SECURITY SUMMIT

Security Summit

Milano 11-12-13 marzo 2025

Clusit 25 2000-2025

ASTREA

AI On-Premise: implementazioni Pratiche per la Sicurezza Aziendale

Alessandro Vallega | Fondatore e Senior partner, *Rexilience.eu*

Roberto Piazzolla | Software Developer/Webmaster, *Clusit*

DI COSA PARLEREMO

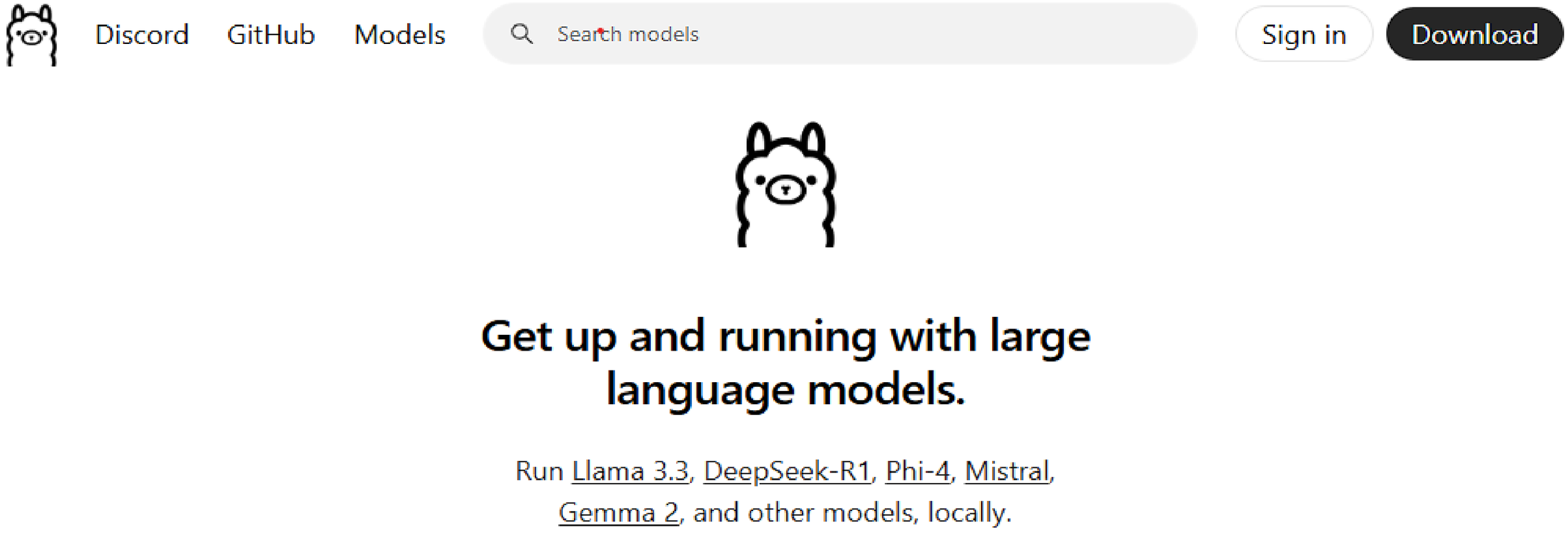
- Perché usare le AI in locale
- Come installare le AI in locale
- Come usarle per programmare
- Considerazioni finali

PERCHÉ USARE LE AI IN LOCALE

- Sicurezza
- Confidenzialità
- Continuità di utilizzo

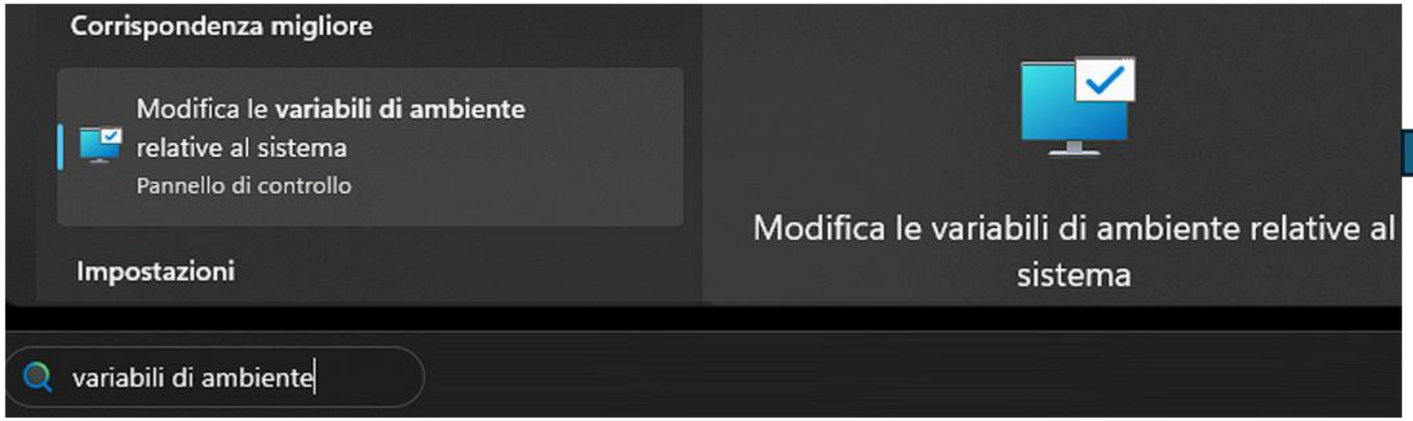
CHE SOFTWARE SERVE?

