Red Hat Summit

Connect

Potenziare l'Innovazione con LLM Open e GenAI su OpenShift AI

Marco Romani, Cloud Architect
Carmelo Valore, AI Engineer
Open Reply

Roma, 7 novembre 2024

REPLY OPEN | Red Hat

Agenda

- GenAI e Large Language Models
- OpenShift AI
- Demo
- Conclusioni
- Q&A

Red Hat Summit

Connect

GenAI e Large Language Models

REPLY OPEN | Red Hat

GenAI e Large Language Models

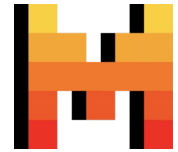
Cosa sono gli LLM?

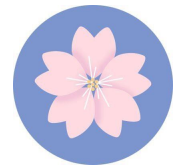
- **Modelli di Intelligenza Artificiale di grande dimensione**, basati su reti neurali di tipo Transformer, in grado di processare e generare contenuti di tipo testuale.
- Sono caratterizzati da **capacità di comprensione e produzione del linguaggio naturale**. Possono comprendere il contesto, rispondere a domande, riassumere testi, tradurre, scrivere testi da zero, ...
- **Accuratezza e prestazioni** crescono con l'aumentare del **dataset di training** e della **potenza computazionale disponibile**.
- Date le loro caratteristiche, sono solitamente messi a disposizione in versioni ***pre-trained***.

Gen AI e Large Language Models

LLM Commerciali vs Open

- **Modelli Commerciali:** soluzioni *ready-to-use* e allo stato dell'arte, generalmente disponibili SaaS con pricing *pay-as-you-go*. Nessun controllo (!).
- **Modelli Open (locali):** soluzioni disponibili per il self-hosting, con controllo quasi totale:
  - **Controllo sull'infrastruttura** se utilizzati in modalità self-hosting.
  - **Pianificazione dei costi** grazie alla possibilità di uscire da logiche PAYG.
  - **Personalizzazione avanzata** grazie alle possibilità di fine-tuning.
  - **Controllo sui dati di training** se il modello è non solo "open", ma "open source".
  - **Continuità** grazie all'indipendenza dalle scelte evolutive del vendor.





REPLY OPEN | Red Hat

GenAI e Large Language Models

Le sfide dei modelli open (locali)

- **Costi di infrastruttura:** per poter fornire prestazioni adeguate a casi d'uso reale, questi modelli richiedono la disponibilità di GPU avanzate:
  - **Nuovi elementi di costo** dell'hardware da considerare, rispetto a infrastrutture tradizionali.
  - Necessario **valutare** con attenzione correttamente il **rapporto costi/benefici**.
  - Questi costi sono **destinati a scendere**, nel tempo, **a parità di caratteristiche del modello**.
- **Costi di operations:** per un'adozione in ambito enterprise e su scala industriale sono richiesti personale preparato e processi adeguati:
  - **MLOps**.

Red Hat Summit

Connect

OpenShift AI

REPLY OPEN | Red Hat

OpenShift AI

Cos'è OpenShift?

- Gestisce il **ciclo di vita** di una **architettura basata sui container**.
- Rappresenta una **piattaforma di sviluppo** per la realizzazione di **applicazioni moderne** e pensate per il **cloud ibrido**.
- Garantisce un'**esperienza di utilizzo consistente** e **indipendente dall'infrastruttura** sottostante, sia essa su public cloud, private cloud o bare metal on-premise.

OpenShift AI

Cos'è?

- **Piattaforma flessibile e scalabile** che abilita la costruzione e l'erogazione di **applicazioni AI-enabled su scala industriale** in ambienti hybrid cloud.
- Basata su OpenShift, fornisce una **toolchain completa e robusta** per lo sviluppo e il deployment di soluzioni basate su AI e ML, offrendo una base solida su cui costituire una **foundation MLOps**.
- Sposta il **focus del team sullo sviluppo di applicazioni che portano valore**, liberandolo dalla gestione dell'infrastruttura e delle risorse.

OpenShift AI

Come ci aiuta?

- **Single-model Serving Runtime:** piattaforma di model serving per modelli di grandi dimensioni basata su KServe.
- Vari runtime per l'esecuzione di modelli: CaiKit-TGIS, OpenVINO, **vLLM**.
- Un ambiente di lavoro per la manipolazione di modelli AI/ML basato su **Jupyter Notebooks**.
- Tutti i classici strumenti di **OpenShift** a support dell'**observability** e del **deployment di componenti applicative** non basate su AI.

...a queste si aggiunge il supporto per le GPU Nvidia offerto da **Node Feature Discovery Operator** e **Nvidia GPU Operator**.

Red Hat Summit

Connect

Demo

REPLY OPEN | Red Hat

Red Hat Summit

Connect

Conclusioni

REPLY OPEN | Red Hat

Conclusioni

Cosa abbiamo visto oggi?

- Cos'è un **Large Language Model** e quali sono i **concetti chiave**.
- Come è possibile **adottare un LLM locale** in un contesto hybrid cloud **grazie ad OpenShift AI**.
- Una **demo dei principali passaggi** necessari **per deployare un'applicazione AI-enabled** basata su LLM locali.
  - Import del modello
  - Configurazione del runtime e deploy del modello
  - Gestione di modelli multipli

Red Hat Summit

Connect

Q&A

REPLY OPEN | Red Hat

Red Hat Summit

Connect

Thank you

REPLY OPEN | Red Hat