



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/>

- ➔ Tramite Docker Engine
- ➔ Tramite Docker Desktop

Caricamento LLM in locale

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

Esecuzione LLM in locale

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

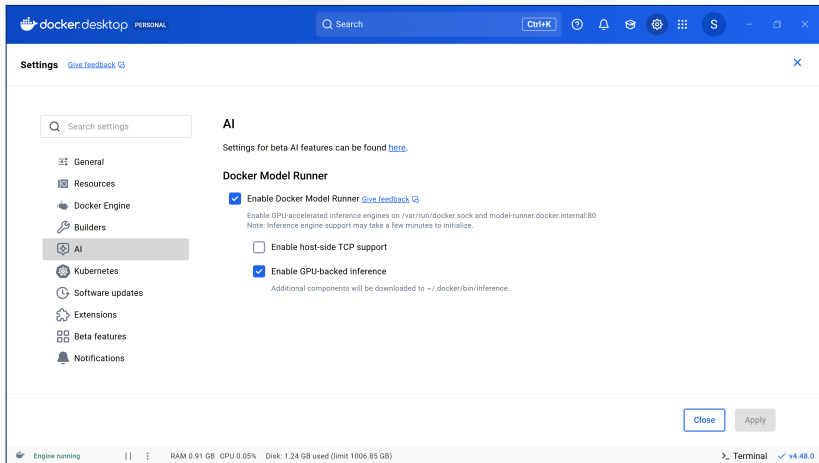
Elenco LLM in locale

```
docker model list
```

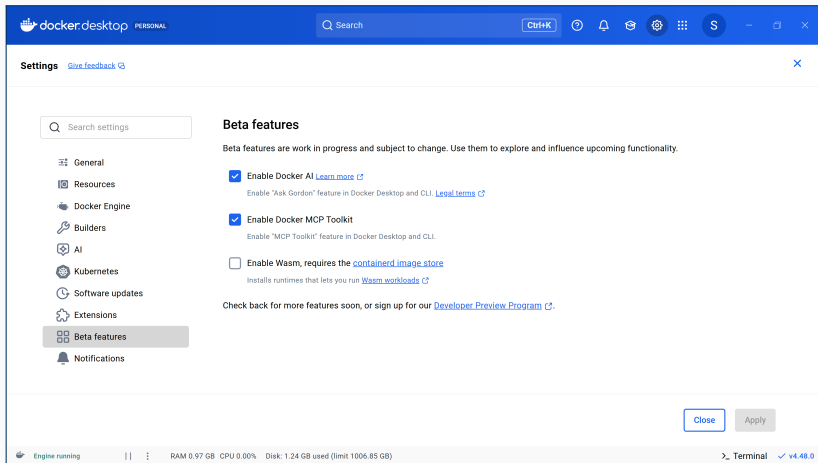
Eliminazione LLM in locale

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

1 Verificare le impostazioni relative ad AI



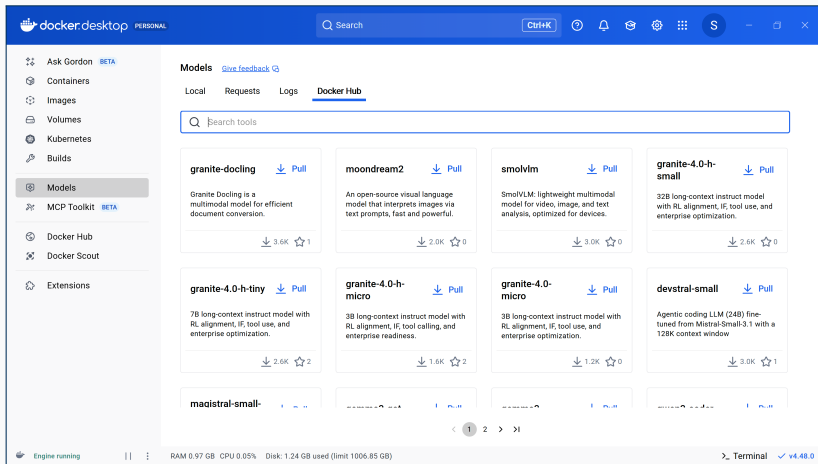
2 Verificare le impostazioni relative alle Beta features



DOCKER MODEL RUNNER

UTILIZZO DI BASE - DOCKER DESKTOP

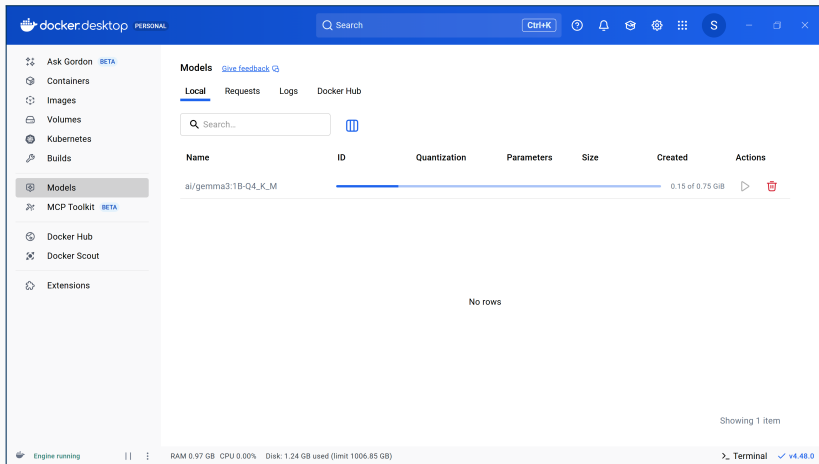
3 Accedere al pannello Docker Hub della sezione Models