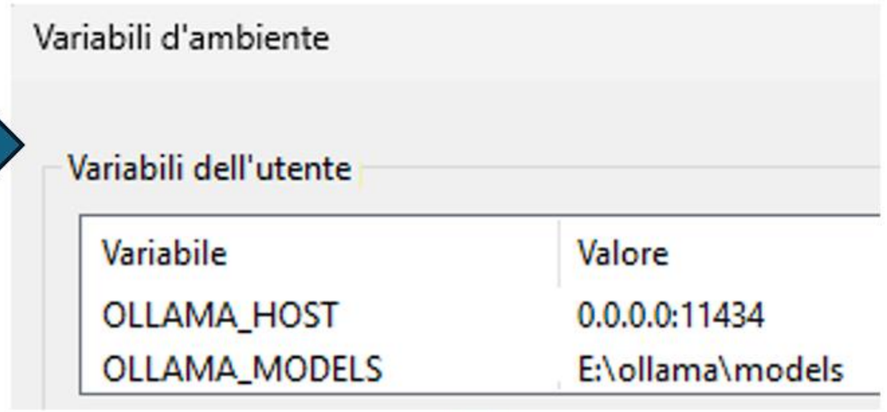
Ollama : un server open-source per le AI in locale

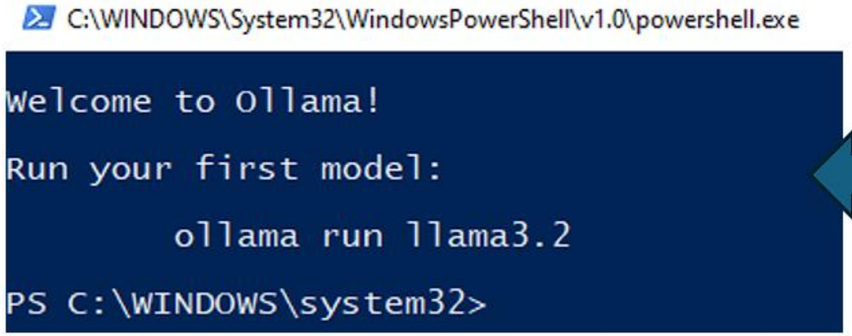


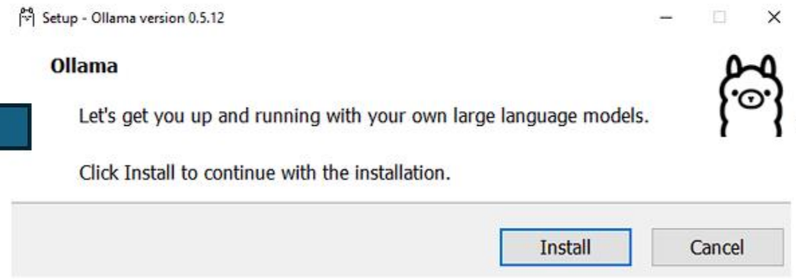
INSTALLARE OLLAMA

Iniziare settando le variabili di ambiente









https://ollama.com/download

I COMANDI DI OLLAMA

C:\WINDOWS\System32\WindowsPowerShell\v1.0\powershell.exe

```
PS C:\WINDOWS\system32> ollama
Usage:
  ollama [flags]
  ollama [command]

Available Commands:
  serve       Start ollama
  create      Create a model from a Modelfile
  show        Show information for a model
  run         Run a model
  stop        Stop a running model
  pull        Pull a model from a registry
  push        Push a model to a registry
  list        List models
  ps          List running models
  cp          Copy a model
  rm          Remove a model
  help        Help about any command

Flags:
  -h, --help      help for ollama
  -v, --version   Show version information

Use "ollama [command] --help" for more information about a command.
```

```
PS C:\WINDOWS\system32> ollama list
NAME    ID    SIZE    MODIFIED
```

I MODELLI AI

https://ollama.com/search

Search models

All | Embedding | Vision | Tools | Popular

deepseek-r1

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

1.5b 7b 8b 14b 32b 70b 671b

20.3M Pulls 29 Tags Updated 2 weeks ago

llama3.3

New state of the art 70B model. Llama 3.3 70B offers similar performance compared to the Llama 3.1 405B model.

