Ollama: Local AI for my Raspberry PI

Nicola Paro

SPONSOR

beanTech
IT moves your business

CONSORZIO
UNIVERSITARIO
DI PORDENONE
MOLTIPLICATORE DI VALORE

OVERNET.
upgrade your digital skills

[stesi]
Powered by Innovation

Perché voglio utilizzare un LLM nel mio software?

Ci serve per davvero

Non è hype e non è «perché lo fanno tutti»

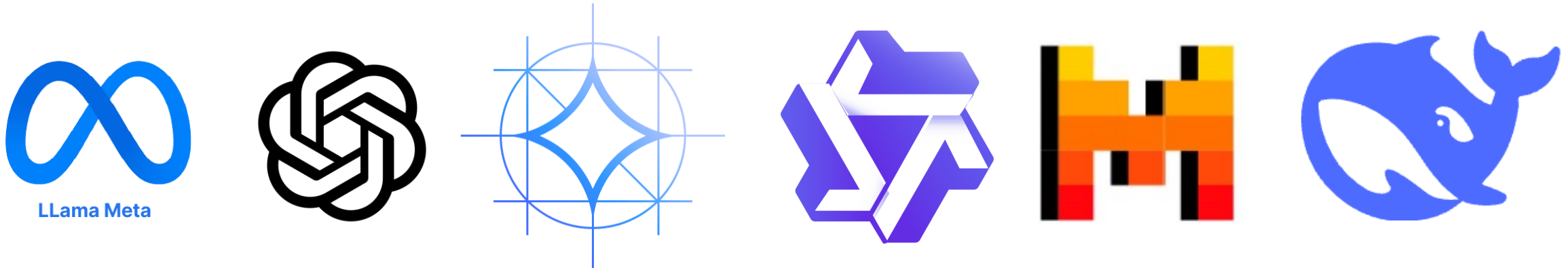
Perché voglio utilizzare un LLM **Locale**?

E’ **davvero** per la privacy

I miei dati sono su un server interno.
I miei clienti usano server interni dedicati anche loro.

Lavoro in scenari disconnessi dal cloud

Che modelli Open Source posso usare in locale?



gpt-oss deepseek-r1 gemma3 embeddinggemma qwen3 deepseek-v3.1 llama3.1 nomic-embed-text llama3.2 mistral qwen2.5 llama3 phi3 llava gemma2 qwen2.5-coder gemma mxbai-embed-large qwen phi4 qwen2 llama2 minicpm-v codellama tinyllama dolphin3 olmo2 mistral-nemo llama3.3 llama3.2-vision deepseek-v3 bge-m3 mistral-small smollm2 llava-llama3 qwq all-minilm mixtral deepseek-coder llama2-uncensored starcoder2 deepseek-coder-v2 codegemma snowflake-arctic-embed phi orca-mini llama4 qwen2.5vl dolphin-mixtral falcon3 openthinker granite3.1-moe granite3.3 gemma3n phi4-reasoning smollm mistral-small3.2 codestral dolphin-llama3 wizardlm2 cogito dolphin-mistral qwen3-coder magistral phi4-mini deepscaler devstral dolphin-phi command-r hermes3 phi3.5 granite3.2-vision yi deepcoder zephyr mistral-small3.1 mistral-large moondream wizard-vicuna-uncensored granite-code starcoder deepseek-llm vicuna openchat deepseek-v2 mistral-openorca codegeex4 openhermes nous-hermes exaone-deep codeqwen qwen2-math snowflake-arctic-embed2 llama2-chinese falcon aya tinydolphin glm4 granite3.2 stable-code nous-hermes2 opencoder neural-chat wizardcoder command-r-plus bakllava bge-large stablelm2 sqlcoder llama3-chatqa llava-phi3 yi-coder granite3.1-dense granite3-dense wizard-math reflection llama3-gradient exaone3.5 dbrx r1-1776 dolphincoder samantha-mistral nemotron-mini tulu3 paraphrase-multilingual starling-lm internlm2 phind-codellama solar xwinlm granite-embedding athene-v2 nemotron llama3-groq-tool-use yarn-llama2 meditron granite3-moe wizardlm-uncensored aya-expanse llama-guard3 smallthinker wizardlm orca2 medllama2 nous-hermes2-mixtral stable-beluga deepseek-v2.5 reader-lm llama-pro yarn-mistral command-r7b shieldgemma phi4-mini-reasoning command-a mathstral nexusraven everythinglm codeup marco-o1 stablelm-zephyr solar-pro duckdb-nsql falcon2 magicoder mistrallite codebooga bespoke-minicheck nuextract wizard-vicuna granite3-guardian megadolphin notux open-orca-platypus2 notus sailor2 firefunction-v2 goliath alfred command-r7b-arabic

Perché **IO** voglio utilizzare un LLM su una Raspberry PI?

Ma perché no?

Scelta del LLM Runner

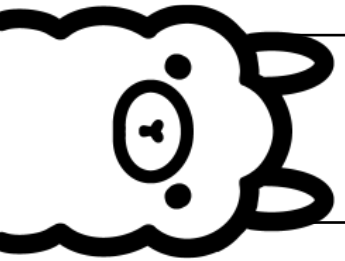
LLM Runners

Ollama	LM Studio	Foundry Local
Aggiornamenti frequenti, documentazione chiara.	Interfaccia desktop molto popolare per chi vuole un'alternativa "ChatGPT offline".	Nuovo, ma in rapida crescita: non ancora "di massa" come Ollama, ma più solido in ottica enterprise.
API REST semplici (localhost:11434).	API OpenAI-compatible, quindi integrabile ovunque senza cambiare librerie.	API REST e SDK OpenAI-compatible.
Community enorme (soprattutto su macOS, ora anche Linux e Windows).	Ampio uso in community e progetti personali/produttivi.	Forte spinta da Microsoft, integrazione con VS Code e strumenti Azure.

LLM Runners, support by OS

	Windows	Linux	Mac
Ollama	✔	✔	✔
Foundry Local	✔	✘	✔
LM Studio	✔	✔	✔

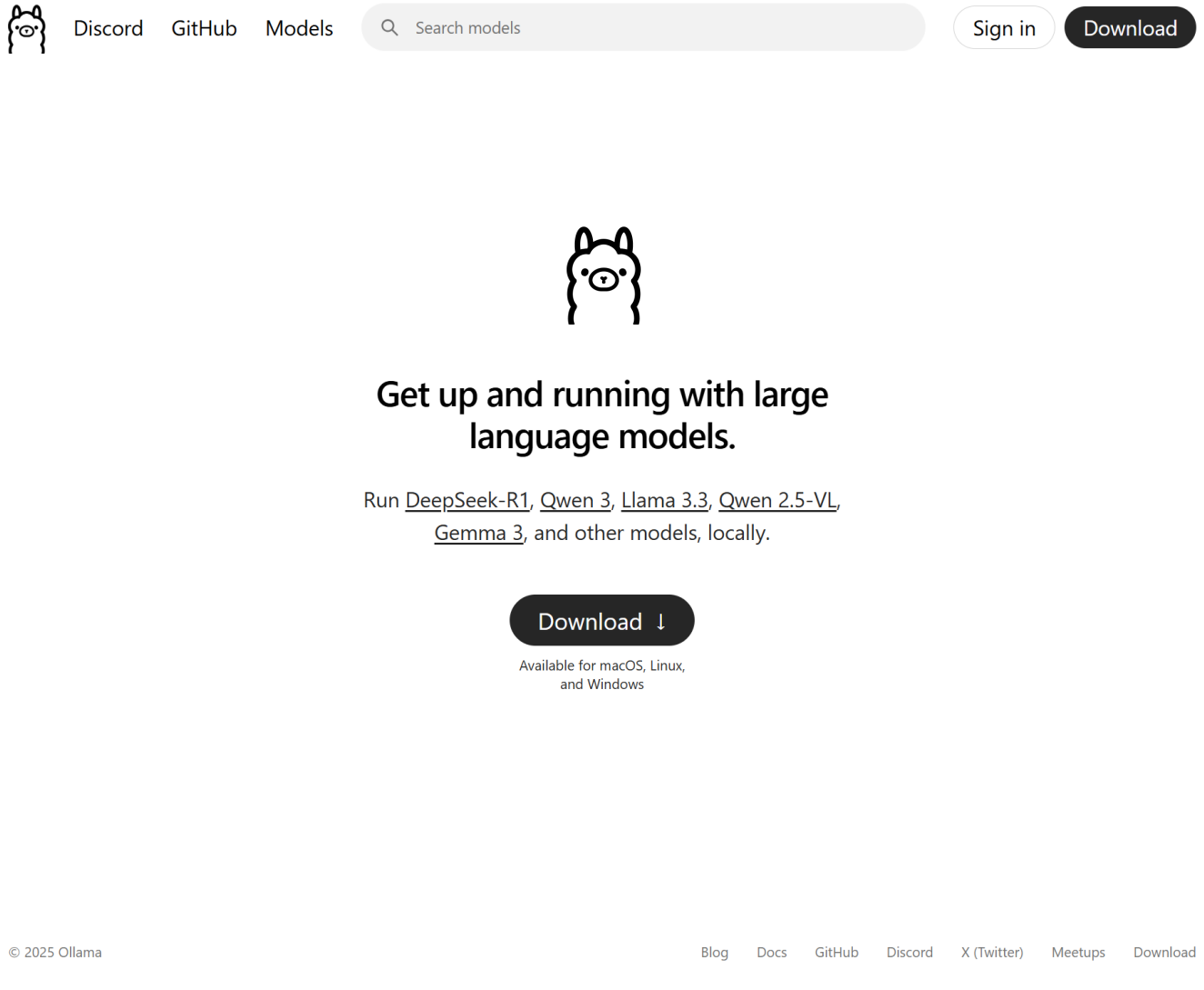
LLM Runners, support by Silicon



	CPU (x86, x64)	CPU (ARM)	GPU	NPU
Ollama	✔	✔	✔	✘
Foundry Local	✔	✔	✔	✔
LM Studio	✔	Solo su Windows	✔	✘

Ollama

- Sviluppato da un team indipendente con background in AI e developer tools.
- Obiettivo di rendere semplice e accessibile l'esecuzione di **LLM in locale**, per una maggiore **privacy**, **controllo** e **autonomia**.



Ollama

- 2023 (inizio): Prime versioni interne
- Estate 2023: Lancio pubblico su GitHub con supporto a modelli GGUF e binari ottimizzati.
- 2024: Integrazione API REST, supporto a GPU e ottimizzazioni multi-piattaforma. Aggiunta di modelli più recenti come Mistral e Gemma.

Ollama

Ollama utilizza container di modelli LLM (.bin o .gguf)

Supporta GPU / CPU a seconda del sistema

Inferenza locale: nessun invio dati a server esterni

Compatibile con diversi modelli (es: LLaMA, Mistral, Gemma...)

Integrazione semplice: CLI, API REST

Supporto a modelli di visione e modelli tool

Costi dell’AI

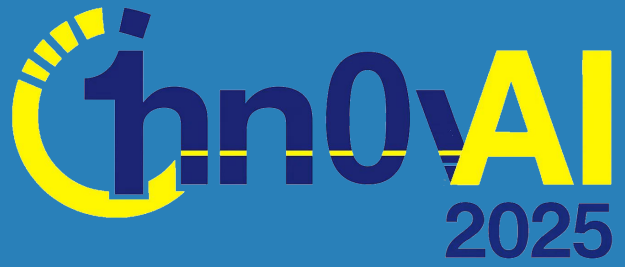
Quanto costa Ollama su Raspberry PI?

