

LINUX DAY TORINO

24 OTTOBRE 2026



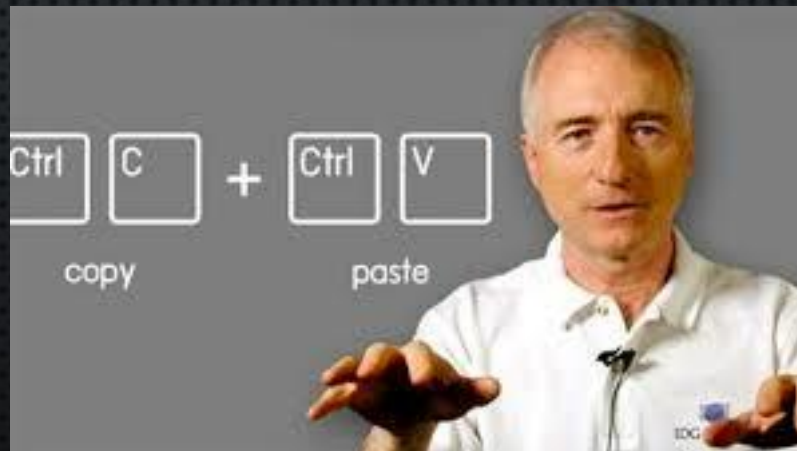
AI PER TUTTI: DAGLI AGENTI AI MODELLI OPEN SOURCE

GIULIO SCARAPPA · LINUX DAY TORINO 2026



"INTELLIGENCE IS WHATEVER
MACHINES HAVEN'T DONE
YET."

LARRY TESLER



CHI SONO

GIULIO SCIARAPPA

ENTERPRISE & CLOUD ARCHITECT

MICROSOFT MVP

TORINO .NET USER GROUP



PRIMA DI INIZIARE: ALZATA DI MANO

1

Hai mai usato
ChatGPT, Copilot,
Gemini o Claude?

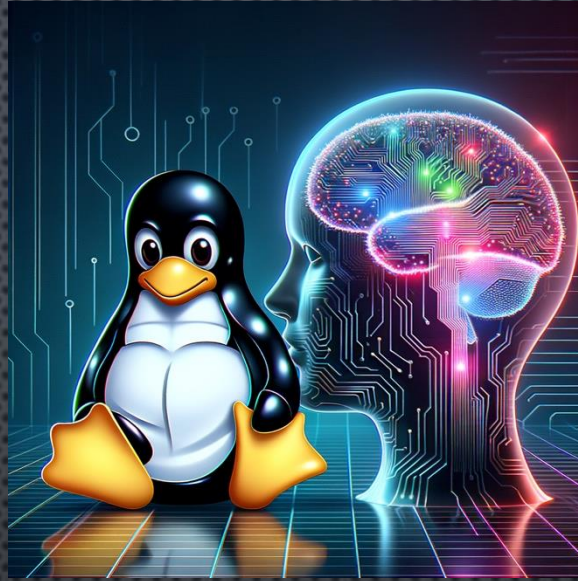
2

Lo usi anche per
lavoro o per studio?

3

Hai mai lasciato
lavorare un agente
da solo sul tuo PC?





L'AI SPIEGATA SEMPLICE

PARTE 1



UN LLM COMPLETA LE FRASI

Domenica vado a pranzo
dalla...

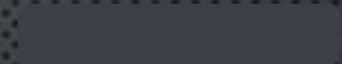


nonna



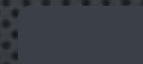
48%

mamma



27%

zia



11%

pizzeria



6%

Probabilità di esempio

Come il T9 del telefono, ma addestrato su miliardi di pagine di testo.
Sceglie una parola, la aggiunge e ricomincia. Così nasce una risposta intera.



TOKEN E CONTESTO

Token

Il modello legge pezzi di parola, non lettere né parole intere.

Linux

Day

Tor

ino

20

26

La divisione qui sopra è indicativa: ogni modello ha la sua.

Si paga e si misura tutto in token: 1.000 token sono più o meno 700 parole.

Contesto

La memoria di lavoro: tutto quello che il modello vede adesso.