tools 70b

1.4M Pulls 14 Tags Updated 2 months ago

deepseek-r1

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

1.5b 7b 8b 14b 32b 70b 671b

20.3M Pulls Updated 2 weeks ago

7b | 29 Tags | `ollama run deepseek-r1`

Updated 5 weeks ago		0a8c26691023 · 4.7GB
model	arch **qwen2** · parameters **7.62B** · quantization **Q4_K_M**	4.7GB
params	{ "stop": ["<｜begin_of_sentence｜>", "<｜end_of_sentence｜>",	148B
template	{{- if .System }}{{ .System }}{{ end }} {{- range $i, $_ := .…	387B
license	MIT License Copyright (c) 2023 DeepSeek Permission is hereby …	1.1kB

LA STRUTTURA DI UN MODELLO AI

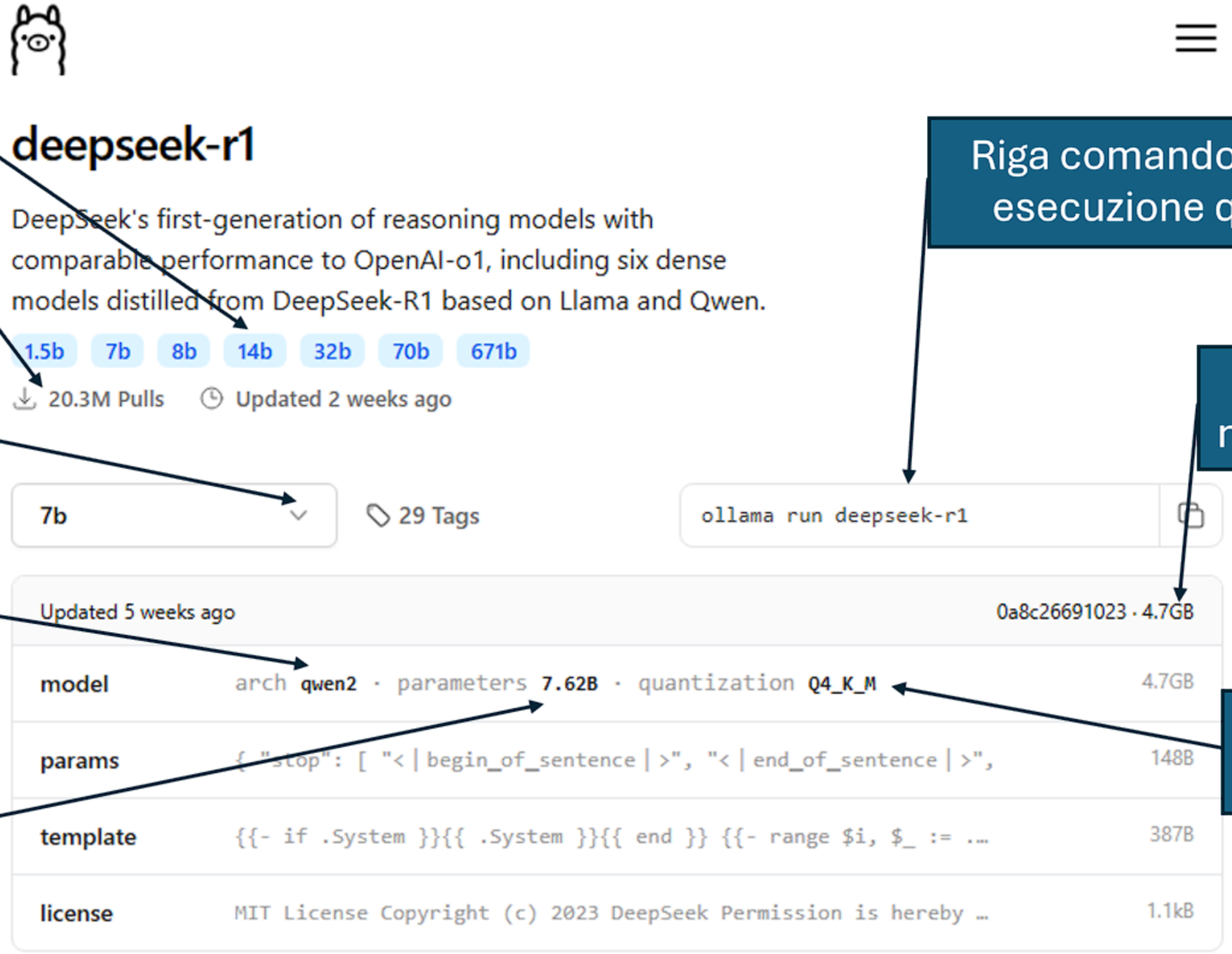
Formati disponibili per questo modello

Quante volte è stato scaricato

Selezione del formato corrente

Motore base da cui è partito l'addestramento