0.13€ / Million Tokens

(in Italia)

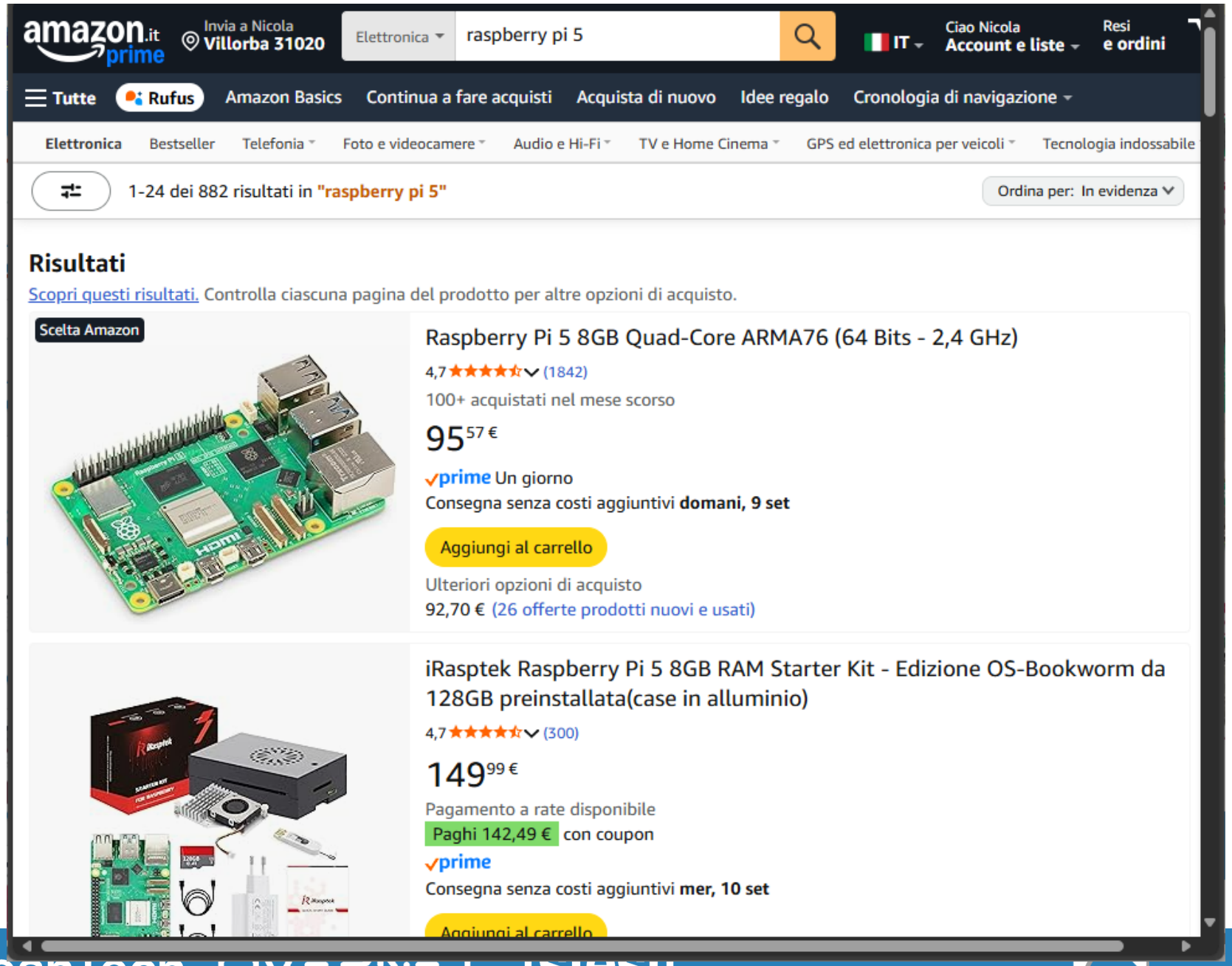
Quanto costa Ollama su Raspberry PI?

	Costo input (€/1k tok)	Costo output (€/1k tok)	Tokens/sec (tipico)	Costo per 1M tok
GPT-4 (Azure, 8k)	~0,028 €	~0,055 €	~15–30 t/s	~82 €
GPT-3.5 Turbo (Azure)	~0,0014 €	~0,0018 €	~50–100 t/s	~1,8 €
GPT-4o-mini (Azure)	~0,00014 €	~0,00055 €	~50–100 t/s	~0,7 €
llama3.2:3b (RPi5, Ollama)	~0,00013–0,00016 € (energia)	~0,00013–0,00016 € (energia)	~4–5 t/s	~0,13–0,16 €



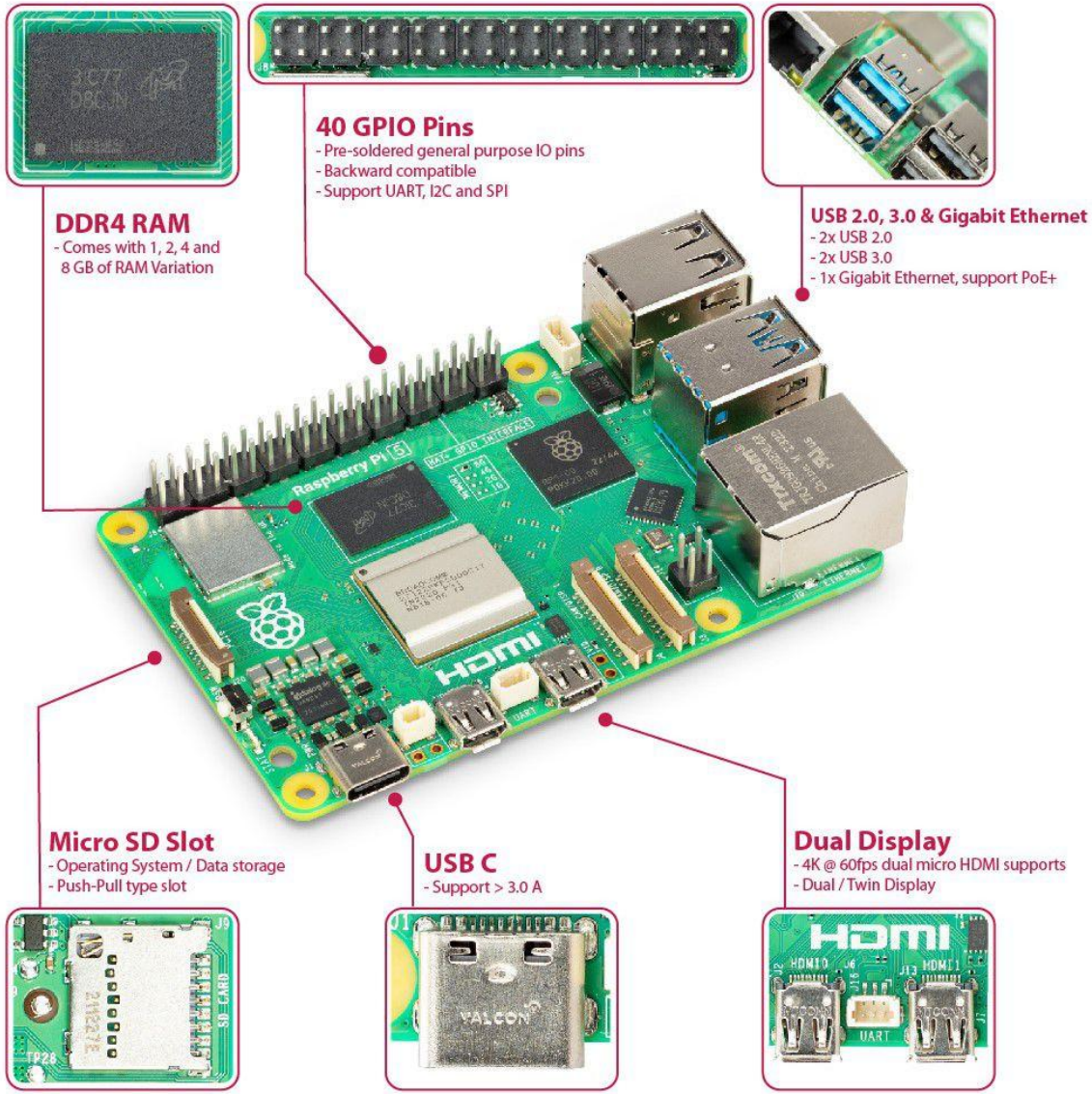
Let’s get started!

Hardware

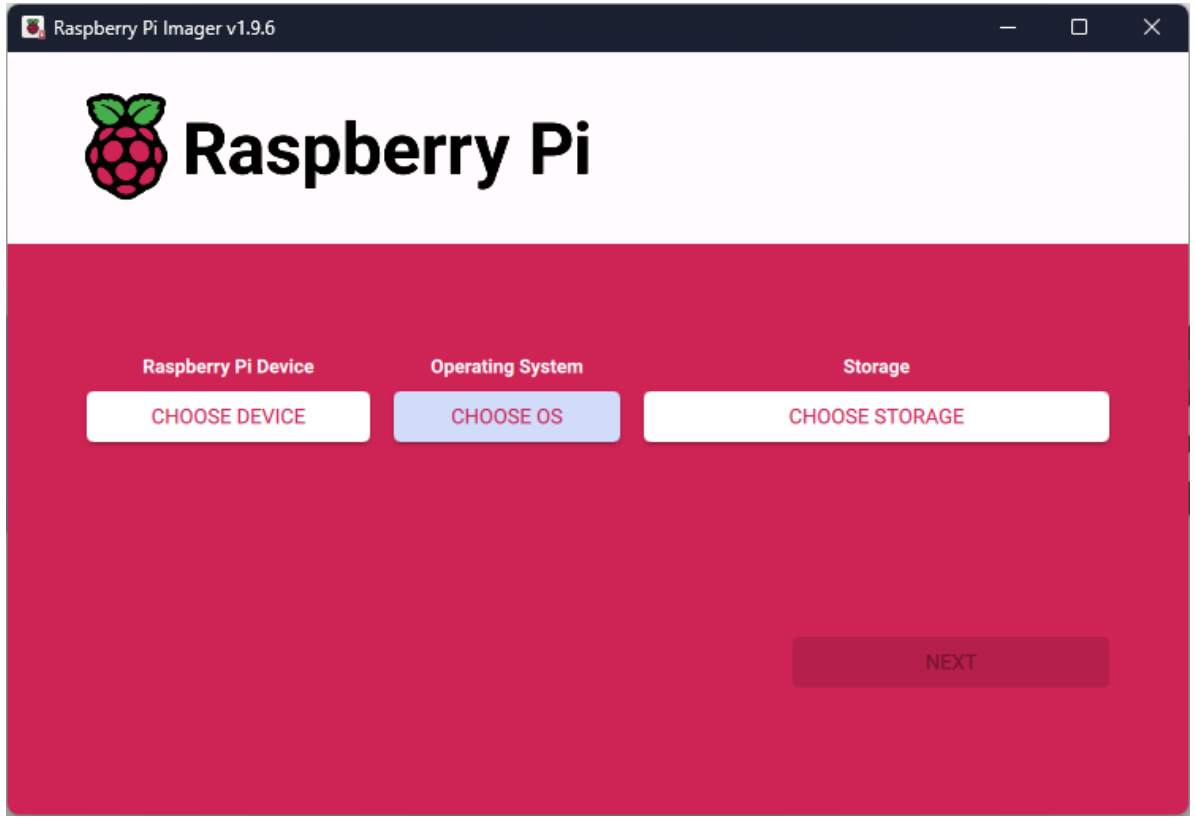

amazon.it prime
Invia a Nicola
Villorba 31020
Elettronica
raspberry pi 5
IT
Ciao Nicola
Account e liste
Resi
e ordini
Tutte
Rufus
Amazon Basics
Continua a fare acquisti
Acquista di nuovo
Idee regalo
Cronologia di navigazione
Elettronica
Bestseller
Telefonia
Foto e videocamere
Audio e Hi-Fi
TV e Home Cinema
GPS ed elettronica per veicoli
Tecnologia indossabile
1-24 dei 882 risultati in "raspberry pi 5"
Ordina per: In evidenza
Risultati
Scopri questi risultati. Controlla ciascuna pagina del prodotto per altre opzioni di acquisto.
Scelta Amazon
Raspberry Pi 5 8GB Quad-Core ARMA76 (64 Bits - 2,4 GHz)
4,7 (1842)
100+ acquistati nel mese scorso
95 57 €
prime Un giorno
Consegna senza costi aggiuntivi domani, 9 set
Aggiungi al carrello
Ulteriori opzioni di acquisto
92,70 € (26 offerte prodotti nuovi e usati)
iRasptek Raspberry Pi 5 8GB RAM Starter Kit - Edizione OS-Bookworm da 128GB preinstallata(case in alluminio)
4,7 (300)
149 99 €
Pagamento a rate disponibile
Paghi 142,49 € con coupon
prime
Consegna senza costi aggiuntivi mer, 10 set
Aggiungi al carrello

Hardware

- Raspberry PI 5
  Praticamente è un mini pc, ma con le GPIO
  - 64bit ARM CPU
  - 16GB RAM
- MicroSD 64GB (o più)
- **Prendete il dissipatore attivo**



