dgroove
Tecnologia emozionale.

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
```

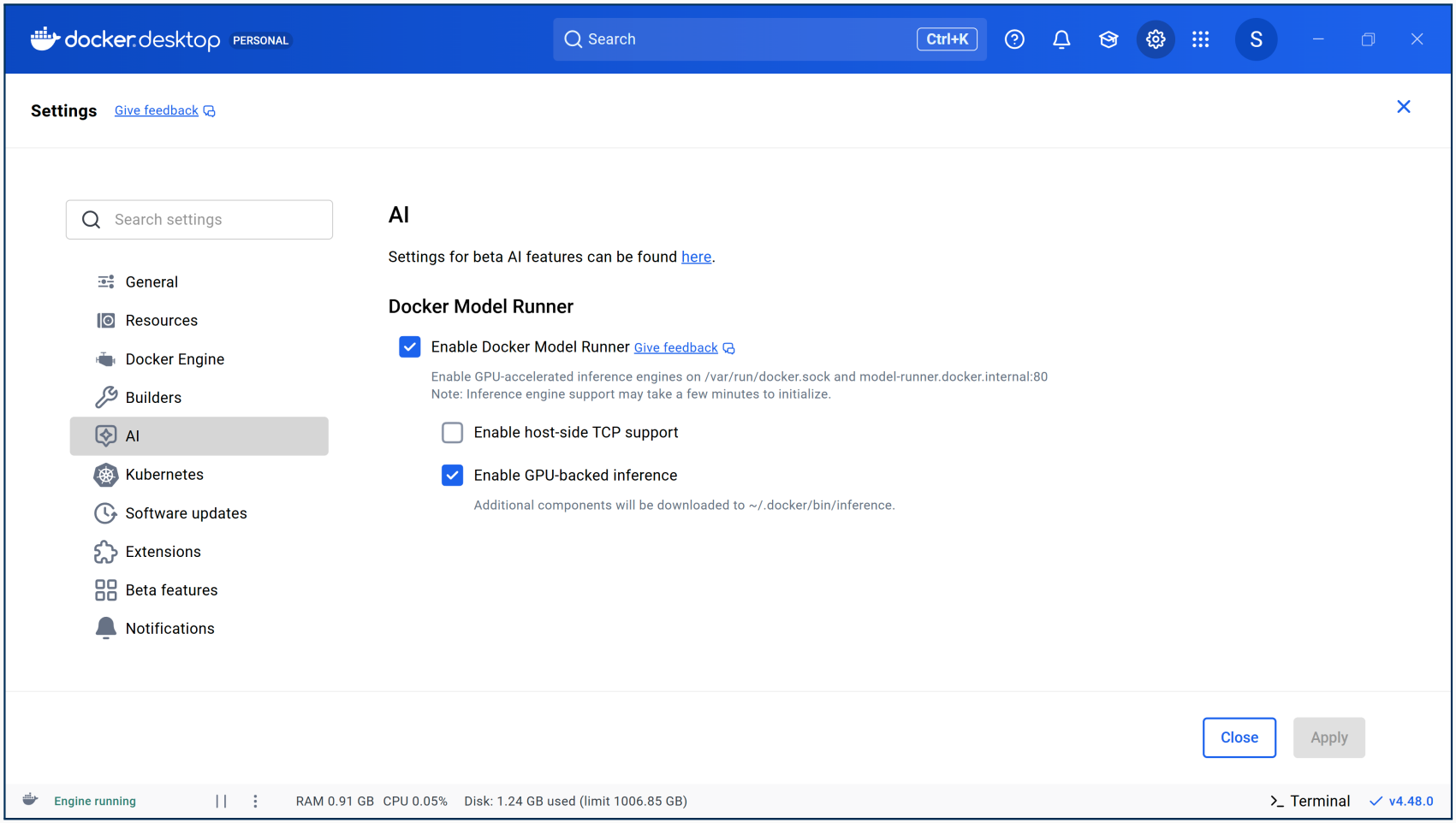
Elenco LLM in locale

```