8K

un capitolo

128K

un libro

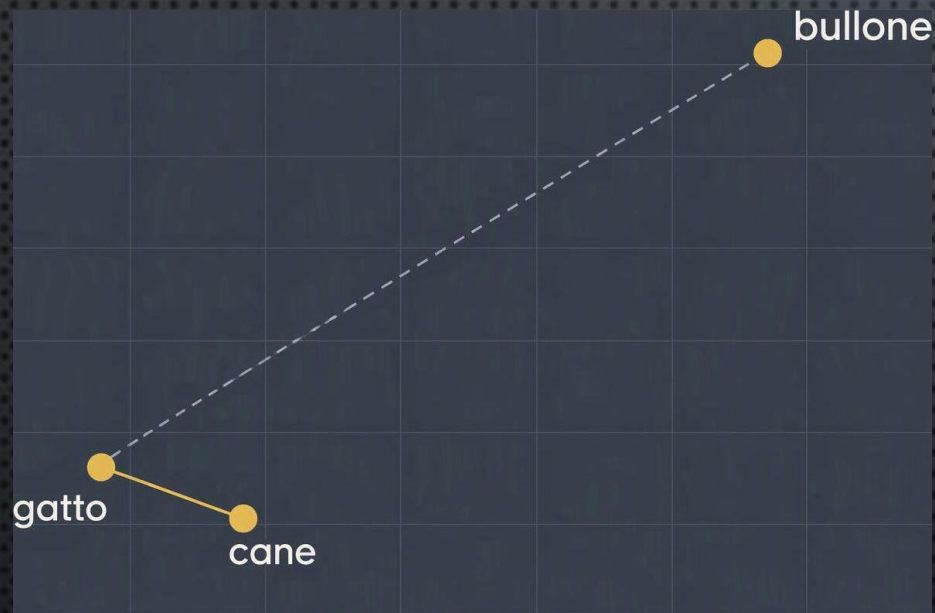
1M

Quando è pieno, le cose più vecchie escono. Tra una chat e l'altra non ricorda nulla.



LA DISTANZA TRA DUE PAROLE

Il modello trasforma le parole in numeri.
Quei numeri sono come coordinate su una mappa.



gatto e cane

VICINI

Compaiono in contesti simili:
per esempio, gli animali di casa.

gatto e bullone

PIÙ LONTANI

Compaiono in contesti diversi:
animali e meccanica.

Misurare la distanza aiuta a capire quanto due parole
si assomigliano nell'uso e nel significato.

Esempio semplificato: la posizione dipende dal modello e dal contesto.



PERCHÉ A VOLTE INVENTA

Allucinazione

Una risposta scritta bene, convincente, ma falsa.

Il modello sceglie le parole più probabili. Non controlla i fatti e non sa di non sapere.

1

Dagli i documenti

Incolla o allega il testo su cui deve lavorare

2

Chiedi le fonti

E aprile, prima di fidarti

3

Fallo cercare

Ricerca web o strumenti per i fatti recenti

4

Calcoli con il codice

Per numeri precisi, fagli scrivere ed eseguire codice



COSA FA BENE, COSA FA MENO BENE

Fa bene

- Riassumere e riscrivere testi
- Tradurre e cambiare tono
- Spiegare codice, errori e concetti
- Fare brainstorming e prime bozze
- Estrarre dati da documenti disordinati

Serve aiuto

- Fatti recenti senza ricerca web
- Calcoli esatti senza strumenti
- Citazioni e numeri a memoria
- Decidere al posto tuo senza contesto
- Sapere cose che non gli hai detto



NON TUTTI I MODELLI SONO UGUALI

Dimensione

piccolo  grande

Pochi miliardi di parametri girano sul portatile. Centinaia di miliardi richiedono un datacenter.

Ragionamento

veloce  pensante

I modelli thinking ragionano prima di rispondere: meglio su logica e codice, ma più lenti.

Input

solo testo  multimodale

I multimodali leggono anche immagini, audio, PDF e schermate.

Accesso

chiuso  aperto

Chiusi: GPT, Claude, Gemini via app o API.
Aperti: scarichi i pesi, come Gemma, Qwen, gpt-oss.





DALLA CHAT AGLI AGENTI

PARTE 2



CHATBOT E AGENTE

Chatbot

Tu chiedi, lui risponde.

Il lavoro lo fai tu: copi, incolli,
provi, torni a chiedere.

Vede solo quello che gli scrivi.