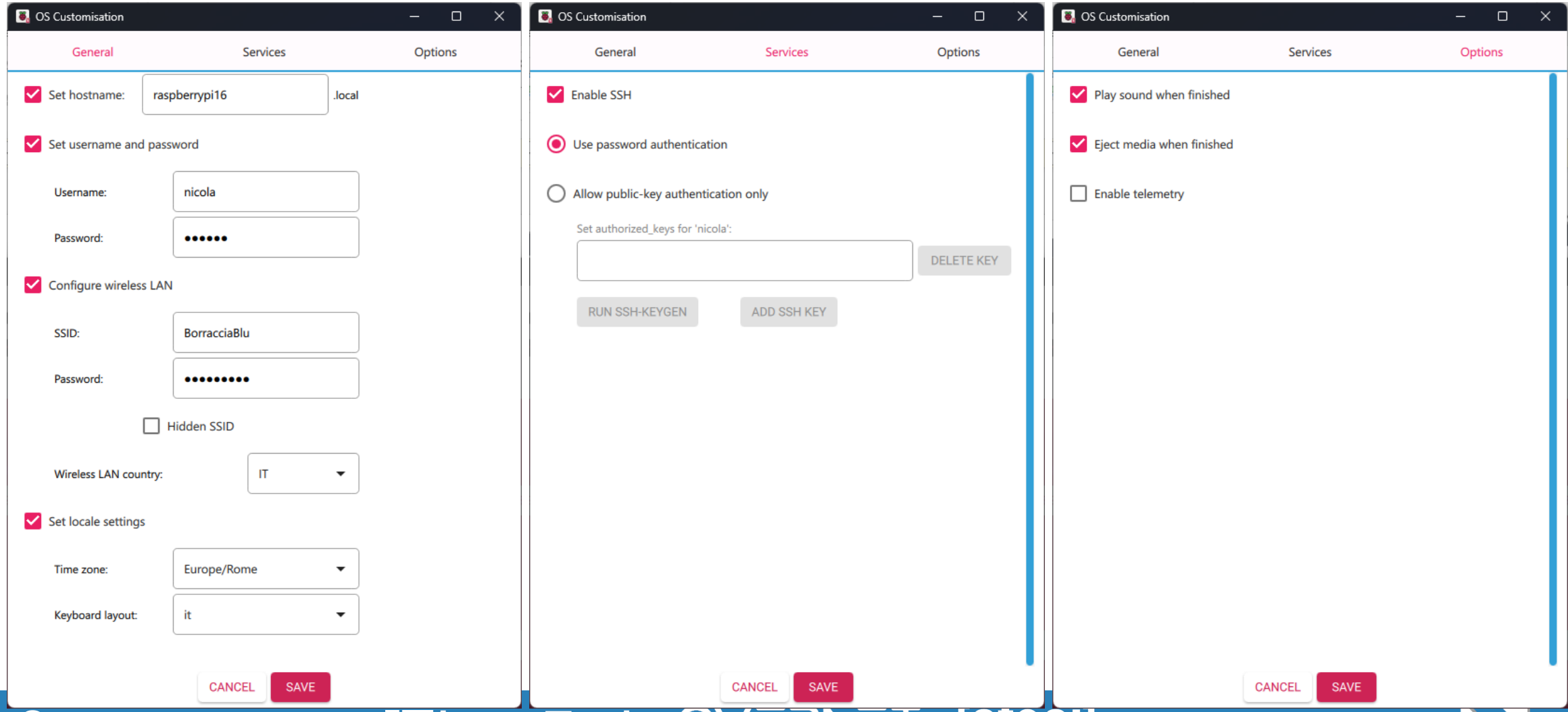
Software



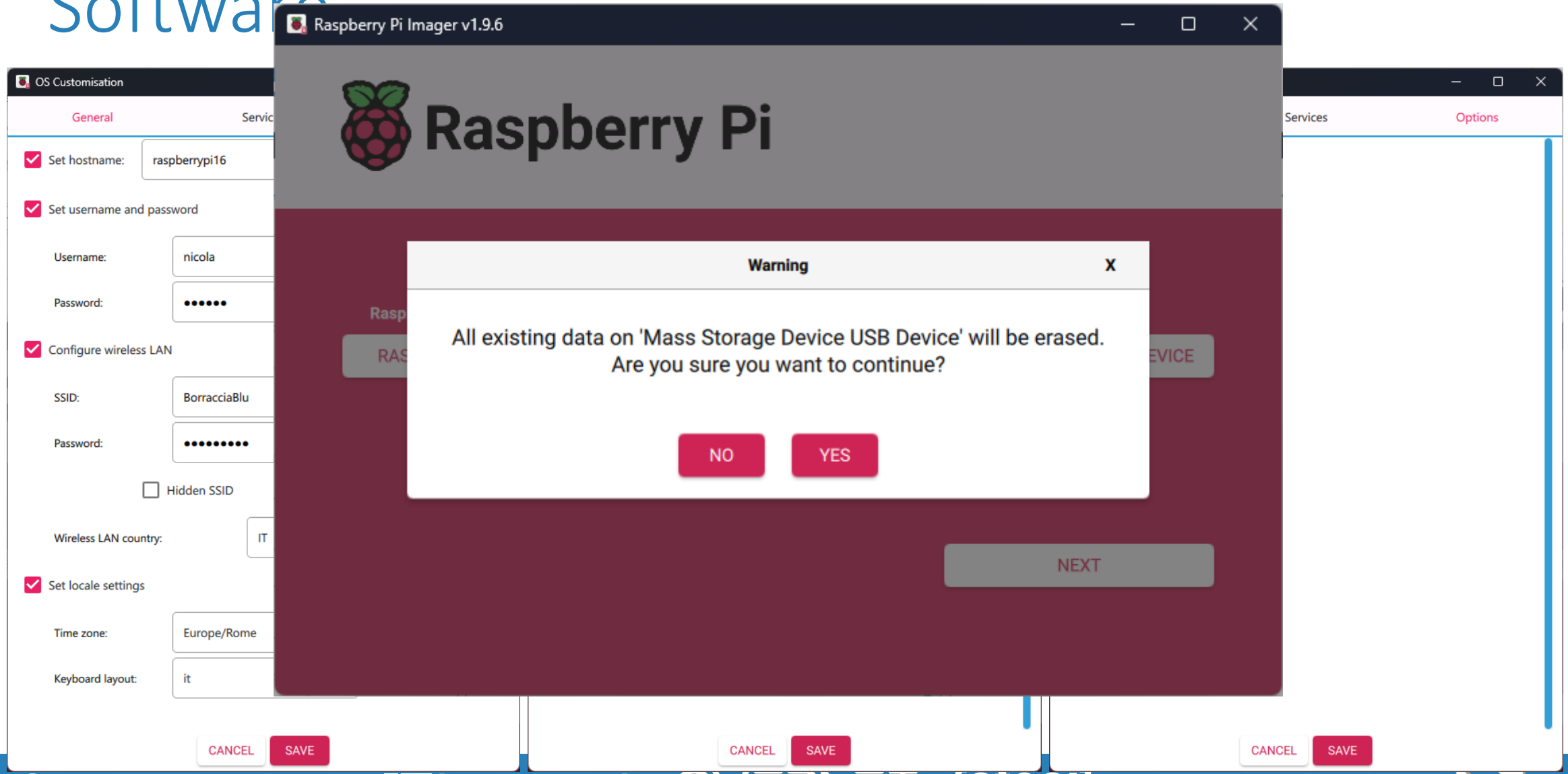
Download Raspberry Pi Imager on your computer

https://www.raspberrypi.com/software/

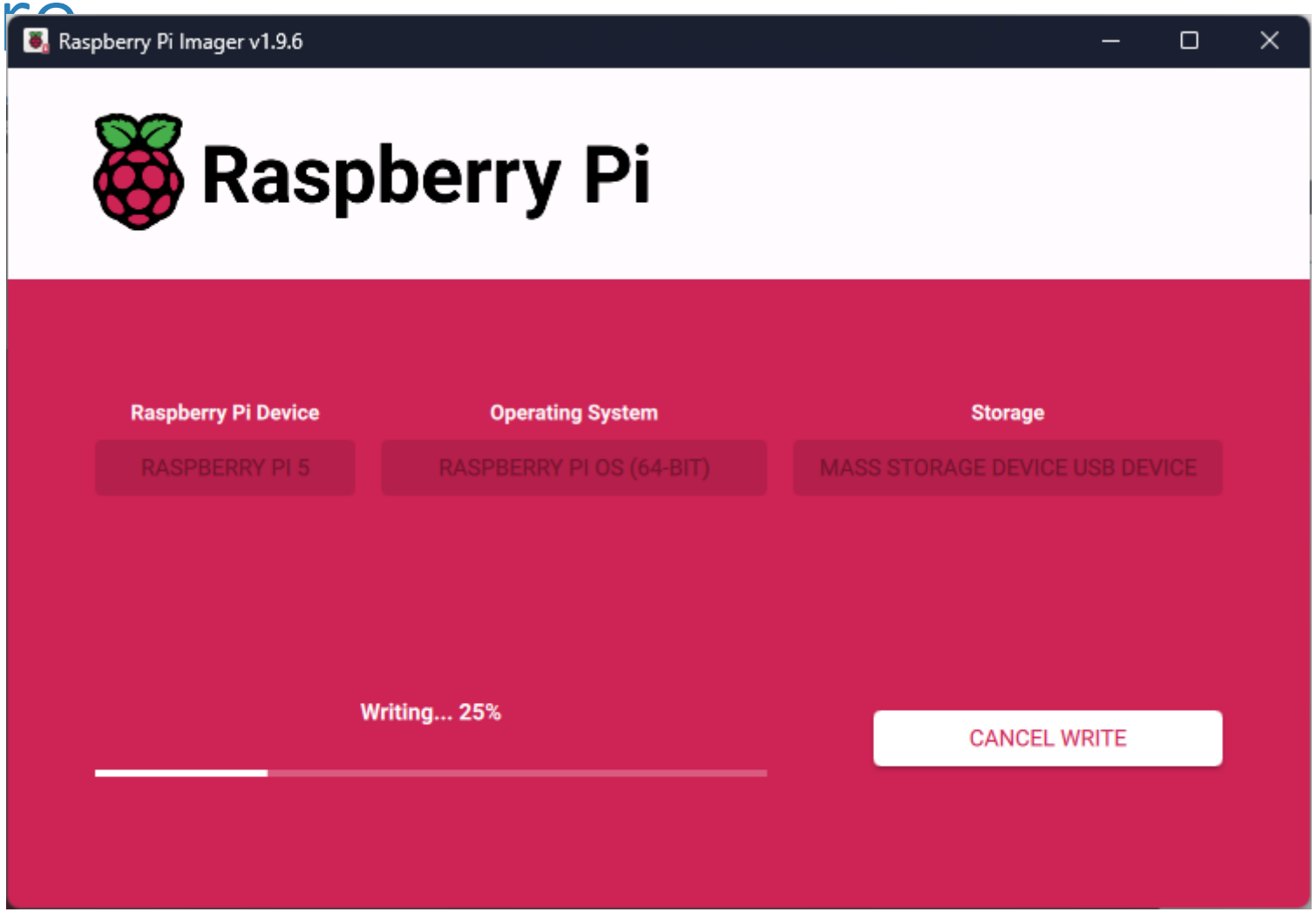
Software


OS Customisation
General
Services
Options
Set hostname:
raspberrypi16
.local
Set username and password
Username:
nicola
Password:
Configure wireless LAN
SSID:
BorracciaBlu
Password:
Hidden SSID
Wireless LAN country:
IT
Set locale settings
Time zone:
Europe/Rome
Keyboard layout:
it
CANCEL
SAVE
OS Customisation
General
Services
Options
Enable SSH
Use password authentication
Allow public-key authentication only
Set authorized_keys for 'nicola':
DELETE KEY
RUN SSH-KEYGEN
ADD SSH KEY
CANCEL
SAVE
OS Customisation
General
Services
Options
Play sound when finished
Eject media when finished
Enable telemetry
CANCEL
SAVE