docker model list
```

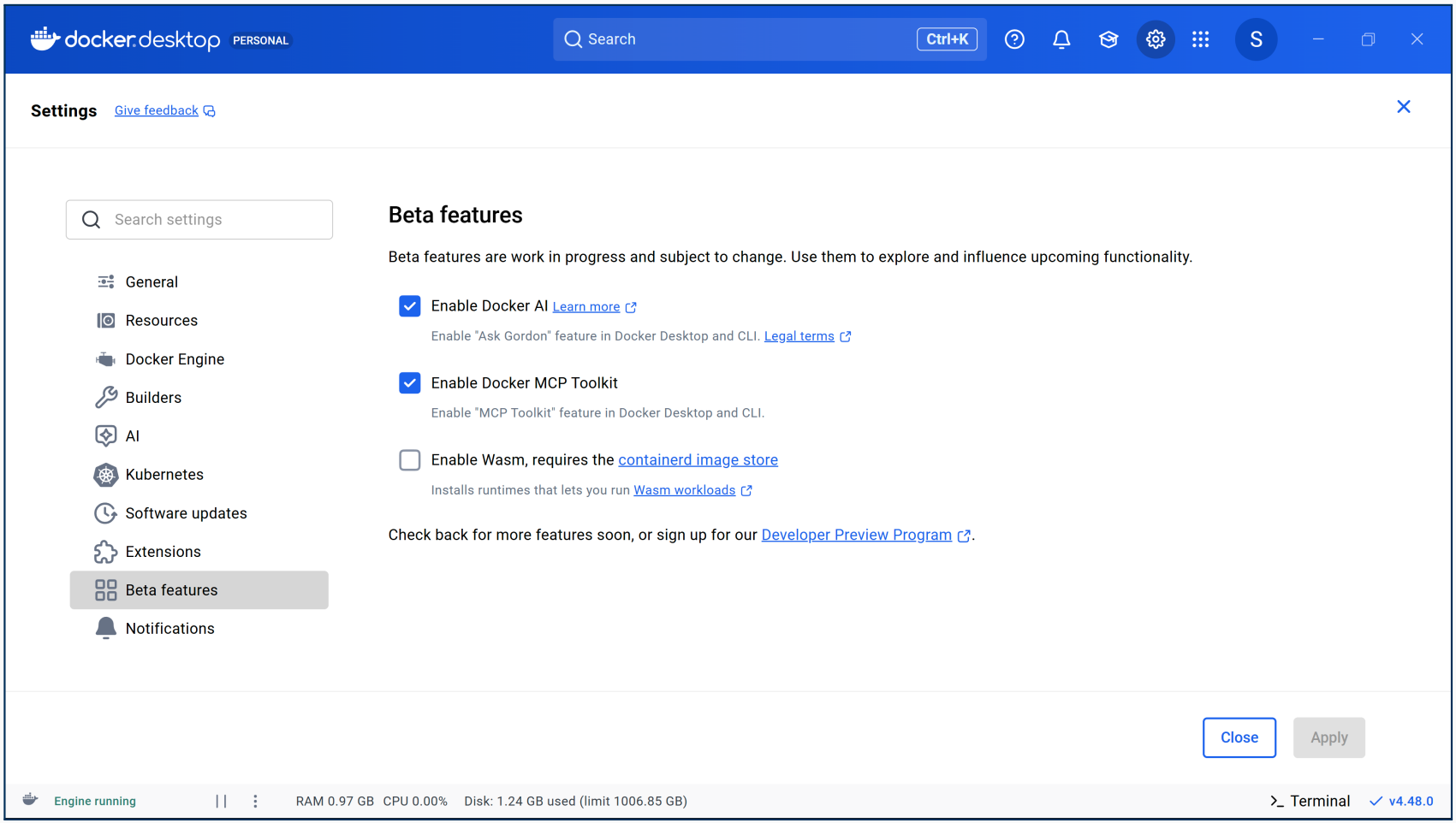
Eliminazione LLM in locale

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

