



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

 SPRING AI E DOCKER MODEL RUNNER

➔ Multi configurazione Ollama + Gemma

- 1 Pull modello LLM ed embedder in Model Runner
- 2 Test del raggiungimento del servizio *chat/embed* con Postman/Insomnia
- 3 Creazione profilo applicativo
- 4 Modifica proprietà applicativo
- 5 Test delle funzionalità con Postman/Insomnia

 Cercare *embedder* su HuggingFace, con task NLP->Sentence similarity e Library->GGUF

Caricamento LLM in locale

```
docker model pull ai/gemma3
docker model pull hf.co/unsloth/embeddinggemma-300m-GGUF
```

⚠ Utilizzare Postman/Insomnia per chiamate di test

Test raggiungimento servizio *chat*

```
curl http://172.17.0.1:12434/engines/llama.cpp/v1/chat/completions \
-H "Content-Type: application/json" \
-d '{
    "model": "ai/gemma3",
    "messages": [{"role": "user", "content": "Ciao! Come ti chiami?"}]
}'
```

Test raggiungimento servizio *embedding*

```
curl http://172.17.0.1:12434/engines/llama.cpp/v1/embeddings \
-H "Content-Type: application/json" \
-d '{
    "model": "hf.co/unsloth/embeddinggemma-300m-gguf",
    "input": "Questa è una frase di prova"
}'
```

Configurazione applicativo

```
spring:  
...  
profiles:  
    active: rag-document-to-vector-store,docker-model-runner  
...
```

File application-docker-model-runner.yml

```
spring:
  ai:
    openai:
      api-key: pippoplutopaperino
      base-url: http://localhost:12434/engines
      chat:
        completions-path: /llama.cpp/v1/chat/completions
        options:
          model: ai/gemma3
          temperature: 0.2
    embedding:
      embeddings-path: /llama.cpp/v1/embeddings
      options:
        model: hf.co/unsloth/embeddinggemma-300m-gguf # The model name must be taken from
                                                # 'docker model list'
```



<https://github.com/simonescannapieco/spring-ai-advanced-dgroove-venis-code.git>

Branch: 10-spring-ai-gemma-ollama-docker-model-runner