Software

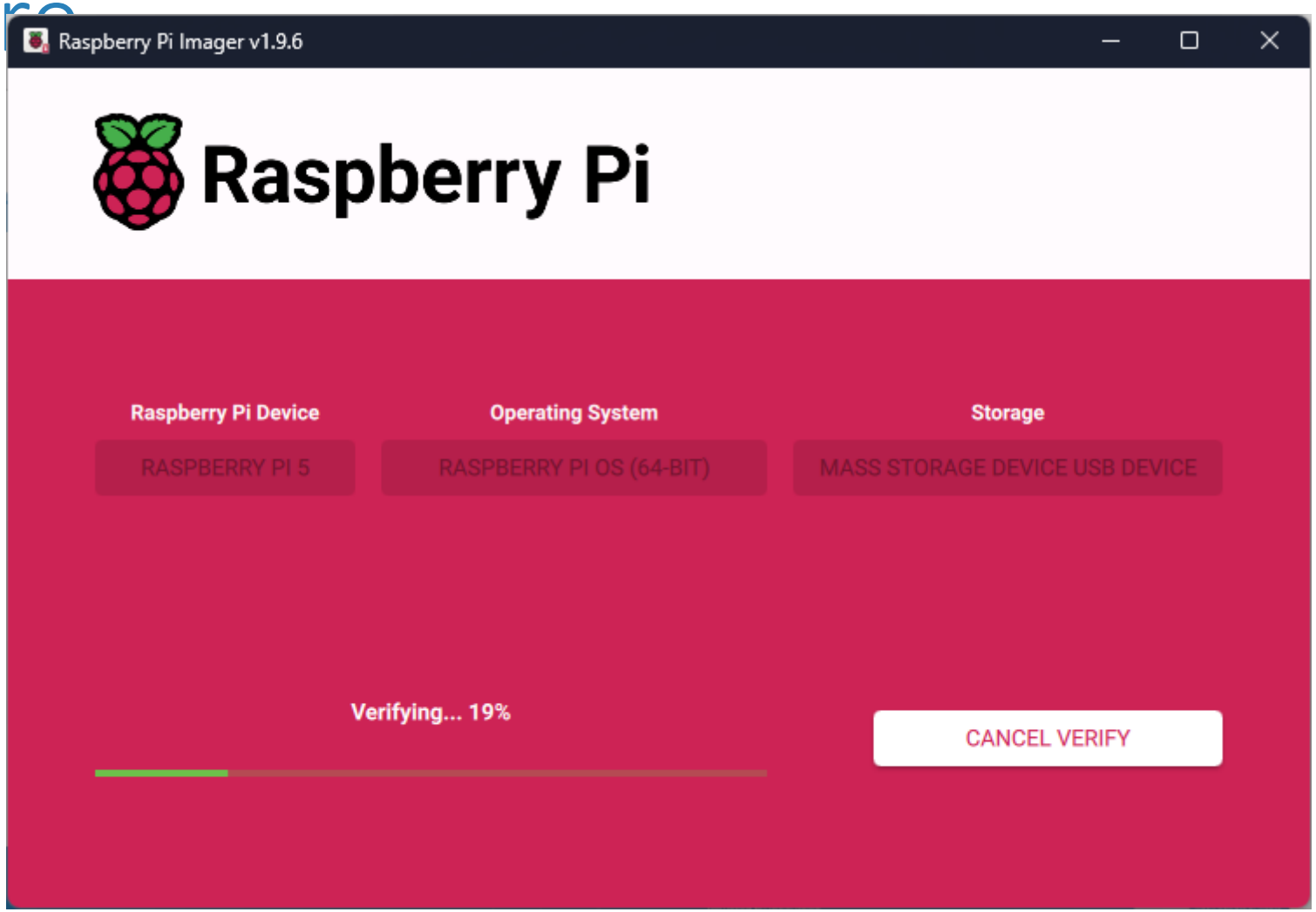

OS Customisation
General
Servic
Set hostname:
raspberrypi16
Set username and password
Username:
nicola
Password:
Configure wireless LAN
SSID:
BorracciaBlu
Password:
Hidden SSID
Wireless LAN country:
IT
Set locale settings
Time zone:
Europe/Rome
Keyboard layout:
it
CANCEL
SAVE
Raspberry Pi Imager v1.9.6
Raspberry Pi
Warning
All existing data on 'Mass Storage Device USB Device' will be erased.
Are you sure you want to continue?
NO
YES
NEXT
Services
Options
CANCEL
SAVE
CANCEL
SAVE

Software


Raspberry Pi Imager v1.9.6
Raspberry Pi
Raspberry Pi Device
RASPBERRY PI 5
Operating System
RASPBERRY PI OS (64-BIT)
Storage
MASS STORAGE DEVICE USB DEVICE
Writing... 25%
CANCEL WRITE

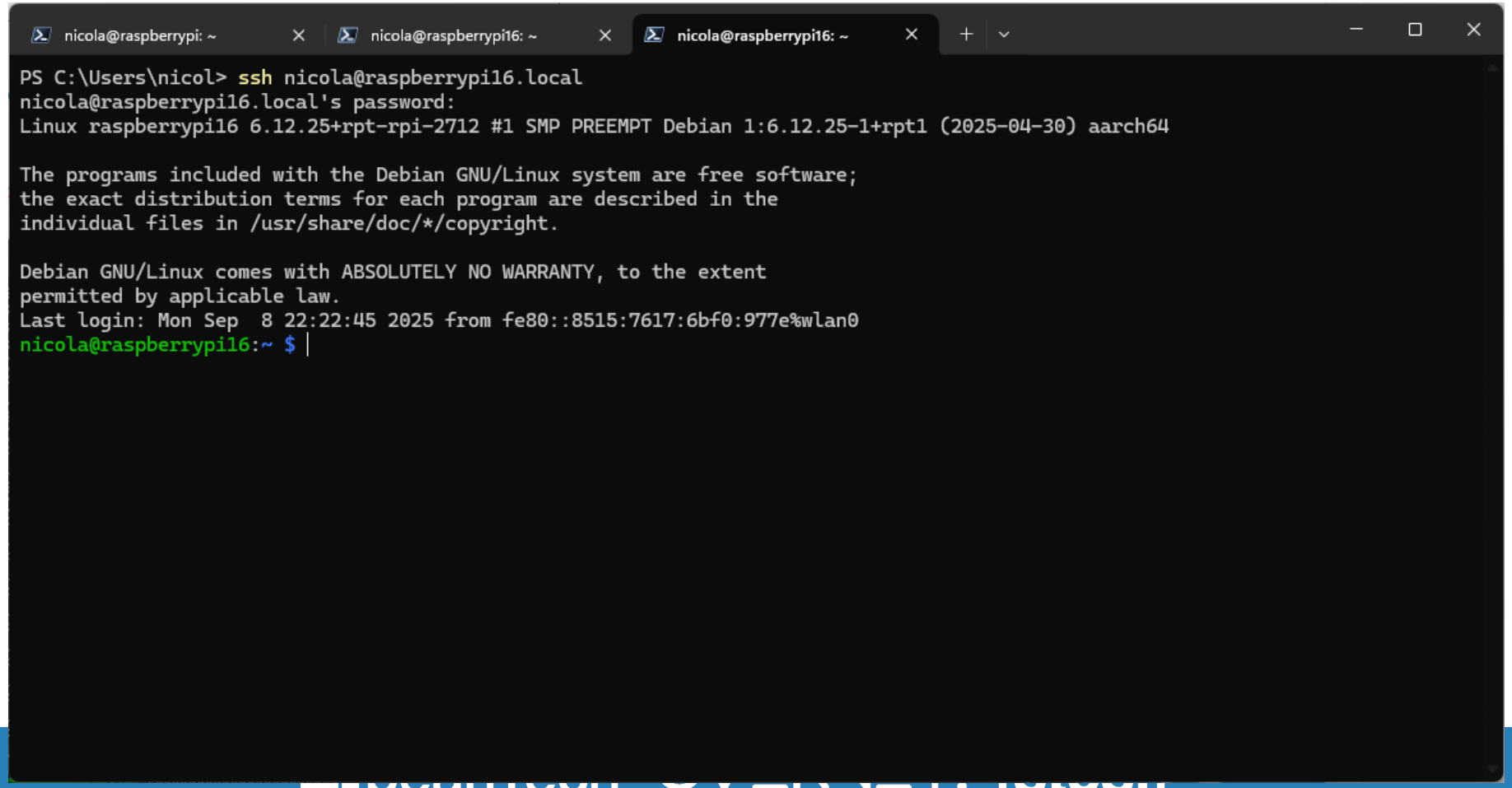

A Few
Moments Later

Software


Raspberry Pi Imager v1.9.6
Raspberry Pi
Raspberry Pi Device
RASPBERRY PI 5
Operating System
RASPBERRY PI OS (64-BIT)
Storage
MASS STORAGE DEVICE USB DEVICE
Verifying... 19%
CANCEL VERIFY