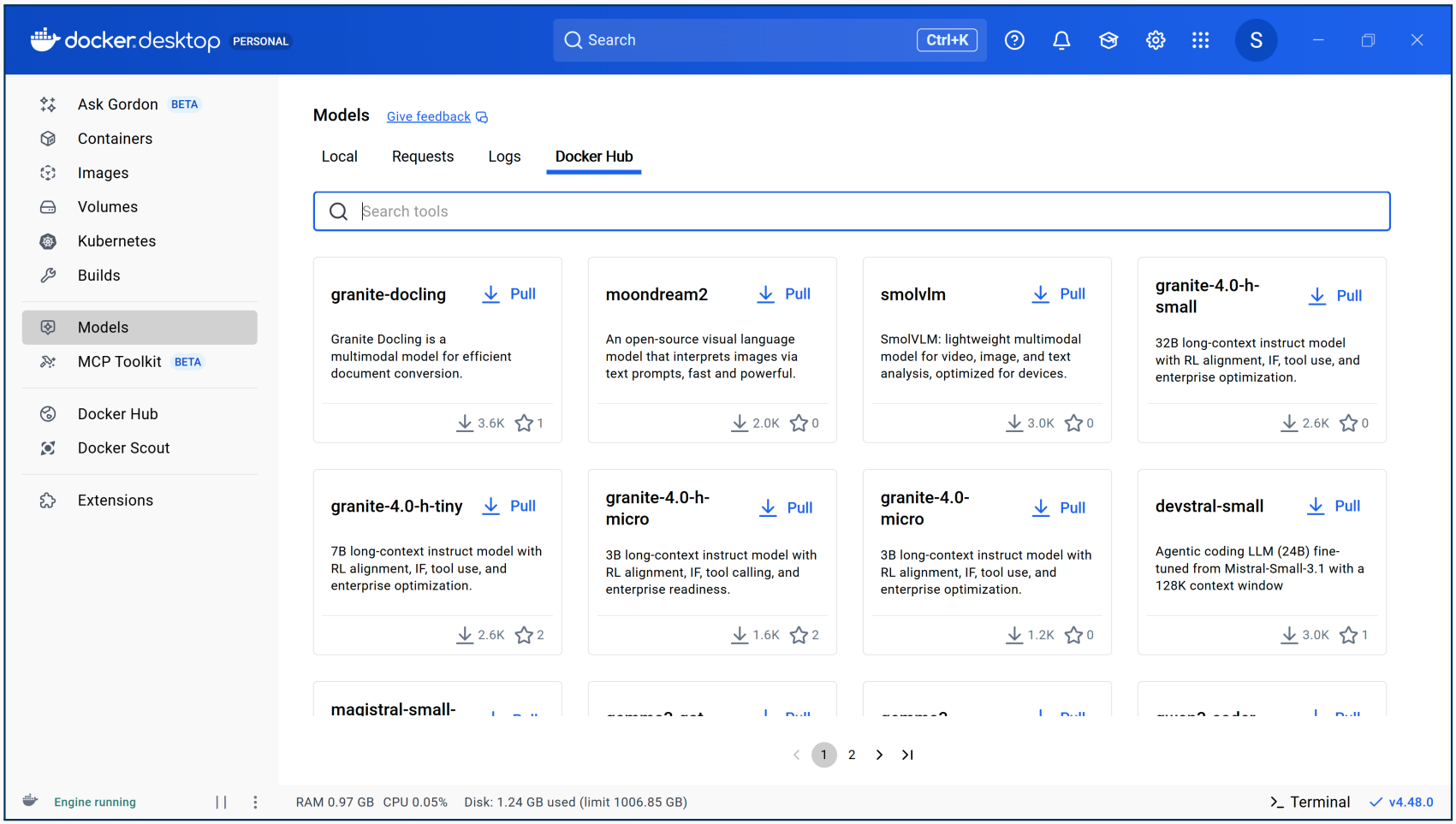
1 Verificare le impostazioni relative ad `AI`



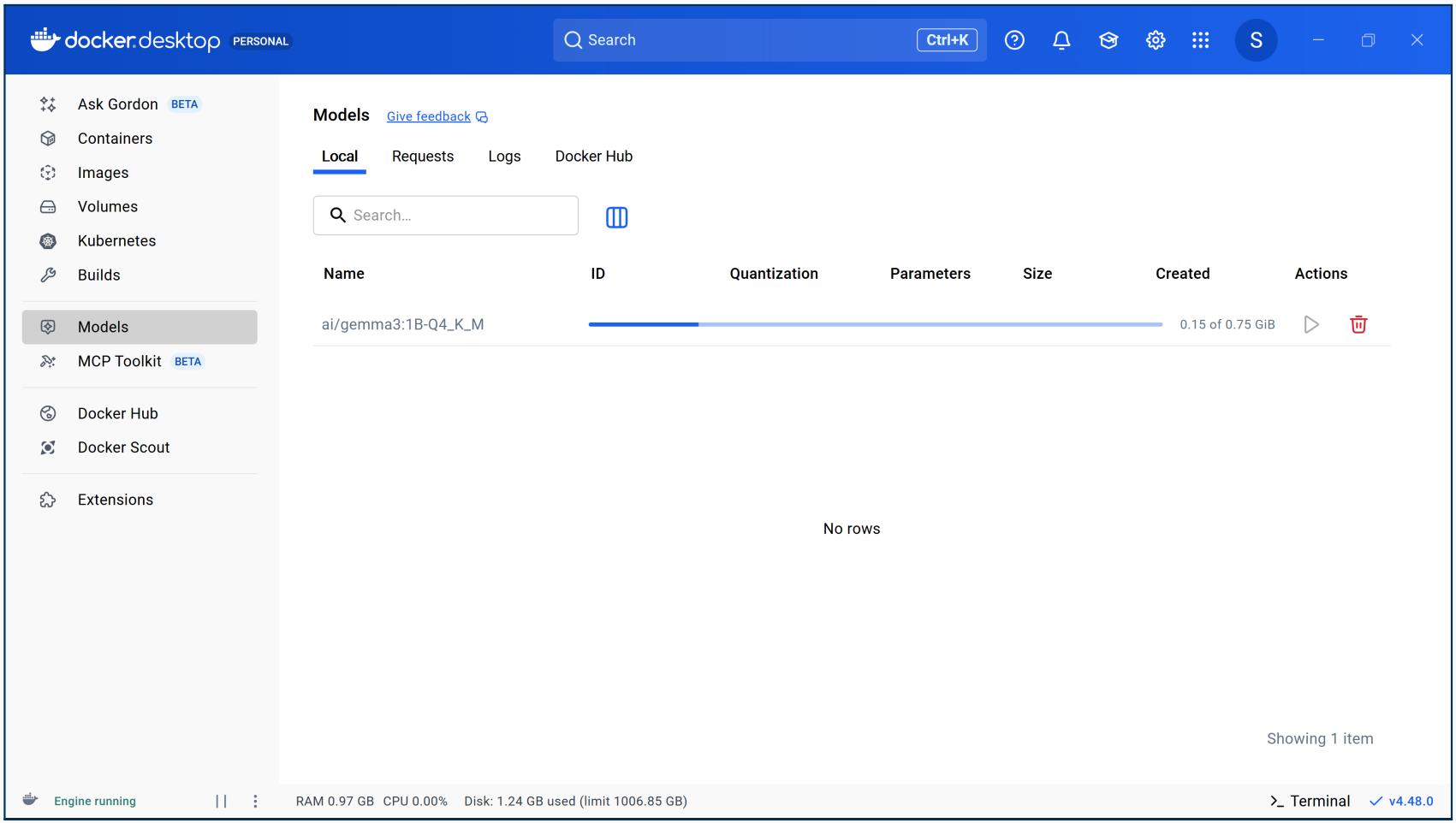
2 Verificare le impostazioni relative