Agente: lavora in un ciclo



Si ferma a obiettivo raggiunto, o quando serve il tuo ok.



AGENTS.MD: IL README PER L'AI

```
# AGENTS.md

## Progetto
Sito del Linux Day, Hugo + Markdown

## Comandi
- Anteprima: hugo server
- Build: hugo --minify

## Regole
- Testi in italiano
- Non modificare /static/loghi
- Chiedi prima di cancellare file
```

Cos'è

Un file Markdown nella cartella del progetto

Quando

L'agente lo legge a inizio sessione

Cosa ci metti

Comandi, convenzioni, cose da non toccare

Chi lo usa

Oltre 60.000 progetti, ora alla Linux Foundation



SKILL: COMPETENZE SU RICHIESTA

```
slide-linux-day/  
  SKILL.md  
  scripts/genera.py  
  assets/template.pptx
```

```
---  
name: slide-linux-day  
description: crea slide con  
  il template del Linux Day  
---  
1. Apri assets/template.pptx  
2. Lancia scripts/genera.py
```

1

Scoperta

All'avvio legge solo nome e descrizione di ogni skill

2

Attivazione

Se il compito corrisponde, carica le istruzioni complete

3

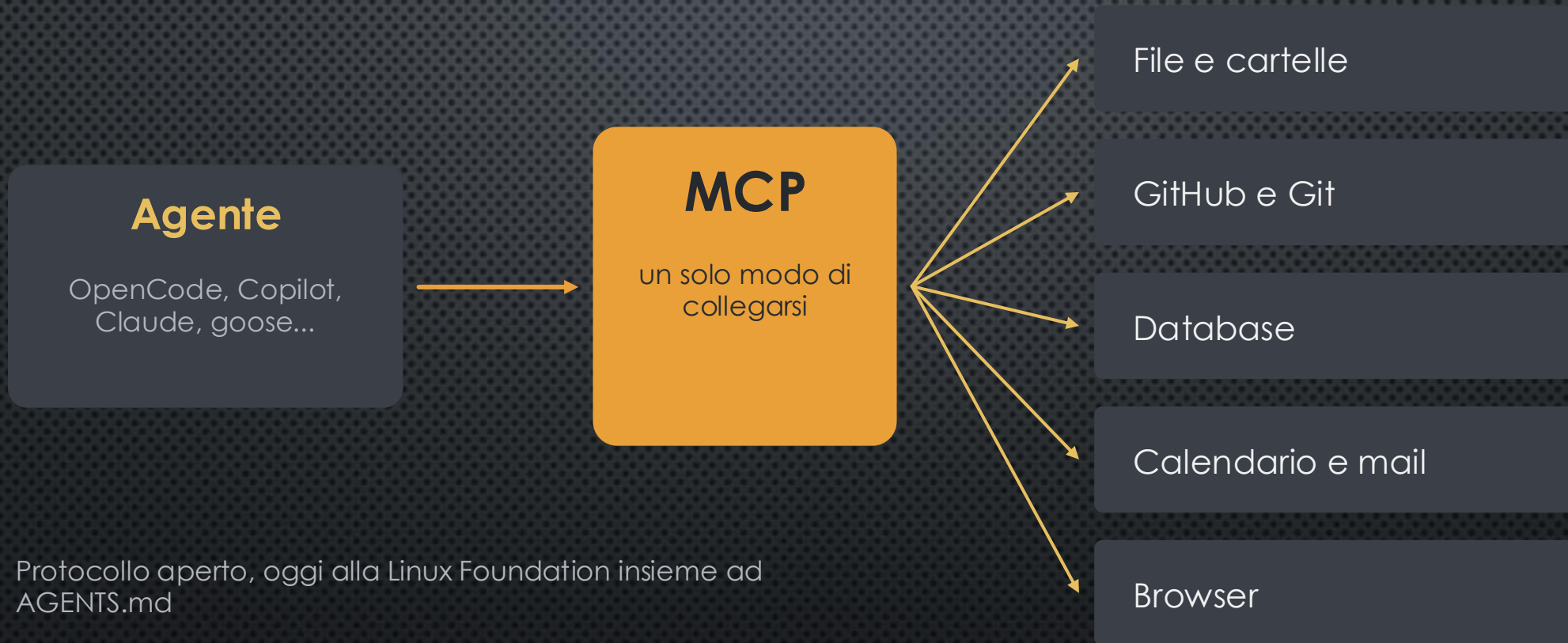
Esecuzione

Segue le istruzioni e usa script e file della cartella

Standard aperto: agentskills.io



MCP: LA PRESA USB DEGLI AGENTI



METTIAMO TUTTO INSIEME

Agente

chi coordina e gira il ciclo ·
OpenCode, Copilot, Codex, goose

Strumenti e MCP

le mani · file, terminale, web, app

Skill

i manuali · SKILL.md caricati quando servono

AGENTS.md

le regole del posto · come si lavora in questo
progetto

Modello

il cervello · in cloud oppure sul tuo PC

*Il cervello si può
cambiare*





MODELLI OPEN IN LOCALE

PARTE 3



OPEN, MA QUANTO?

Locale ≠ open source

Chiuso

Lo usi solo via app o API.
Non vedi e non scarichi nulla.

GPT, Claude, Gemini

Open weight

Scarichi i pesi e li usi dove vuoi. La licenza può avere limiti.

Gemma, Qwen, gpt-oss, Llama