Miliardi di parametri



Riga comando per mandare in esecuzione questo modello

Quanta RAM / VRAM è necessaria per utilizzarlo

Tipologia di quantizzazione

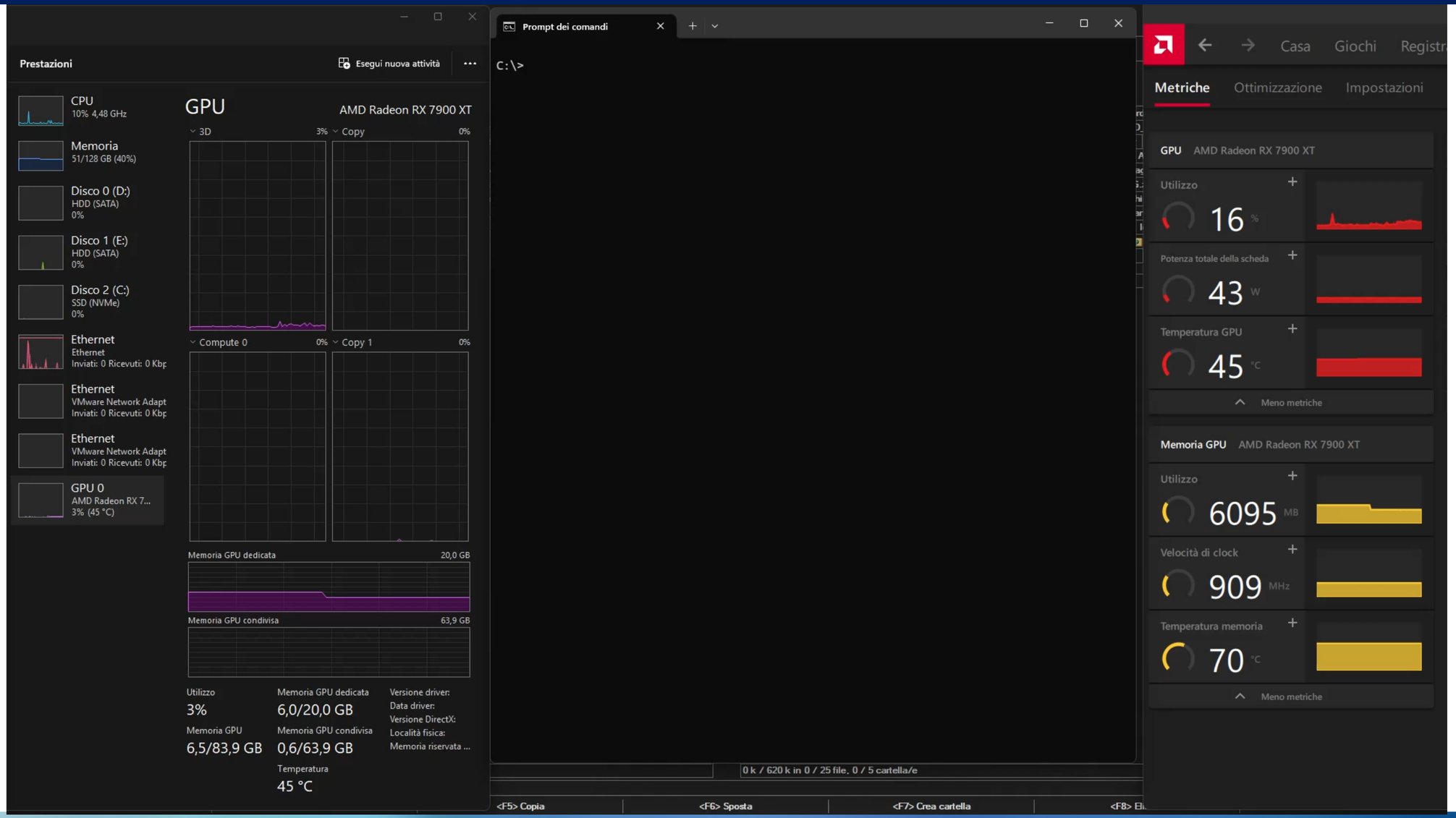
COME SCEGLIERE IL MODELLO AI CORRETTO?

- Per funzionare, un modello AI deve poter essere caricato interamente nella memora RAM (o, ancora meglio, nella VRAM).
- Se esiste un modello specializzato per la vostra specifica esigenza, utilizzate preferibilmente quello.
- Scaricate modelli AI differenti e testateli con lo stesso prompt, per vedere i risultati.
- Provate anche i modelli AI che nel nome contengono uno dei seguenti termini: *Abliterated* o *Ablated* o *Uncensored*.
- Se siete indecisi tra due modelli che sembrano simili, prendete il più nuovo (a meno che non abbia troppi pochi utilizzatori).