alle `Beta features`



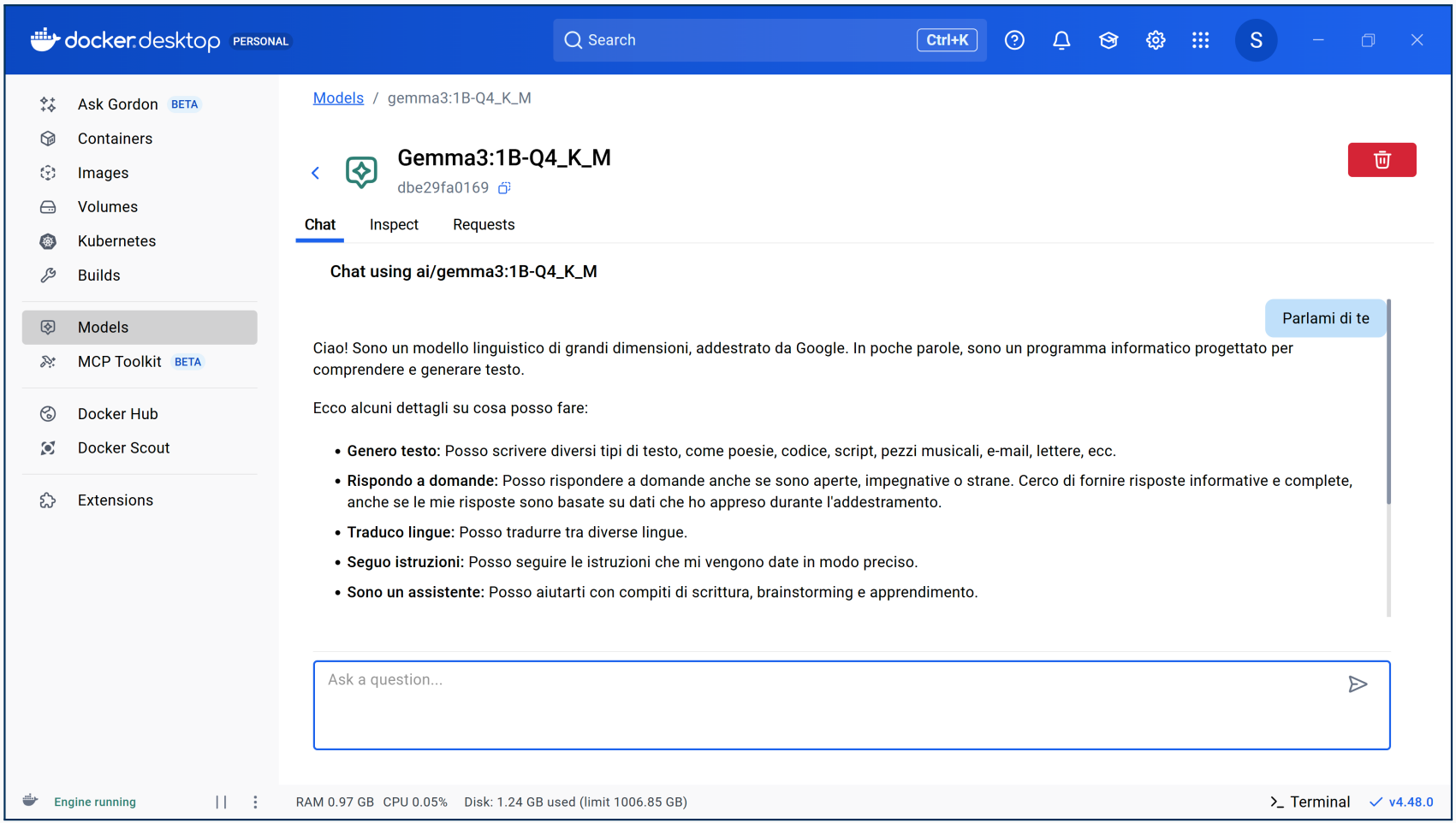
3 Accedere al pannello `Docker Hub` della sezione `Models`



4 Selezionare il modello ed eventuale versionamento quantizzato



5 Utilizzare il modello da linea di comando integrata



- 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
```