Open source

Pesi, codice di training e informazioni sui dati.
Definizione OSI 2024.

OLMo, alcuni modelli di ricerca



PERCHÉ FARLI GIRARE IN LOCALE

1

Privacy

I dati restano sul tuo computer

2

Costo

Nessun abbonamento, nessun costo a token

3

Autonomia

Funziona offline, anche senza rete

4

Controllo

Scegli tu versione e aggiornamenti

I limiti

I modelli che stanno su un portatile sono più piccoli dei migliori in cloud. Serve memoria: RAM o VRAM.



PARAMETRI E QUANTIZZAZIONE

Parametri: i numeri che il modello ha imparato. 8B = 8 miliardi.

Quantizzazione: salvare quei numeri con meno bit.

Come un MP3: un po' meno qualità, molto meno spazio.

Regola veloce: $\text{parametri} \times \text{bit} \div 8 = \text{GB}$, più un margine.

Un modello da 8 miliardi di parametri

16 bit

originale

~16 GB

8 bit

Q8

~8,5 GB

4 bit

Q4, il più usato

~5 GB

Valori indicativi, solo i pesi del modello



QUANTA MEMORIA SERVE

Memoria*	Esempi su Ollama	Per cosa
8 GB	<code>qwen3.5:4b</code> , <code>llama3.2:3b</code>	chat, riassunti, traduzioni
16 GB	<code>qwen3.5:9b</code> , <code>gemma4:12b</code> , <code>gemma4:e4b</code>	assistente quotidiano, codice semplice
24-32 GB	<code>qwen3.8:27b</code> , <code>gemma4:26b</code> <code>gpt-oss:20b</code> , <code>mistral-small13.2:24b</code>	agenti di coding, immagini
96 GB e più	<code>gpt-oss:120b</code> · <code>qwen3.5:122b</code>	ragionamento e codice complesso

*Fasce indicative, modelli quantizzati e contesto breve. RAM unificata (Mac) o VRAM (GPU).
Serve margine per sistema e cache. Per contesti lunghi occorre più memoria.
Kimi-VL 16B e Kimi Linear 48B: runtime e memoria da verificare per la versione scelta.
Kimi K3: su Ollama via cloud (`kimi-k3:cloud`), senza caricare i pesi in memoria locale.