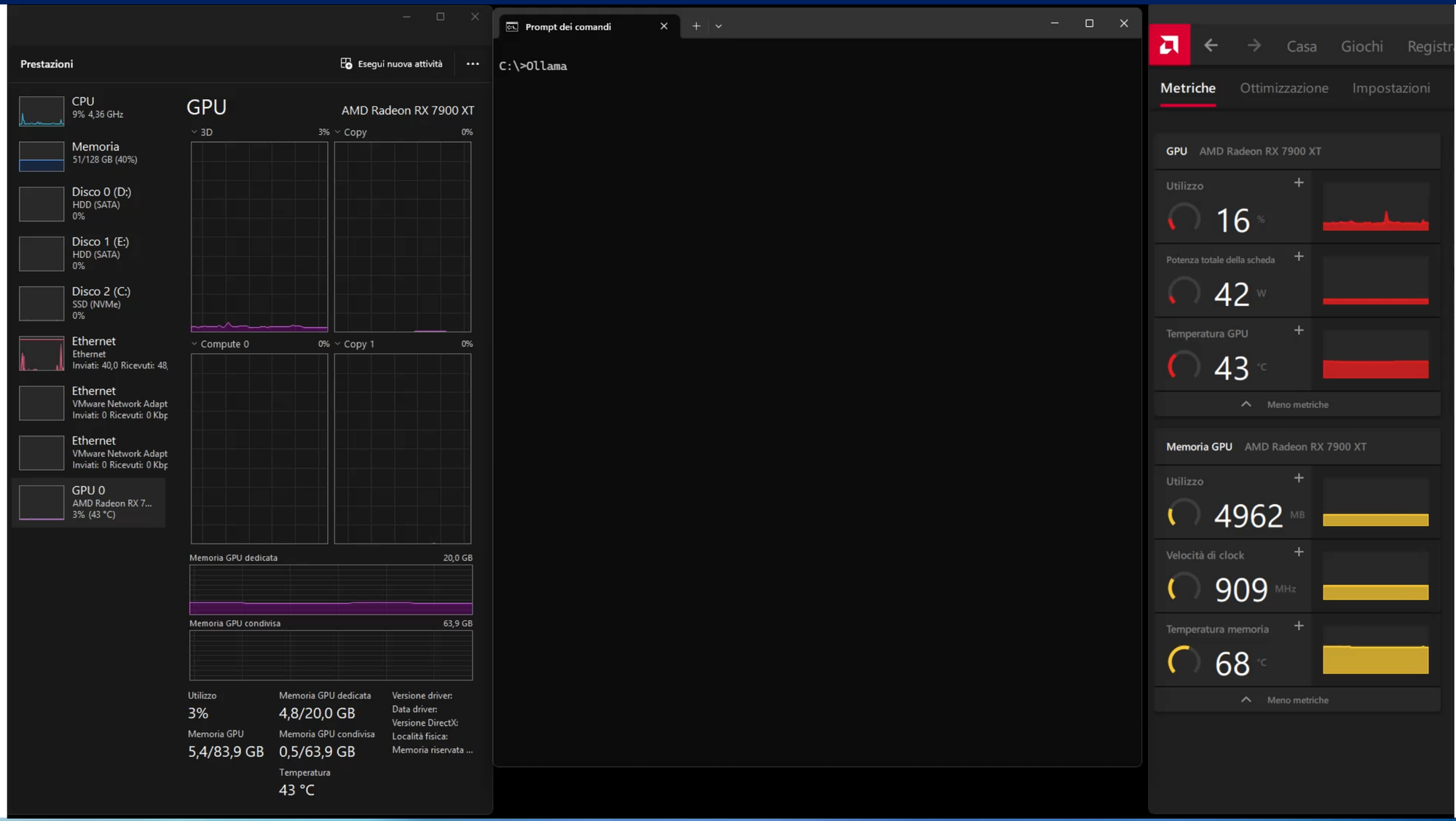
OK, MA TU QUALI USI?

- qwen2.5-coder: per programmare
- phi4: motore AI «non ragionante»
- deepseek-r1: motore «ragionante»
- jimscard/whiterabbit-neo: motore AI specializzato in «*offensive and defensive cybersecurity*»

