LINUX DAY TORINO 25/10/25

OPEN SOURCE LLM E COME SFRUTTARLI

GIULIO SCIARAPPA

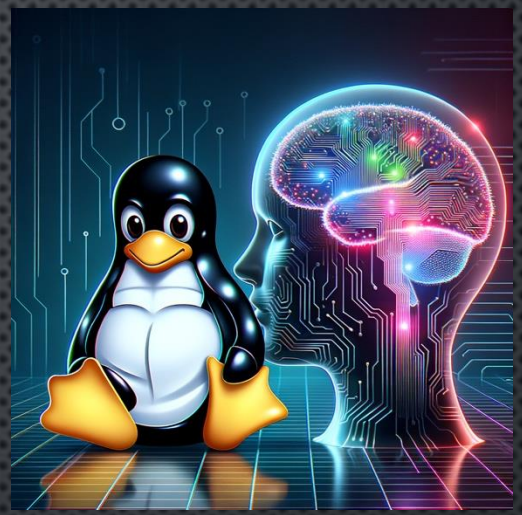
CHI SONO

Giulio Sciarappa

Cloud Enterprise Architect

LINUX
DAY

LLM: COSA SONO?



- LLM = Large Language Model
- Addestrato su miliardi di token testuali con pesi differenti
- Motore di completamento del linguaggio naturale – «prevedono la prossima parola»
- Basato su transformer e self-attention
- Esempi: ChatGPT, Copilot, Claude...

LINUX DAY

LLM FANTASTICI E DOVE TROVARLI

- Miliardi di parametri richiedono calcoli complessi, e geometrici
- Le normali cpu NON sono ottimizzate per questi calcoli
- Si usano quindi le GPU, ottimizzate per calcoli vettoriali/geometrici
- Problema di memoria -> caricare miliardi di parametri, dove? - > RAM (video)

LLM FANTASTICI E DOVE TROVARLI

- Grossi datacenter e fondi di investimento si sono fatti avanti e ospitano i modelli più rinomati
- Pro:
  - Costo accessibile
  - Facilità di utilizzo
- Contro:
  - Quanto scrivi è di loro proprietà
  - Potenzialmente fanno training di nuovi modelli con i dati da te inviati

LINUX DAY

LLM LOCALI, OPEN SOURCE

- Locale != Open source
- Open Source = sono noti i pesi e il codice che li traina, rendendo possibile il fine tuning

- Problema-> ospitarli localmente richiede potenza

LINUX DAY

Perché open source? (e locali…)

- Autonomia: nessuna dipendenza da API o cloud
- Privacy: i dati restano sul tuo PC
- Personalizzazione: puoi modificare e fine-tuning
- Costo zero: nessuna subscription
- Community driven: migliora costantemente

LINUX DAY

PARAMETRI ≠ BIT

PARAMETRI (WEIGHTS)

- SONO I *VALORI NUMERICI* CHE IL MODELLO APPRENDE DURANTE L'ADDESTRAMENTO
- OGNI PARAMETRO RAPPRESENTA UNA CONNESSIONE (PESO) TRA NEURONI
- UN MODELLO DA **7B** = 7 MILIARDI DI *PESI*
- OGNI PESO È UN NUMERO IN *FLOATING POINT* (ES. FP16 O QUANTIZZATO A 4-8 BIT)

BIT

- UNITÀ DI MISURA DELLA MEMORIA (CAPACITÀ DI ARCHIVIAZIONE)
- DETERMINANO *QUANTO SPAZIO* OCCUPANO I PARAMETRI
- ES. 7 MILIARDI DI PARAMETRI × 4 BYTE = **~28 GB DI MEMORIA**

PARAMETRI ≠ BIT

Modello	Parametri	Precisione	Peso su disco
Phi-3 Mini	3B	4bit	~ 1,5GB
Mistral 7B	7B	8 bit	~ 7 GB
LLaMA 3 8B	8B	16 bit	~ 16 GB

PARAMETRI ≠ BIT

PARAMETRI (WEIGHTS)

- SONO I *VALORI NUMERICI* CHE IL MODELLO APPRENDE DURANTE L'ADDESTRAMENTO
- OGNI PARAMETRO RAPPRESENTA UNA CONNESSIONE (PESO) TRA NEURONI
- UN MODELLO DA **7B** = 7 MILIARDI DI *PESI*
- OGNI PESO È UN NUMERO IN *FLOATING POINT* (ES. FP16 O QUANTIZZATO A 4-8 BIT)

BIT

- UNITÀ DI MISURA DELLA MEMORIA (CAPACITÀ DI ARCHIVIAZIONE)
- DETERMINANO *QUANTO SPAZIO* OCCUPANO I PARAMETRI
- ES. 7 MILIARDI DI PARAMETRI × 4 BYTE = **~28 GB DI MEMORIA**

LINUX DAY

DEMO TIME

LINUX
D A Y

Motori di inferenza

- Un motore di inferenza è il software che carica i pesi del modello, gestisce la cache, esegue il forward-pass e restituisce output al prompt.
- Non basta "avere il modello": serve un engine ottimizzato per latenza, throughput, memoria e hardware specifico (CPU/GPU/quantizzazione).
- La scelta del motore può fare la differenza su **velocità** (Time to First Token), **uso memoria**, **scalabilità**.
- In locale, con hardware limitato, sono spesso usati motori leggeri/ottimizzati (es. llama.cpp) che permettono esecuzione su CPU.

OLLAMA

- *Runtime leggero per modelli LLM locali*
- Installazione con un solo comando
- Gestione di modelli (download, cache, run)
- API REST locali compatibili con OpenAI
- Supporto CPU/GPU (llama.cpp)

OLLAMA

- sfrutta quantizzazioni GGUF → versioni ottimizzate per CPU/GPU consumer;
- fornisce un'API REST compatibile OpenAI, ma localmente (porta :11434);
- gestisce cache, modelli e memoria tramite un livello astratto (non devi compilare nulla tu);
- su macOS sfrutta Metal / MPS, su Linux e Windows CUDA o ROCm (se disponibili).

OLLAMA

Installazione su Linux

curl -fsSL https://ollama.com/install.sh | sh

Su Windows e Mac con package

https://ollama.com/download

NB: I modelli vengono scaricati in ~/.ollama/models

OLLAMA

- OLLAMA RUN LLM

DEMO TIME

LINUX
D A Y

ALTERNATIVE OSS

GUI

- **Open WebUI** → interfaccia web per Ollama (pip install open-webui)
- **LM Studio** → alternativa cross-platform

IDE

- Plugin VS Code "Continue"
- Chat AI locale via API

Data Lab

- Notebooks con langchain, llama-index, transformers

DEMO TIME

LINUX
D A Y

Ottimizzare le performance

- Quantizzazione (Q4_0, Q5_K_M, ecc.)
- GPU vs CPU fallback
- Cache dei token
- Riduzione context window per prompt lunghi
- Evitare output streaming se non necessario

SICUREZZA E PRIVACY

- Nessun dato inviato al cloud
- Gestione log locale
- Modelli verificabili e auditabili
- Backup di modelli e pesi su storage personale

DOVE TROVARE I MODELLI

https://ollama.com/library

https://huggingface.co/models

https://github.com/open-webui/open-webui

RIEPILOGO

- Capito cosa sono gli LLM
- Visti i principali modelli open source
- Installato Ollama
- Eseguito un modello locale
- Creato il tuo primo assistente privato

DOMANDE?

LINUX
DAY

GRAZIE!

LINUX
D A Y