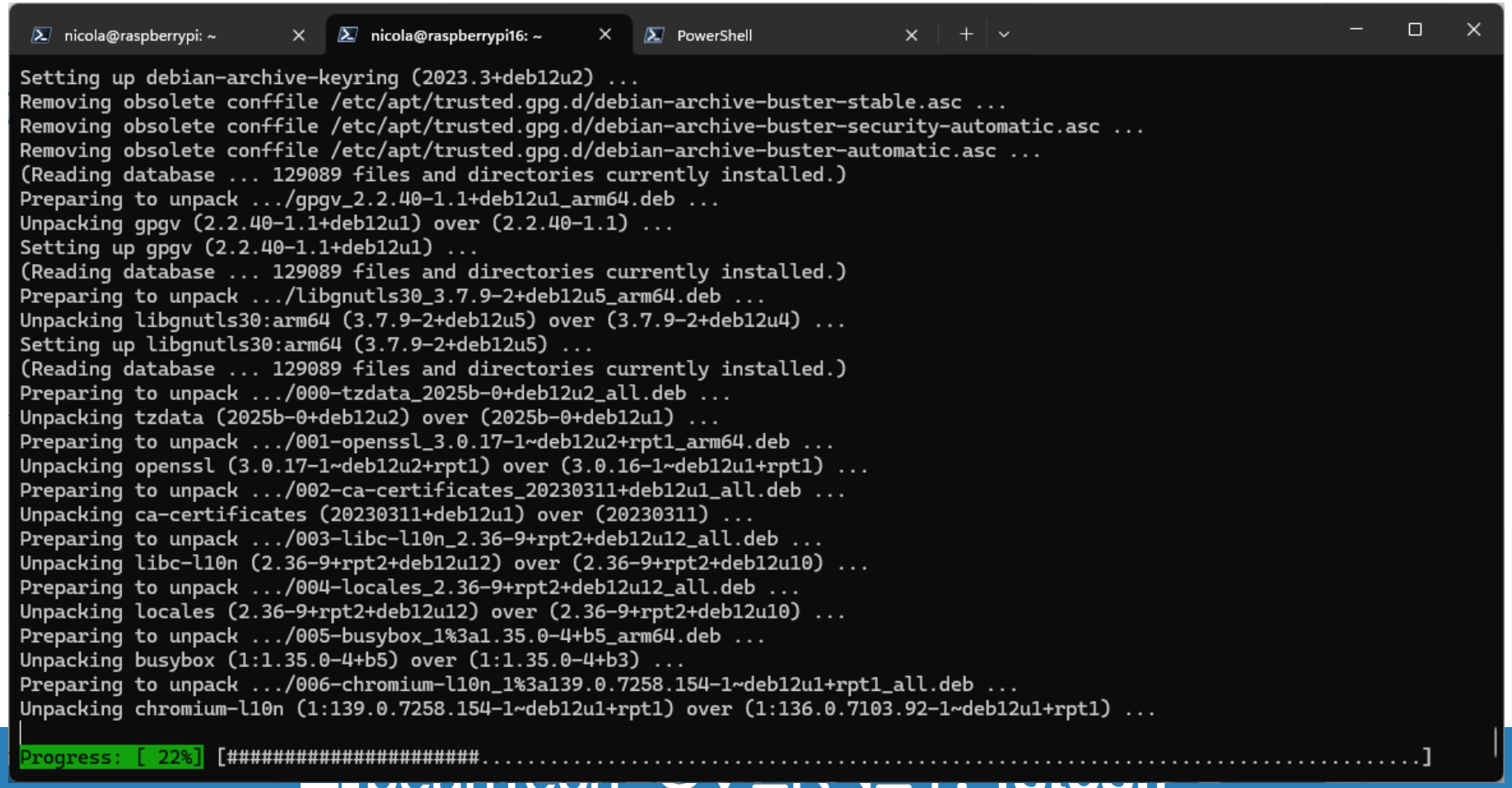
Connect via ssh

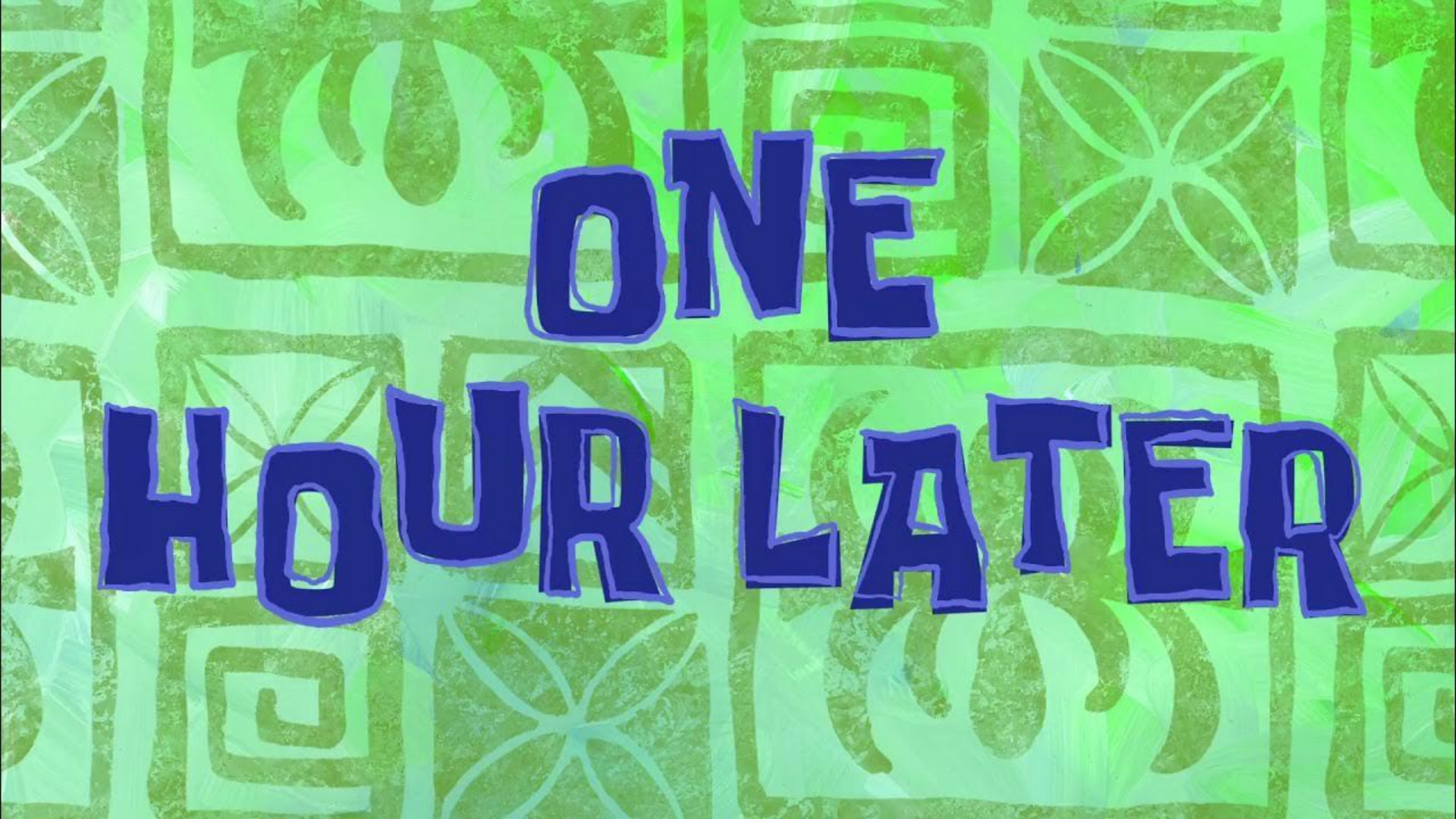
ssh <user>@<hostname>



Update

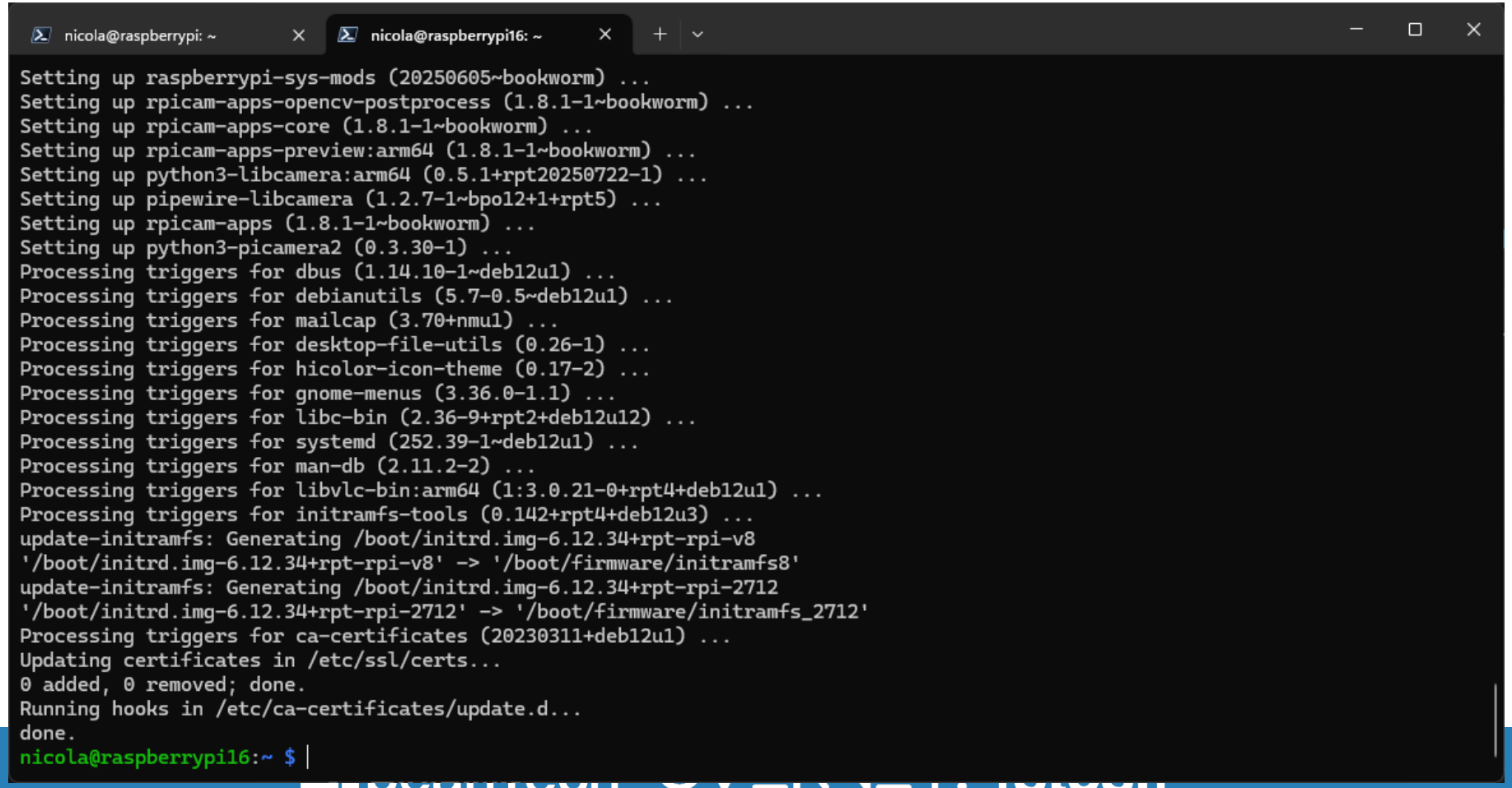
sudo apt update && sudo apt upgrade -Y




ONE
HOUR LATER

Update

sudo apt update && sudo apt upgrade -Y

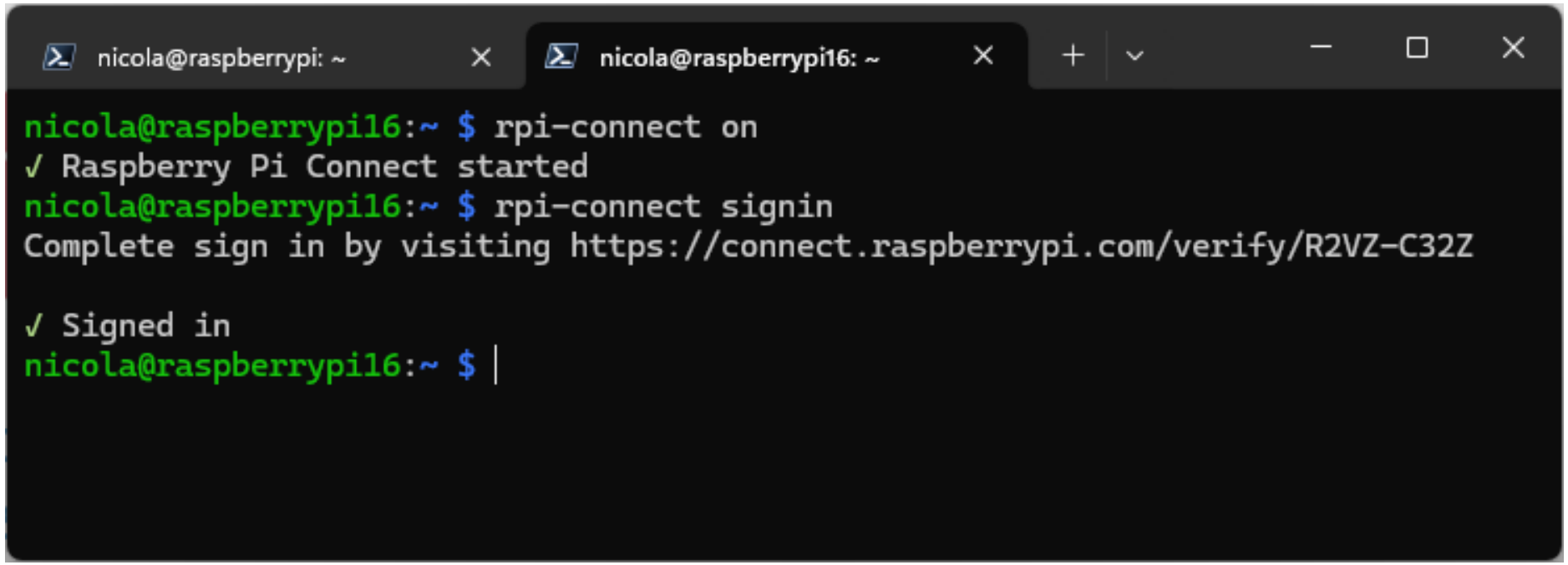


Raspberry Pi Connect

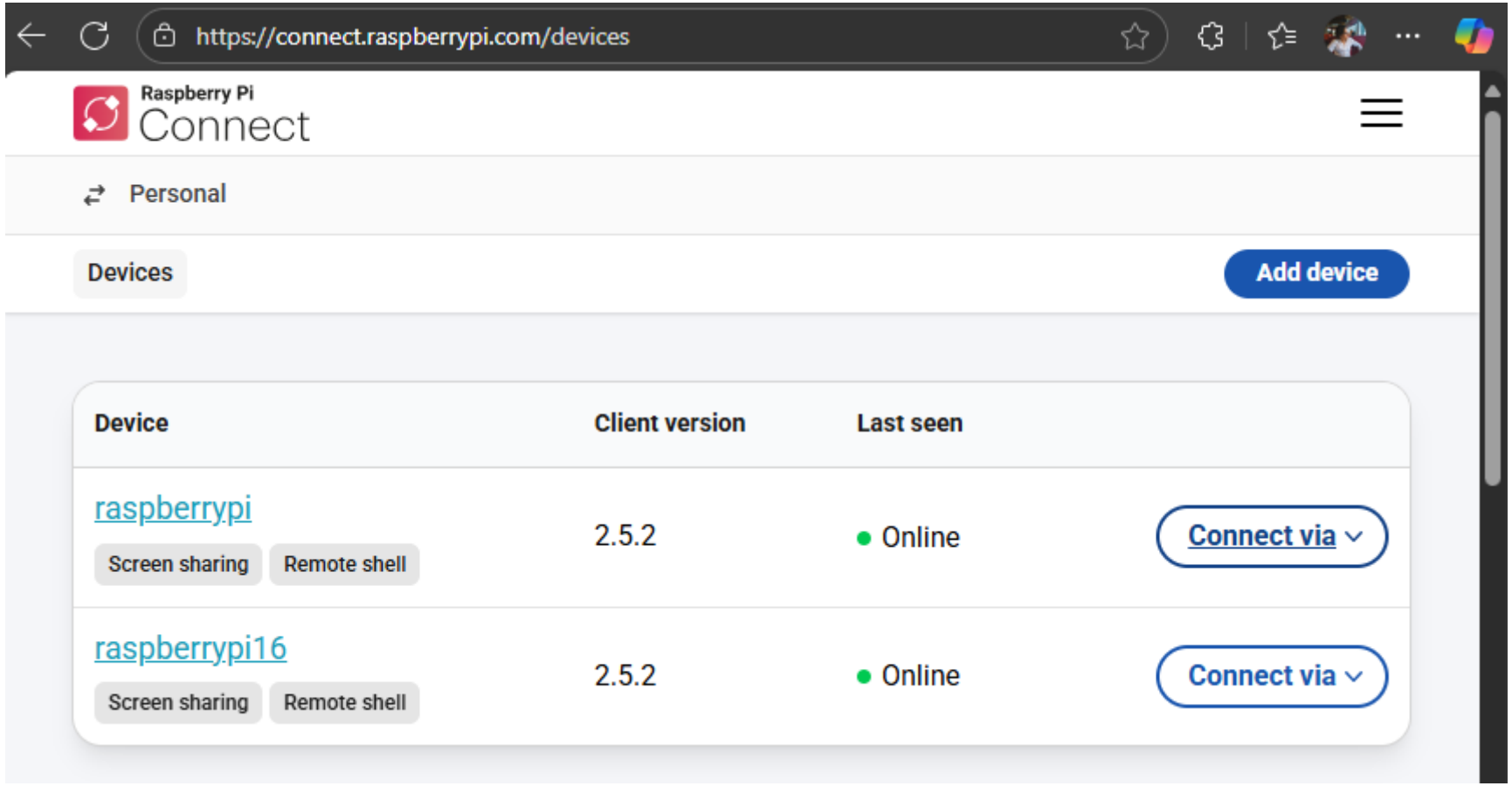
Raspberry Pi Connect gives you free, simple, out-of-the-box access to your Raspberry Pi from anywhere in the world.

Secure remote access solution for Raspberry Pi OS, allowing you to connect to your Raspberry Pi desktop and command line directly from any browser.