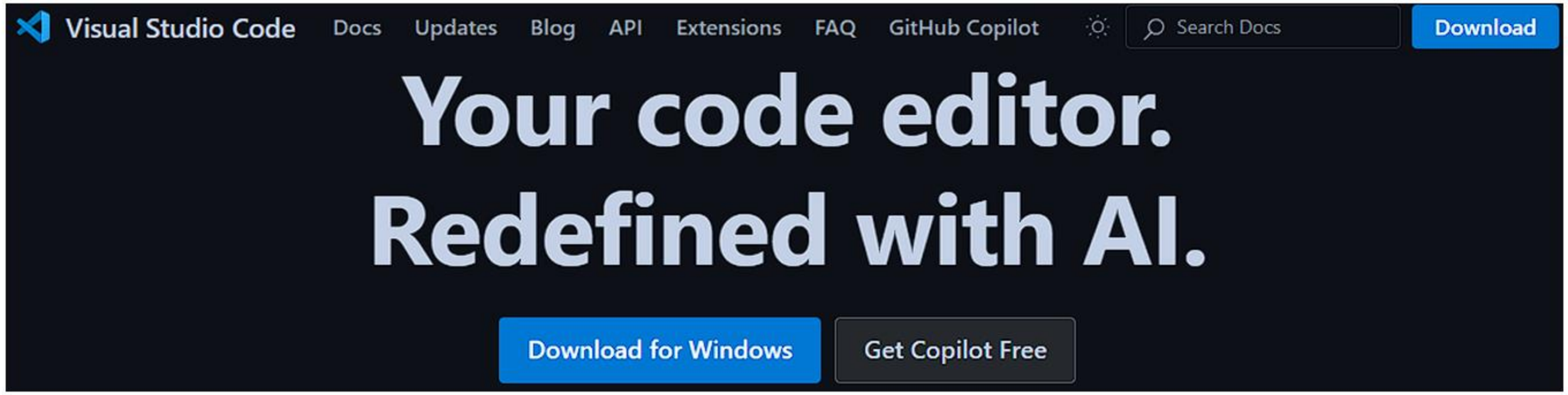
SCARICHIAMO UN MODELLO AI



FACCIAMO UN PRIMO TEST

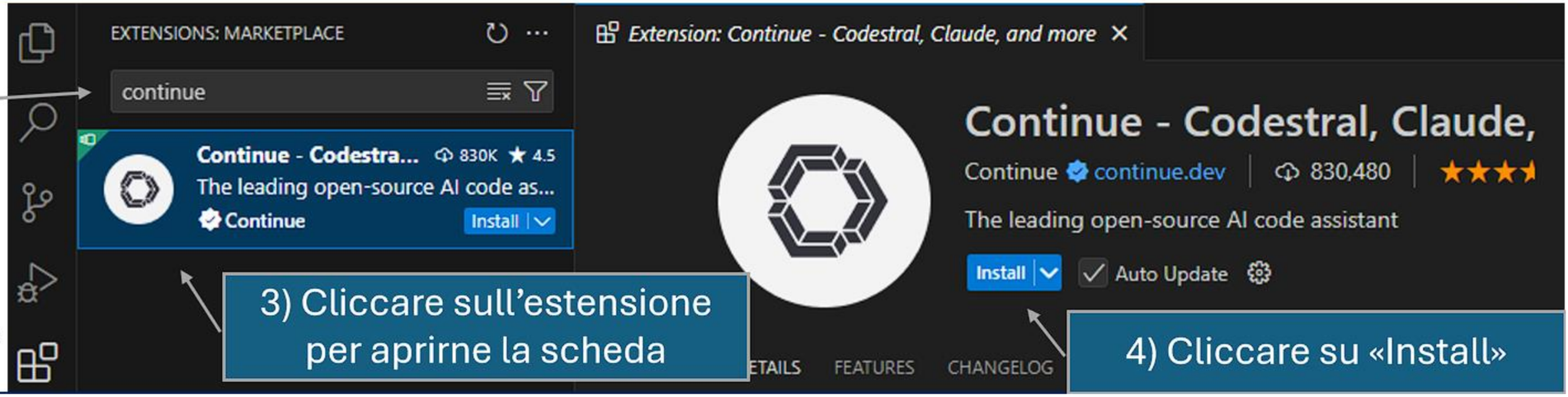


Scaricare, installare ed eseguire Visual Studio Code



2) Digitare il nome per trovare l'estensione «Continue»

1) Cliccare su «Extensions Marketplace»



3) Cliccare sull'estensione per aprirne la scheda

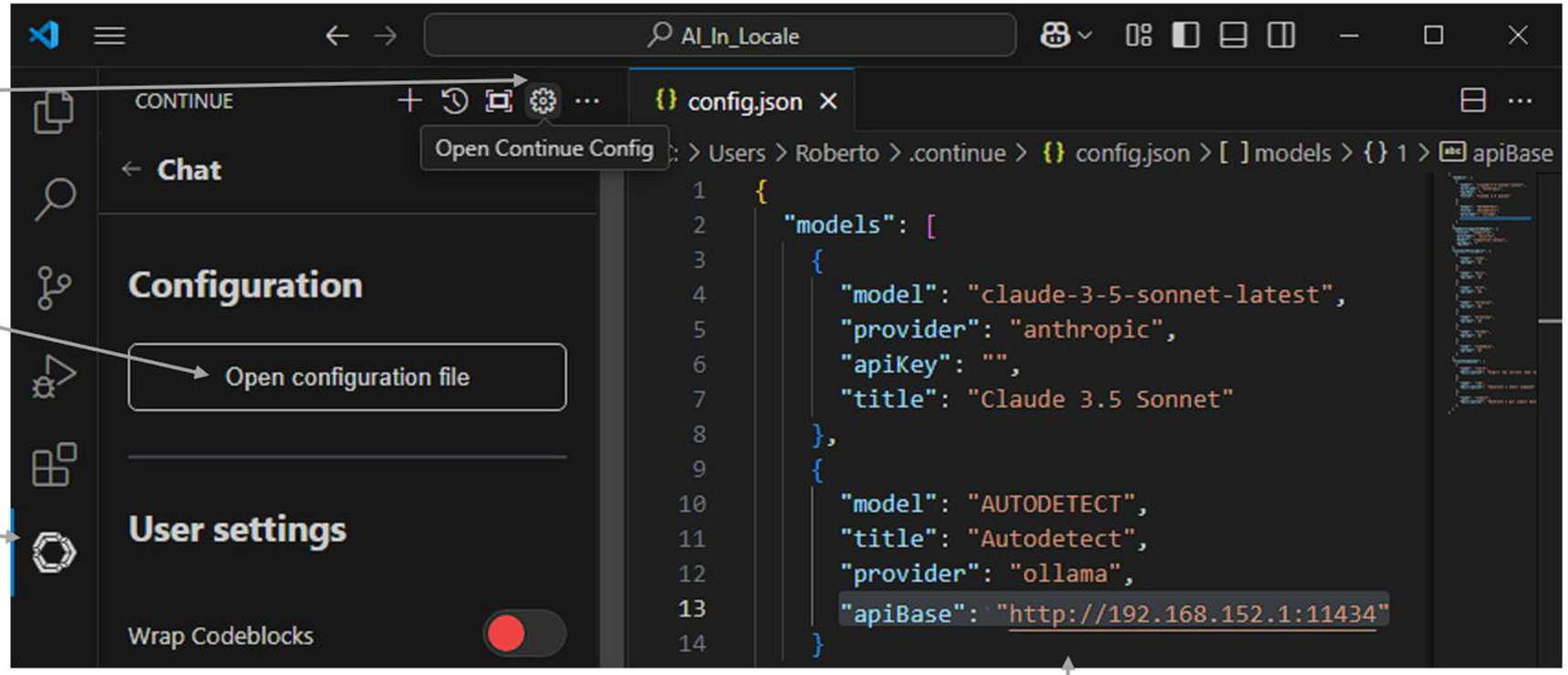
4) Cliccare su «Install»