The screenshot shows the Docker Desktop application window. The top bar is blue with the Docker logo, 'docker.desktop PERSONAL', a search bar, and a 'Ctrl+K' button. The left sidebar contains a list of navigation items: Ask Gordon (BETA), Containers, Images, Volumes, Kubernetes, Builds, Models (selected), MCP Toolkit (BETA), Docker Hub, Docker Scout, and Extensions. The main content area is titled 'Models' with a 'Give feedback' link. Below the title are tabs for 'Local', 'Requests', 'Logs', and 'Docker Hub' (selected). A search bar labeled 'Search tools' is present. The models are displayed in a grid of cards. Each card shows the model name, a 'Pull' button, a description, and download statistics (downward arrow, size, and star rating). The models shown are: granite-docling (3.6K, 1 star), moondream2 (2.0K, 0 stars), smolvlm (3.0K, 0 stars), granite-4.0-h-small (2.6K, 0 stars), granite-4.0-h-tiny (2.6K, 2 stars), granite-4.0-h-micro (1.6K, 2 stars), granite-4.0-micro (1.2K, 0 stars), devstral-small (3.0K, 1 star), and macistral-small. At the bottom, a status bar shows 'Engine running', system resources (RAM 0.97 GB, CPU 0.05%, Disk 1.24 GB used), and a 'Terminal' button with a version indicator 'v4.48.0'.

Model Name	Download Size	Stars
granite-docling	3.6K	1
moondream2	2.0K	0
smolvlm	3.0K	0
granite-4.0-h-small	2.6K	0
granite-4.0-h-tiny	2.6K	2
granite-4.0-h-micro	1.6K	2
granite-4.0-micro	1.2K	0
devstral-small	3.0K	1
macistral-small	-	-

4 Selezionare il modello ed eventuale versionamento quantizzato



The screenshot shows the Docker Desktop interface with the 'Models' section selected in the left sidebar. The 'Local' tab is active, displaying a table of models. The table has columns: Name, ID, Quantization, Parameters, Size, Created, and Actions. One model is listed: 'ai/gemma3:1B-Q4_K_M'. The 'Quantization' column shows a progress bar. The bottom status bar indicates 'Engine running', 'RAM 0.97 GB', 'CPU 0.00%', and 'Disk: 1.24 GB used (limit 1006.85 GB)'.

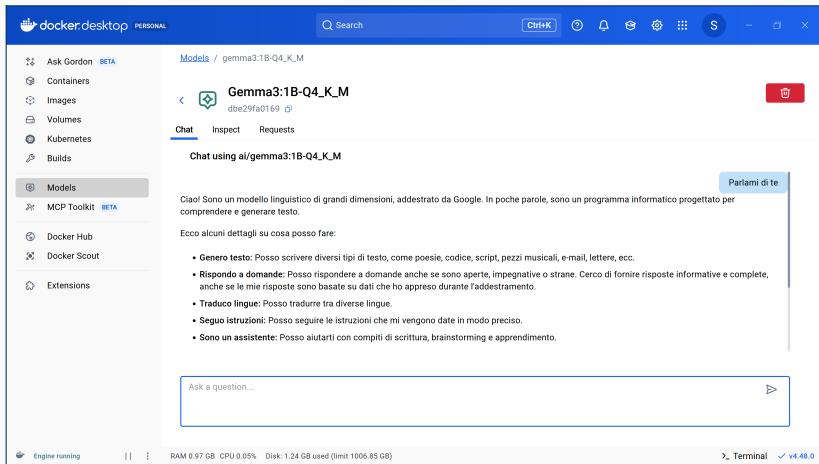
Name	ID	Quantization	Parameters	Size	Created	Actions
ai/gemma3:1B-Q4_K_M		<div></div>		0.15 of 0.75 GiB		 

No rows

Showing 1 item

Engine running | RAM 0.97 GB CPU 0.00% Disk: 1.24 GB used (limit 1006.85 GB) Terminal v4.48.0

5 Utilizzare il modello da linea di comando integrata



The screenshot shows the Docker Desktop application window. On the left is a sidebar with navigation options: Ask Gordon (BETA), Containers, Images, Volumes, Kubernetes, Builds, Models (selected), MCP Toolkit (BETA), Docker Hub, Docker Scout, and Extensions. The main area displays the 'Models' section for 'gemma3:1B-Q4_K_M'. Below this, the 'Gemma3:1B-Q4_K_M' model card is shown with a 'Chat' tab selected. The chat interface displays a message from the model: 'Ciao! Sono un modello linguistico di grandi dimensioni, addestrato da Google. In poche parole, sono un programma informatico progettato per comprendere e generare testo.' followed by a list of capabilities. At the bottom, there is a text input field with the placeholder 'Ask a question...' and a send button. The status bar at the bottom indicates 'Engine running', system resources (RAM 0.97 GB, CPU 0.05%, Disk 1.24 GB used), and a terminal icon.

docker.desktop PERSONAL

Search Ctrl+K

Models / gemma3:1B-Q4_K_M

Gemma3:1B-Q4_K_M
dbe29fa0169

Chat Inspect Requests

Chat using ai/gemma3:1B-Q4_K_M

Parlami di te

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
```