I MODELLI APERTI DEL 2026

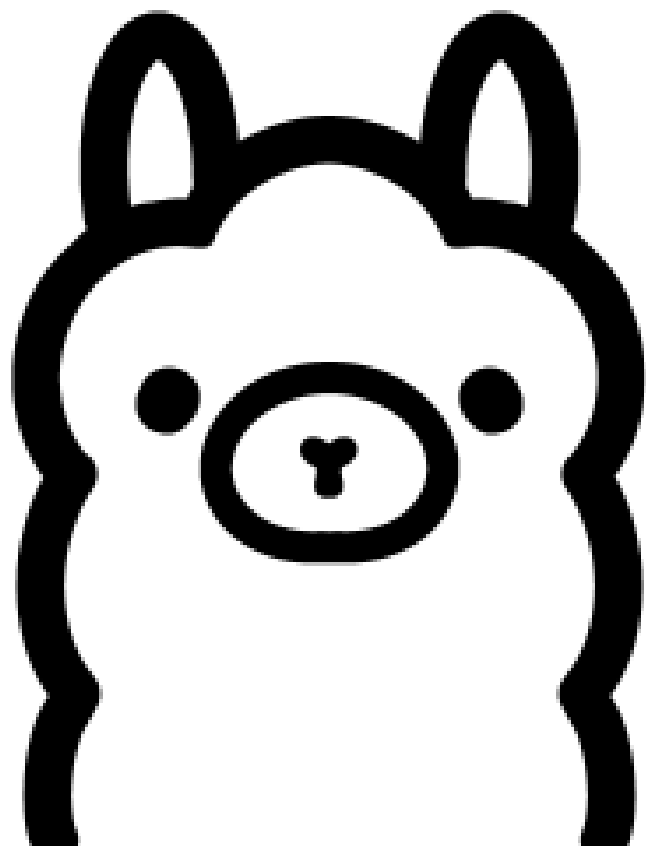
Famiglia	Di chi	Taglie	Punti di forza
Gemma 4	Google	e2b → 31b	multimodale, anche per telefoni
Qwen 3.5 / 3.6 / 3.8	Alibaba	0.8B-2.4T	3.8: coding e agenti, anche 27B
gpt-oss	OpenAI	20b, 120b	ragionamento e strumenti; Apache 2.0
Mistral Small	Mistral AI	24b	l'opzione europea; Apache 2.0
Granite 4.1	IBM	3b → 30b	compatto, pensato per le aziende
DeepSeek, GLM	vari	centinaia di B	vicini ai top, ma da datacenter
Kimi VL / Linear / K3	Moonshot AI	16B / 48B 2.8T	VL: immagini, Linear: contesti lunghi K3: agenti multimodali, cloud/server





DEMO TIME





OLLAMA

- 🐧 SCARICA E AVVIA MODELLI CON UN COMANDO
- 🐧 API LOCALE SU :11434, COMPATIBILE OPENAI E ANTHROPIC
- 🐧 SU MAC USA MLX, SU LINUX CUDA O ROCm
- 🐧 OLLAMA LAUNCH COLLEGA UN AGENTE A UN MODELLO

I COMANDI DELLA DEMO

```
# installazione su Linux (Mac e Windows: ollama.com/download)
curl -fsSL https://ollama.com/install.sh | sh

# scaricare e usare un modello
ollama run gemma4:e4b
ollama list          # modelli scaricati
ollama ps            # modelli caricati in memoria

# la stessa cosa via API, da qualsiasi programma
curl localhost:11434/api/chat -d '{"model":"gemma4:e4b",
  "messages":[{"role":"user","content":"Ciao!"}]}'

# un agente di coding con un modello locale
ollama launch opencode
```



COSA PROViamo

1

Chat con un modello piccolo

gemma4:e4b, risposta quasi istantanea

2

Stessa domanda, modello più grande

confronto tra velocità e qualità

3

Gli mostriamo un'immagine

un modello multimodale descrive una foto

4

Lo chiamiamo via API

curl sulla porta 11434, come farebbe un'app

5

Un agente tutto offline

ollama launch opencode, legge AGENTS.md e lavora sul progetto



NON SOLO OLLAMA

Open WebUI

Interfaccia web stile ChatGPT sopra Ollama

LM Studio

App desktop per scaricare e provare modelli

llama.cpp

Il motore sotto molti di questi strumenti

vLLM

Server per servire modelli a tanti utenti

OpenCode · goose

Agenti open source, anche con modelli locali

Continue

Assistente AI dentro VS Code



PER APPROFONDIRE

Modelli

`ollama.com/library`

`huggingface.co/models`

Agenti

`agents.md`

`agentskills.io`

`modelcontextprotocol.io`

Strumenti

`github.com/open-webui/open-webui`

`opencode.ai`

`lmstudio.ai`



RIEPILOGO

- 1 Un LLM completa testo: utile, ma va verificato
- 2 Un agente usa strumenti e lavora in un ciclo
- 3 AGENTS.md, skill e MCP sono standard aperti
- 4 Open weight e open source non sono la stessa cosa
- 5 Con Ollama un modello locale si prova in 5 minuti





DOMANDE?





GRAZIE!