Collegare «Continue» alla propria installazione di Ollama

2) Cliccare sull'icona dell'ingranaggio per aprire la configurazione

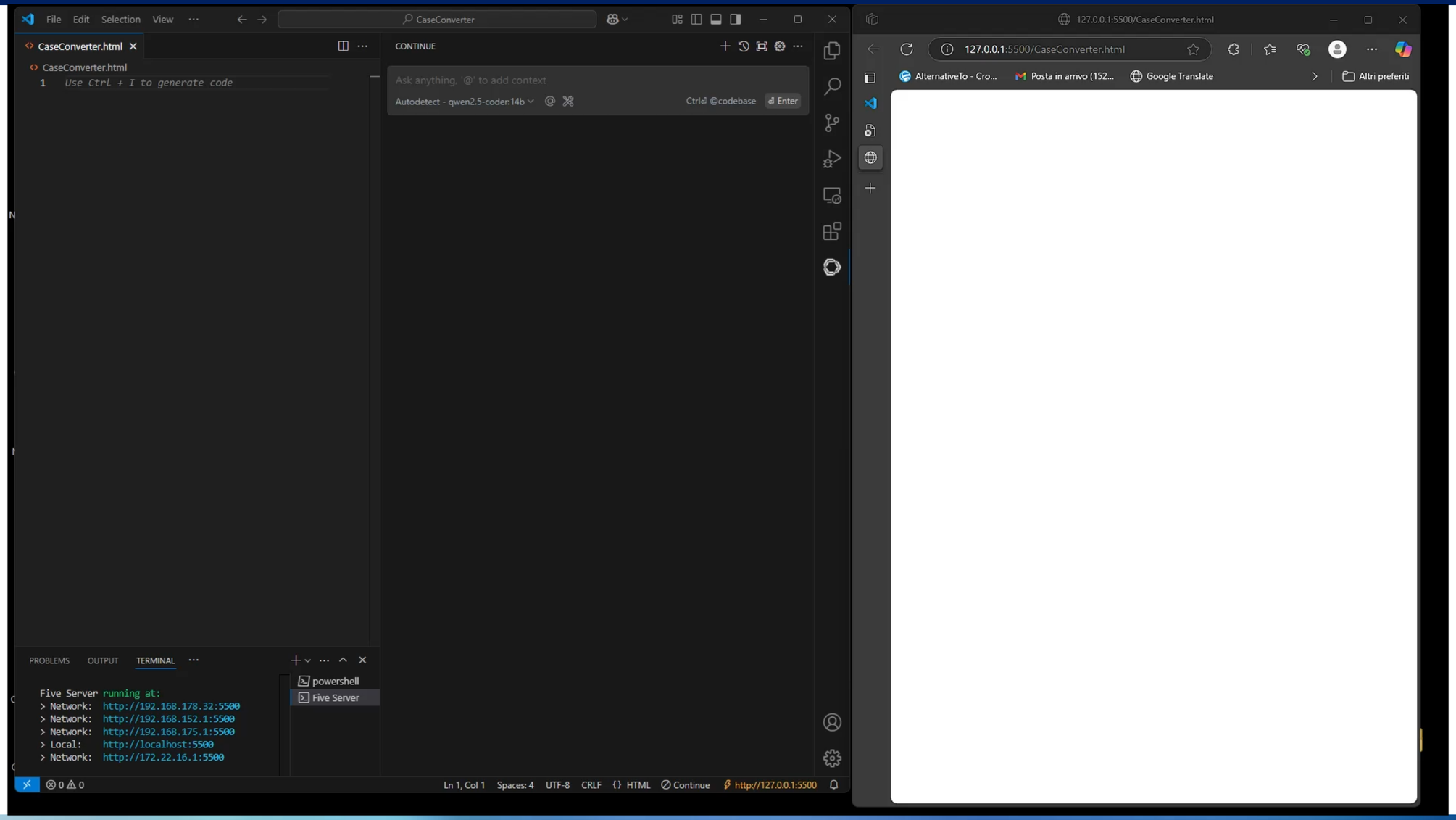
3) Cliccare sul pulsante «Open configuration file»

