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](#) [Legal terms](#)
Enable "Ask Gordon" feature in Docker Desktop and CLI. [Legal terms](#)
- Enable Docker MCP Toolkit
Enable "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 with the 'Models' section selected in the sidebar. The main area displays a grid of AI models from the Docker Hub:

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

At the bottom, there are navigation buttons for pages 1, 2, and more, along with resource usage information (Engine running, RAM, CPU, Disk) and a terminal link.

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		Edit Delete

Below the table, it says "No rows". At the bottom right, it indicates "Showing 1 item". 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 "Spring AI - Corso avanzato" icon.

5 Utilizzare il modello da linea di comando integrata

The screenshot shows the Docker Desktop interface. On the left, a sidebar menu includes 'Ask Gordon' (BETA), 'Containers', 'Images', 'Volumes', 'Kubernetes', and 'Builds'. The 'Models' option is selected and highlighted in grey. Below it are 'MCP Toolkit' (BETA), 'Docker Hub', 'Docker Scout', and 'Extensions'. The main panel displays a container named 'Gemma3:1B-Q4_K_M' (dbe29fa0169). The 'Chat' tab is active, showing the text 'Chat using ai/gemma3:1B-Q4_K_M'. A message from the model says: 'Ciao! Sono un modello linguistico di grandi dimensioni, addestrato da Google. In poche parole, sono un programma informatico progettato per comprendere e generare testo.' Below this, another message says: 'Ecco alcuni dettagli su cosa posso fare:' followed by a bulleted list: '• 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.', and '• Sono un assistente: Posso aiutarti con compiti di scrittura, brainstorming e apprendimento.'. At the bottom, there is a text input field labeled 'Ask a question...' with a send button, and a status bar at the bottom showing 'Engine running', 'RAM 0.97 GB CPU 0.05% Disk: 1.24 GB used (limit 1006.85 GB)', 'Terminal v4.48.0', and a network icon.

→ 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
```