Raspberry Pi Connect



Raspberry Pi Connect



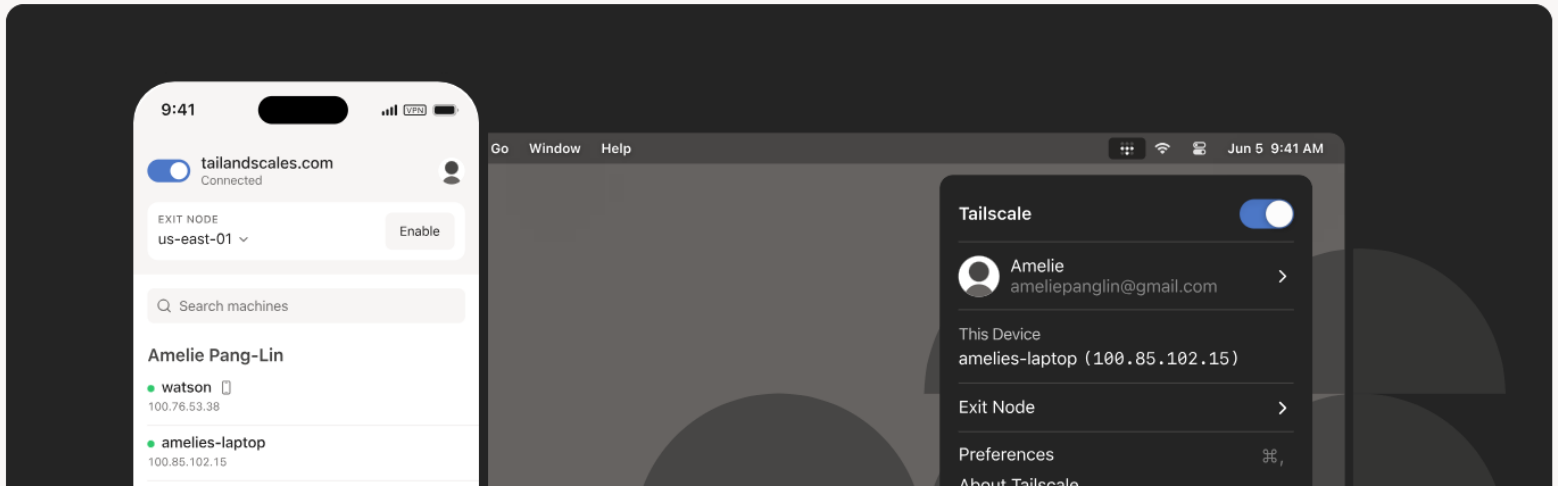
VPN?

tailscale

Your legacy VPN belongs in the past

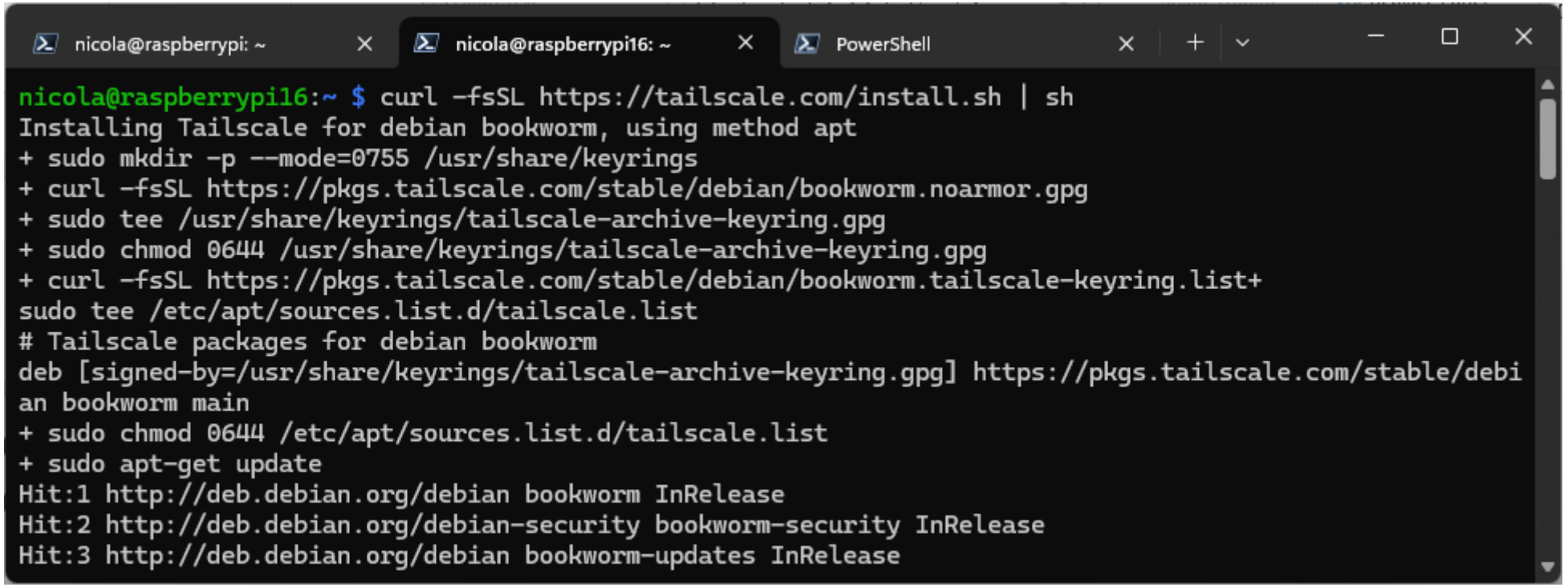
Fast, seamless device connectivity — no hardware, no firewall rules, no wasted time.

Start connecting devices → Contact sales



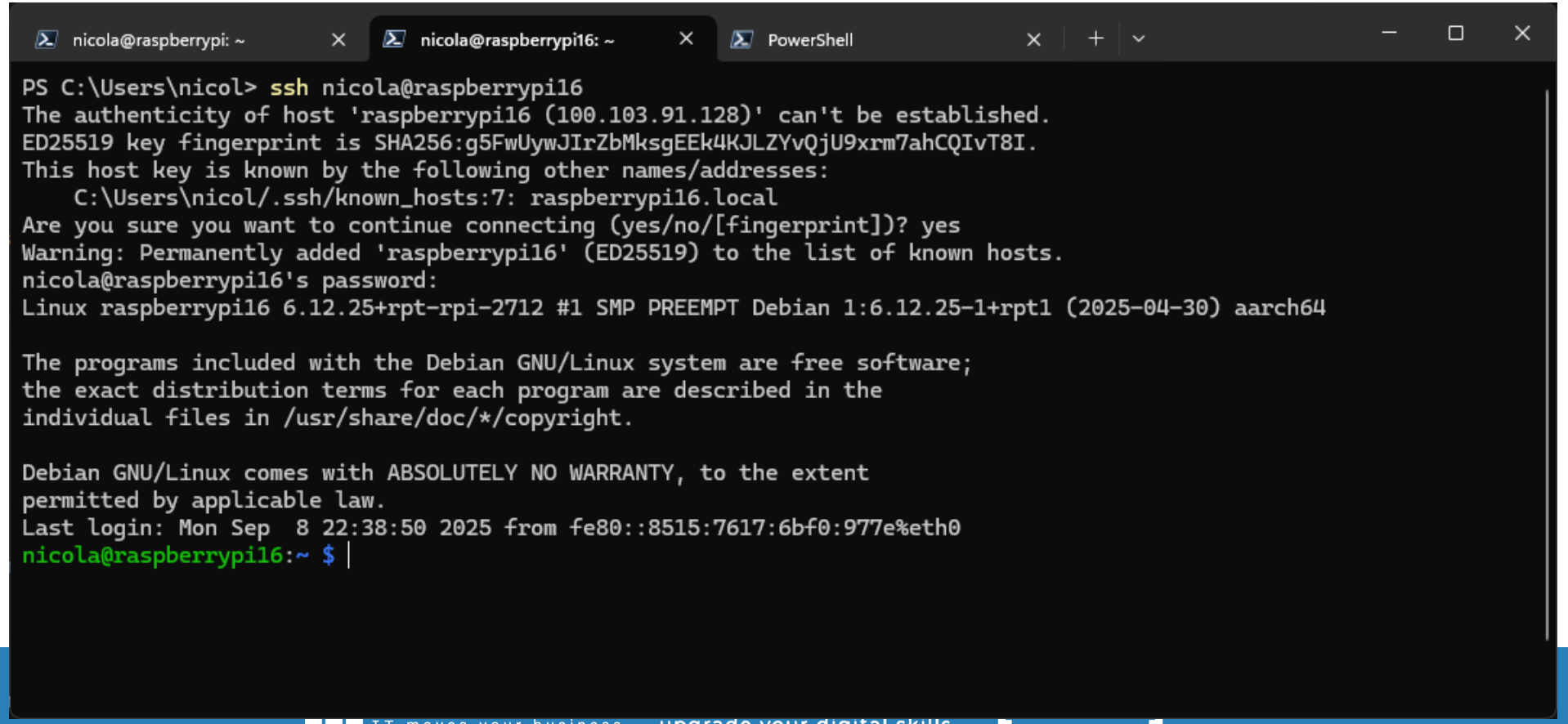
VPN?

tailscale



VPN?

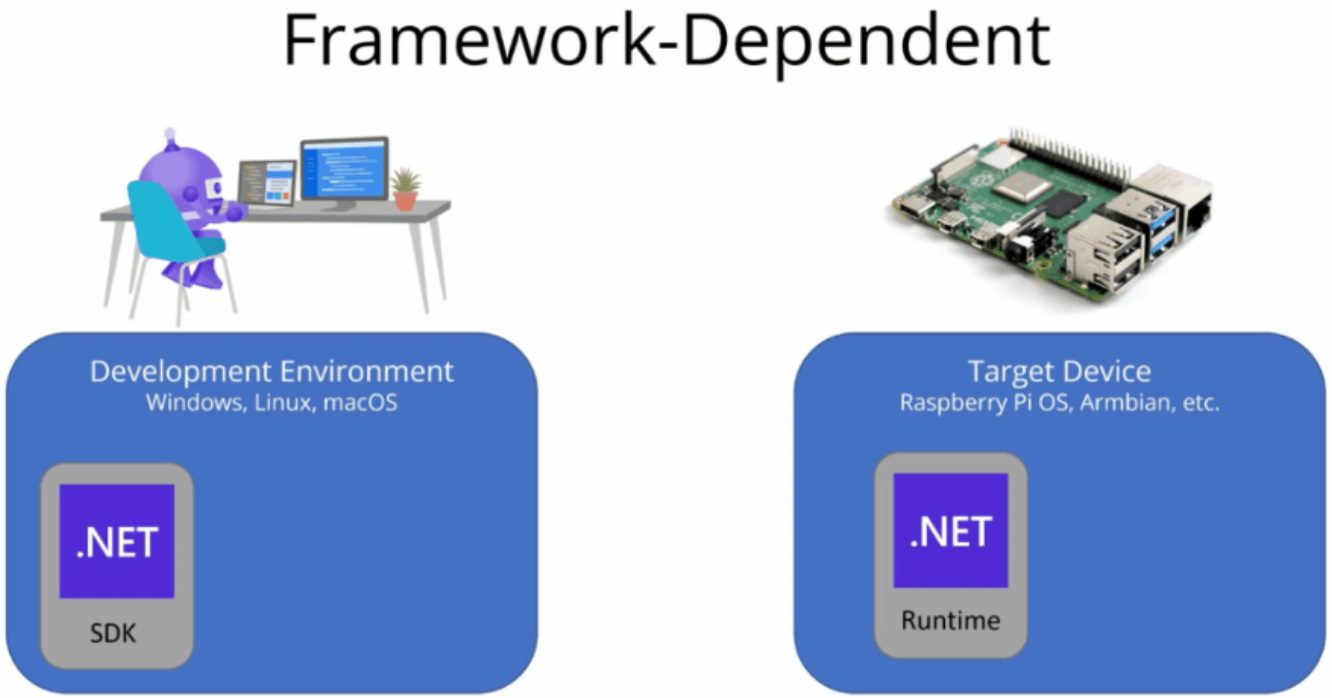
tailscale



Ricapitolando…

- SSH (abilitato di default)
  - Impostazione configurabile tramite Raspberry PI Imager
- RPI Connect
  - Preinstallato con le ultime versioni, installabile tramite apt
  - Connessione «RDP» ed «SSH» alla Raspberry via browser
- Tailscale
  - Zero configuration VPN per creare una rete locale con tutti i device all'interno dei quali è installato