1) Cliccare sull'icona di «Continue»

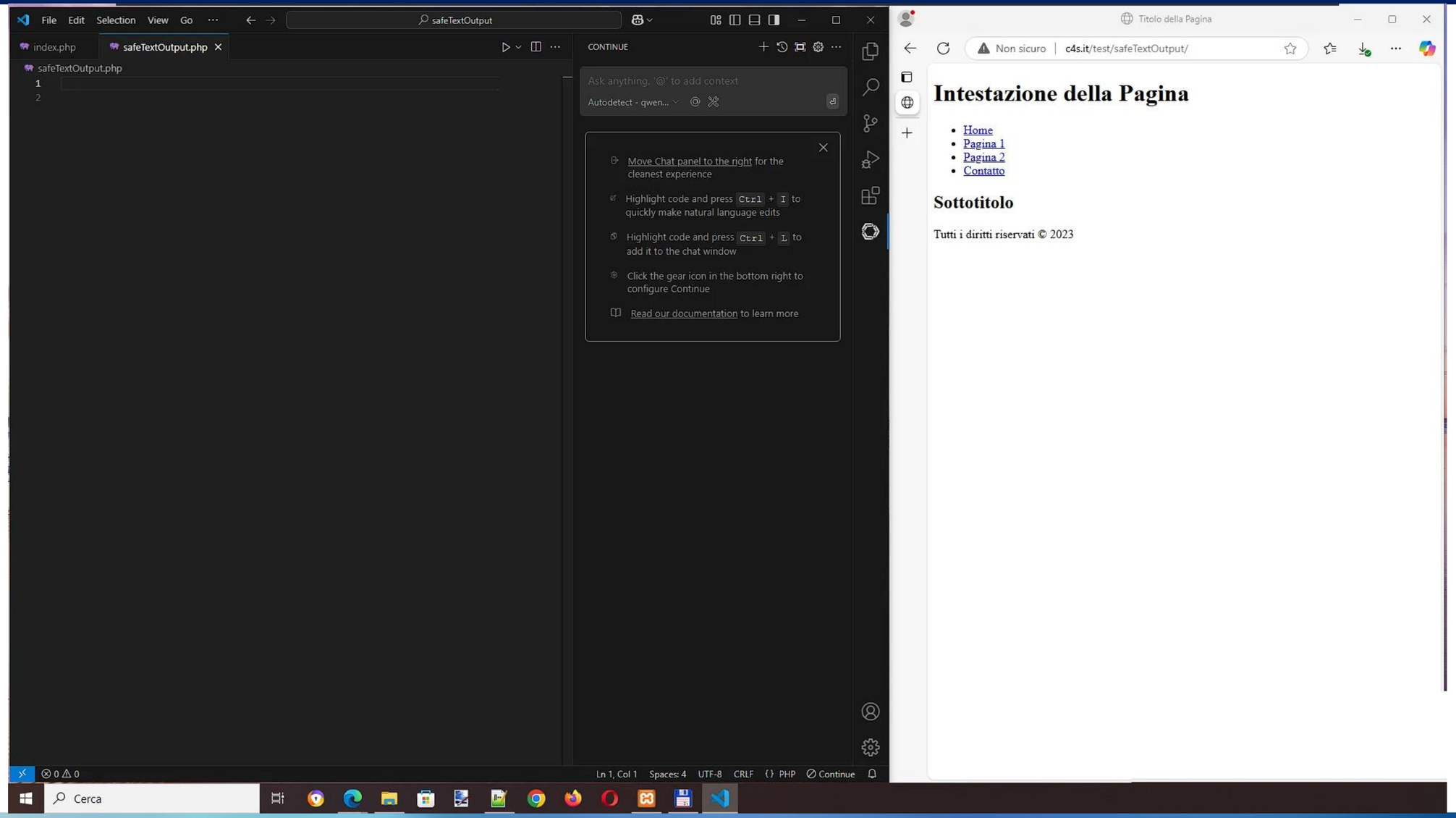


4) Compilare la sezione «AUTODETECT» come nell'esempio, usando l'indirizzo IP locale della propria macchina, seguito dalla porta: 11434

FACCIAMOGLI SCRIVERE UNA SEMPLICE UTILITY



UNA FUNZIONE PHP CHE EVITA GLI ATTACCHI XSS



PROGRAMMARE CON LE AI: CONSIDERAZIONI FINALI

- Si passa da sviluppatori a responsabili di progetto
- L’italiano e l’inglese diventano linguaggi di programmazione
- Servono comunque degli sviluppatori Senior per il controllo; ma se i Junior non servono più, come li creiamo i nuovi Senior?

ABBIAMO SOLO SCALFITO LA SUPERFICIE

Q&A