Installiamo .NET

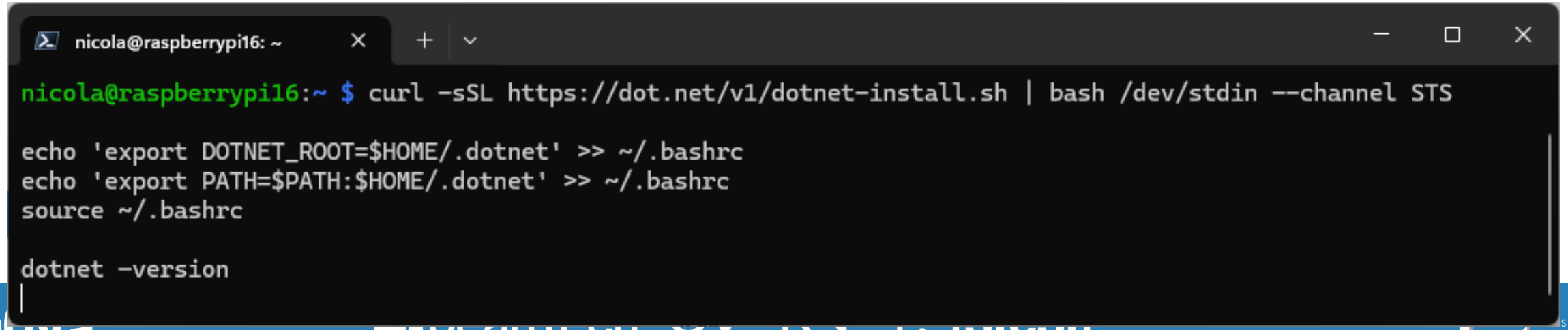


Installiamo .NET

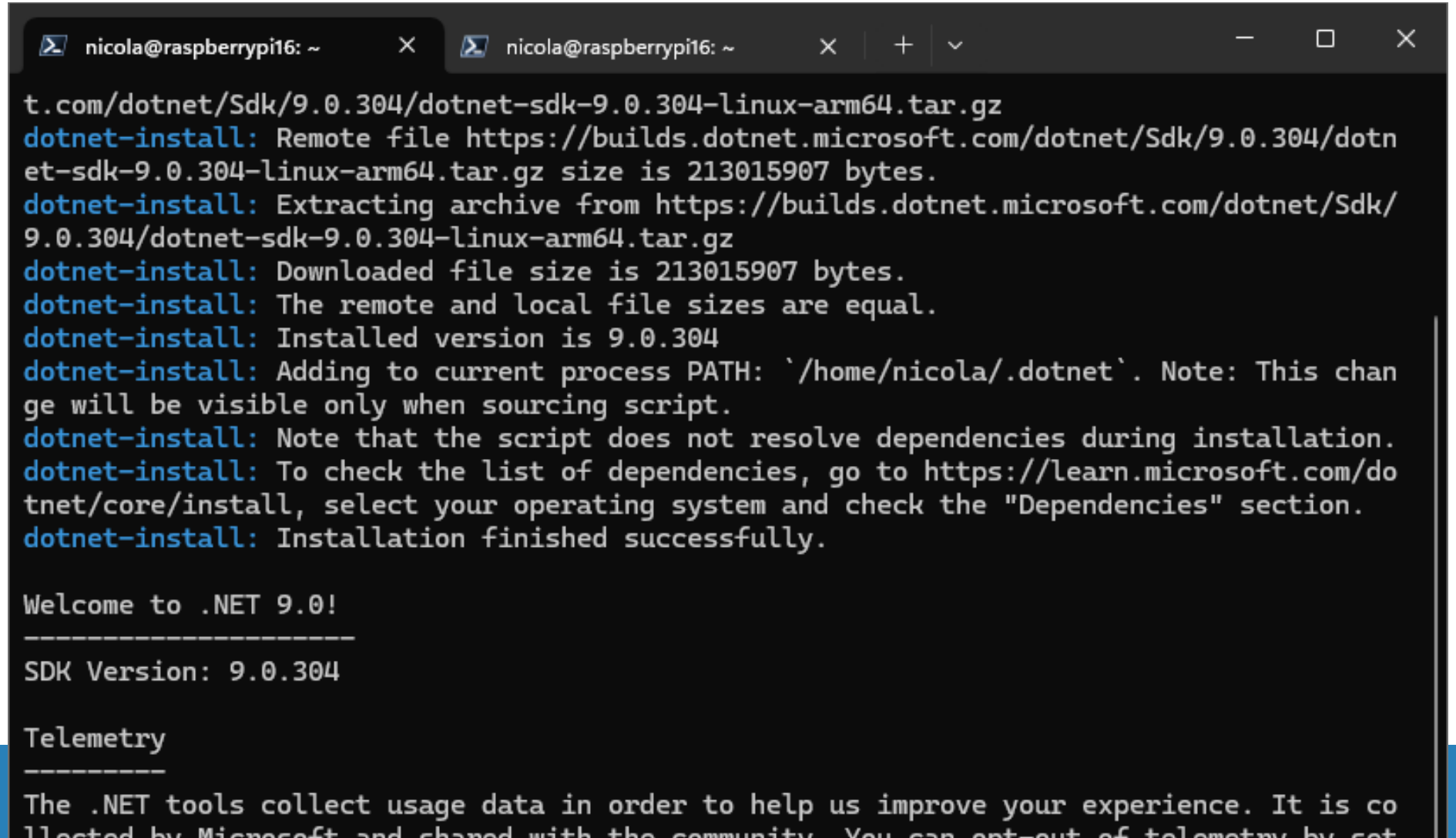
curl -sSL https://dot.net/v1/dotnet-install.sh | bash /dev/stdin --channel STS

echo 'export DOTNET_ROOT=$HOME/.dotnet' >> ~/.bashrc
echo 'export PATH=$PATH:$HOME/.dotnet' >> ~/.bashrc
source ~/.bashrc

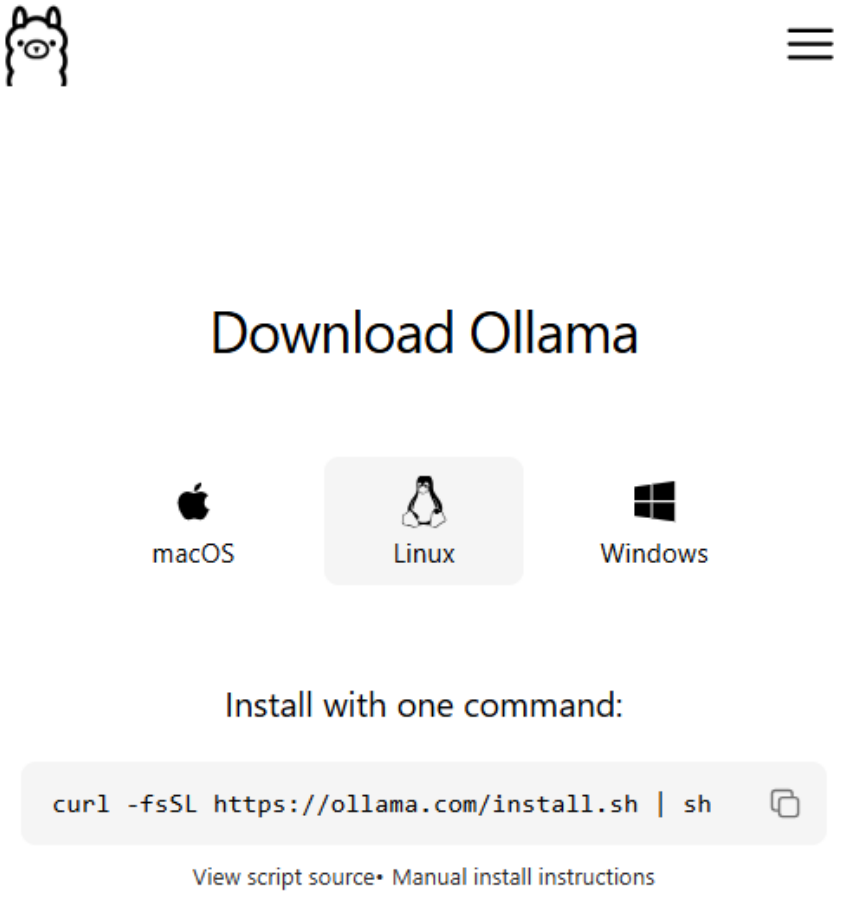
dotnet --version

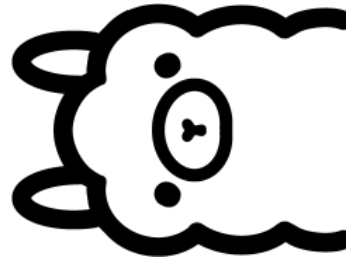


Installiamo .NET



Installiamo Ollama



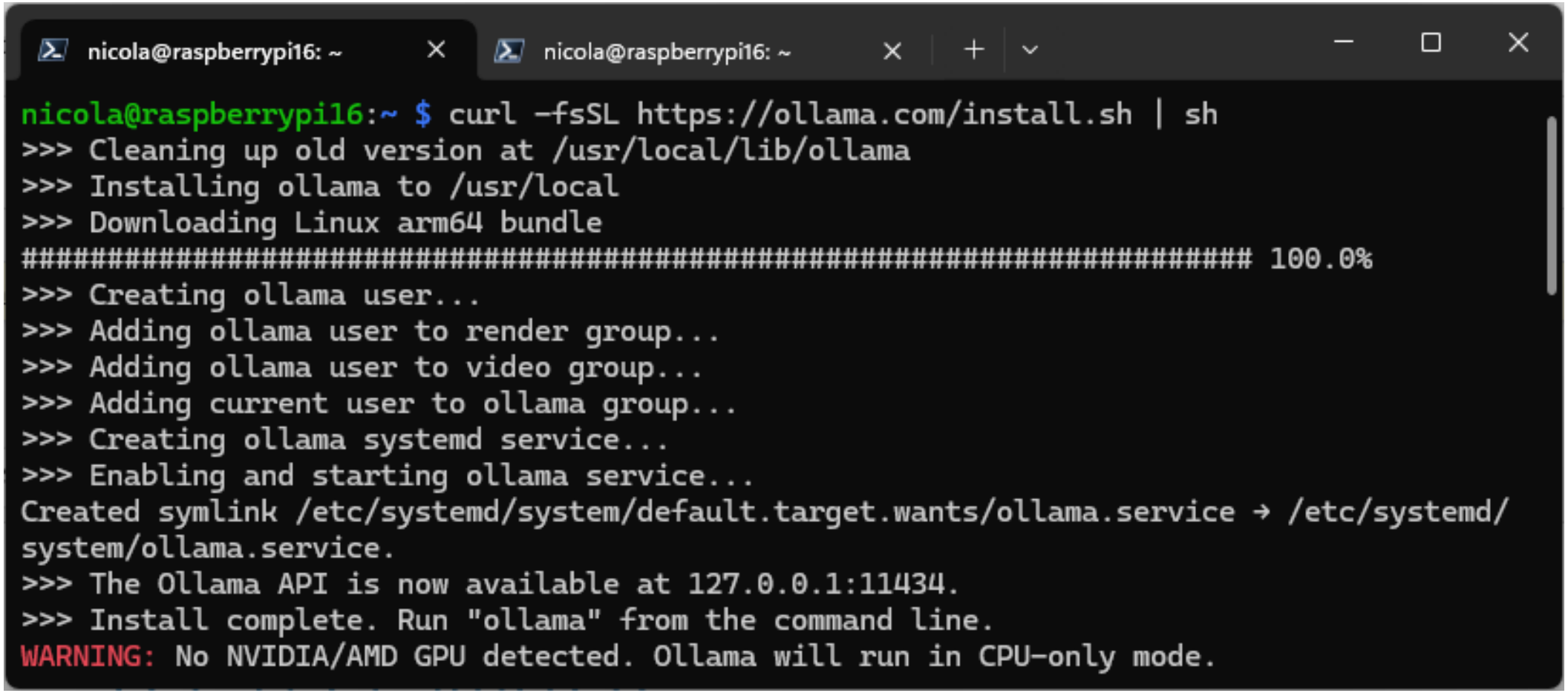


Installiamo Ollama

```
nicola@raspberrypi16:~ $ curl -fsSL https://ollama.com/install.sh | sh
>>> Installing ollama to /usr/local
>>> Downloading Linux arm64 bundle
##############                                                    16.0%
```


One Month
Later...

Installiamo Ollama



Orchestratori

Orchestratori

Permettono di combinare modelli linguistici con codice, strumenti esterni, memoria e flussi logici.

Facilitano la creazione di agenti che possono pianificare, ragionare e agire in modo autonomo.

Offrono librerie e componenti per integrare modelli come GPT, Mistral, Claude, ecc. in app reali.

Orchestratori

Semantic Kernel

LangChain

Orchestratori

		
Linguaggi supportati	C#, Python, Java	Python, JavaScript
Filosofia	Plugin modulari + pianificatori	Catene di chiamate + agenti
Memoria	Embedding + contesto persistente	Memoria conversazionale
Integrazione	Forte con Azure e Microsoft	Ampia con strumenti esterni
Target	Aziende, sviluppatori enterprise	Ricercatori, startup, maker
Use case ideali	Copilot aziendali, workflow complessi	Chatbot, tool AI interattivi

Però …

- Semantic Kernel non è ottimizzato per ambienti con risorse limitate come Raspberry Pi 5.
- Il suo motore di pianificazione e gestione dei plugin può generare prompt complessi e lunghi
- Questo comporta errori di esecuzione, rallentamenti o incompatibilità con modelli ottimizzati per dispositivi ARM.
- Per applicazioni su Raspberry:
  - framework più snelli
  - interfacce dirette con modelli

Orchestratori

Linguaggi supportati	C#, Python, Java	Python, JavaScript
Filosofia	Plugin modulari + pianificatori	Catene di chiamate + agenti
Memoria	Embedding + contesto persistente	Memoria conversazionale
Integrazione	Forte con Azure e Microsoft	Ampia con strumenti esterni
Target	Aziende, sviluppatori enterprise	Ricercatori, startup, maker
Use case ideali	Copilot aziendali, workflow complessi	Chatbot, tool AI interattivi

La soluzione?

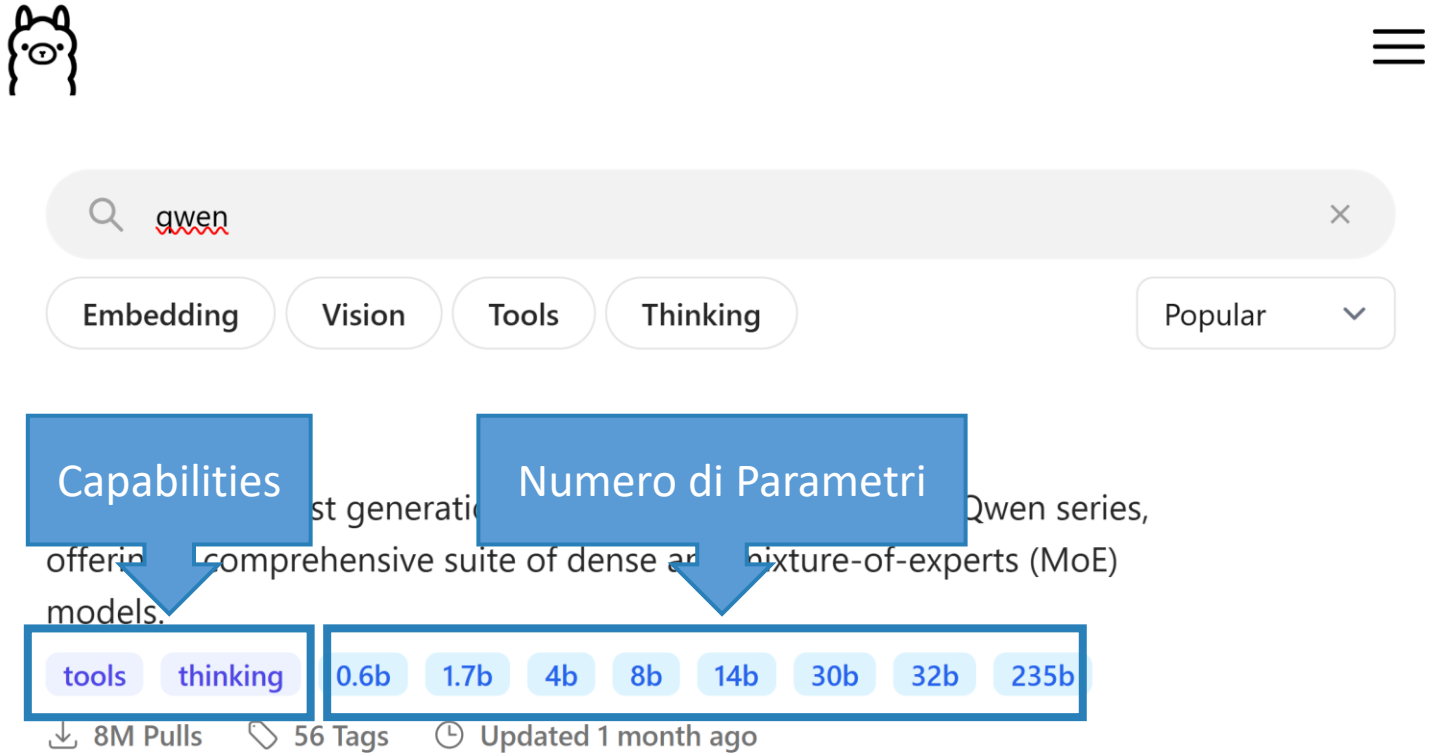
Metodo Coleman

Co’ le man me lo faccio da solo [cit. Massimo Bonanni]

Scelta dei modelli

Scelta dei modelli

https://ollama.com/search



Scelta dei modelli – Numero di Parametri

Pochi Parametri

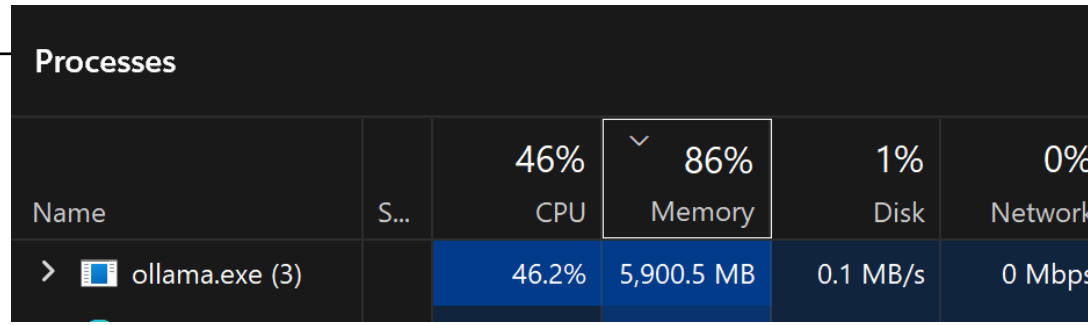
Tanti Parametri

- Velocità di risposta
- Allucinazioni

- Accuratezza della risposta
- Consumo di energia
- Consumo di memoria

Scelta dei modelli – Numero di Parametri

Ollama, Ollama	Yes, papa
Eating RAM?	No, papa
Telling lies?	No, papa
Open your mouth	



Scelta dei modelli – Numero di Parametri

Name	Size	Context	Input
qwen3:latest	5.2GB	40K	Text
qwen3:0.6b	523MB	40K	Text
qwen3:1.7b	1.4GB	40K	Text
qwen3:4b	2.5GB		
qwen3:8b latest	5.2GB		
qwen3:14b	9.3GB		
qwen3:30b	19GB	256K	Text
qwen3:32b	20GB	40K	Text
qwen3:235b	142GB	256K	Text

Tutto il modello viene caricato in RAM

Scelta dei modelli – Capabilities

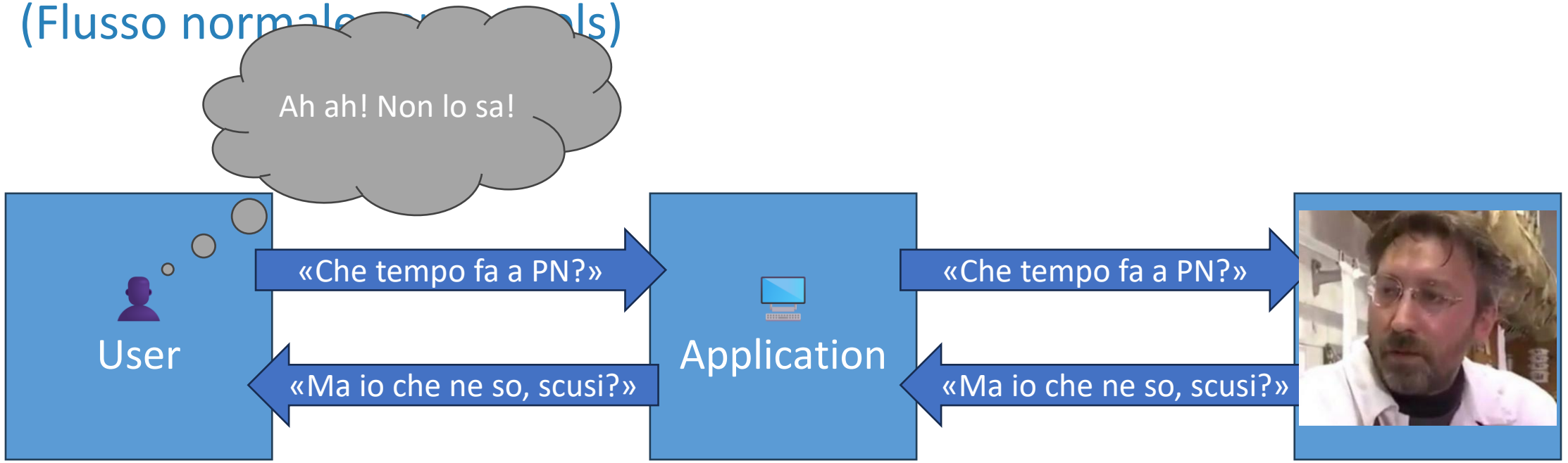
- Embedding
  - Modelli che generano vettori numerici per rappresentare testi, utili in ricerca semantica, raccomandazioni e clustering.
- Vision
  - Modelli capaci di analizzare immagini, riconoscere contenuti visivi e combinarli con testo.
- Thinking
  - Modelli progettati per un ragionamento più accurato e passo-per-passo, ottimizzati per compiti complessi.
- Tools
  - Modelli che possono usare strumenti esterni (API, funzioni, calcoli) oltre al testo.

Cosa sono i Tools?

- Un tool è una funzionalità esterna che il modello può attivare per svolgere compiti specifici che vanno oltre la semplice generazione di testo.
- I tools permettono al modello di interagire con il mondo esterno o di eseguire operazioni complesse che richiedono capacità specifiche.

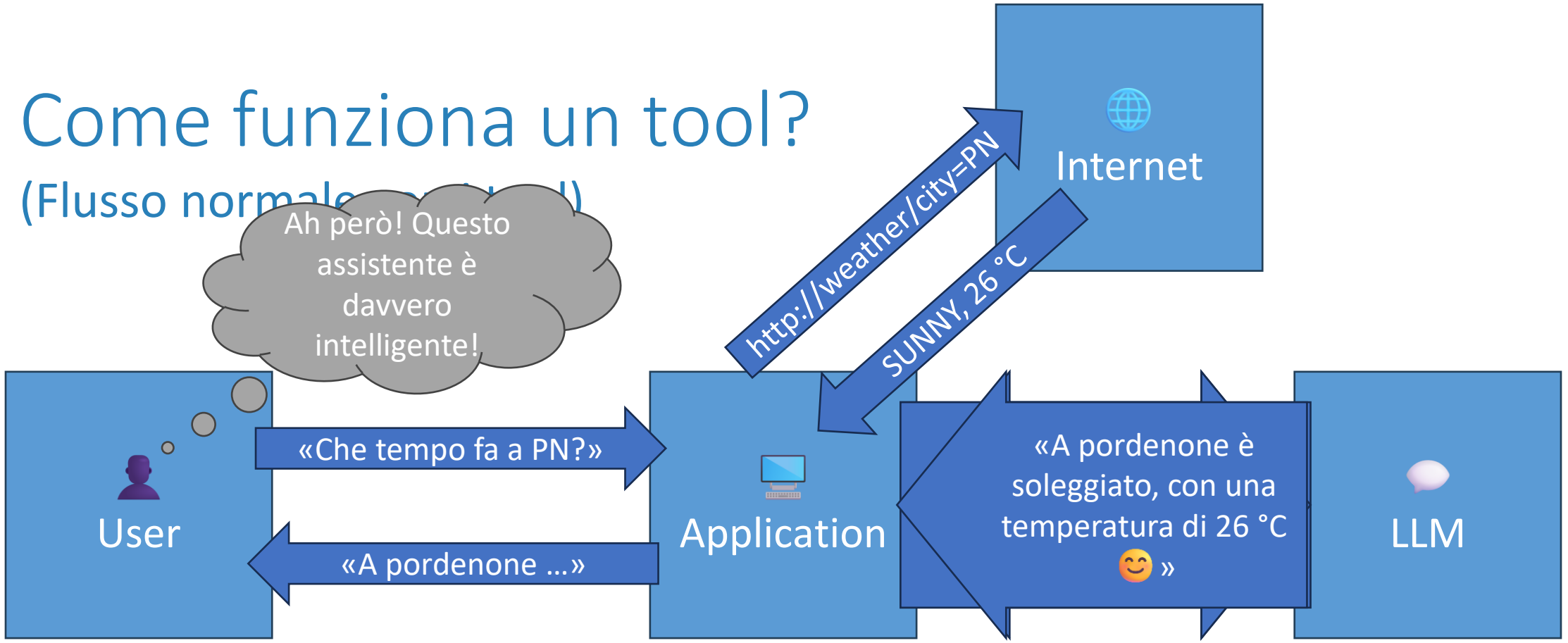
Come funziona un tool?

(Flusso normale [illegible]ols)



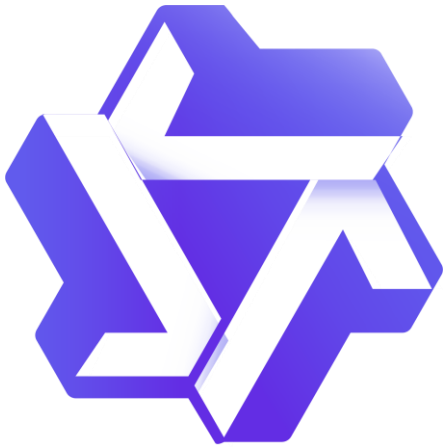
Come funziona un tool?

(Flusso normale con tool)

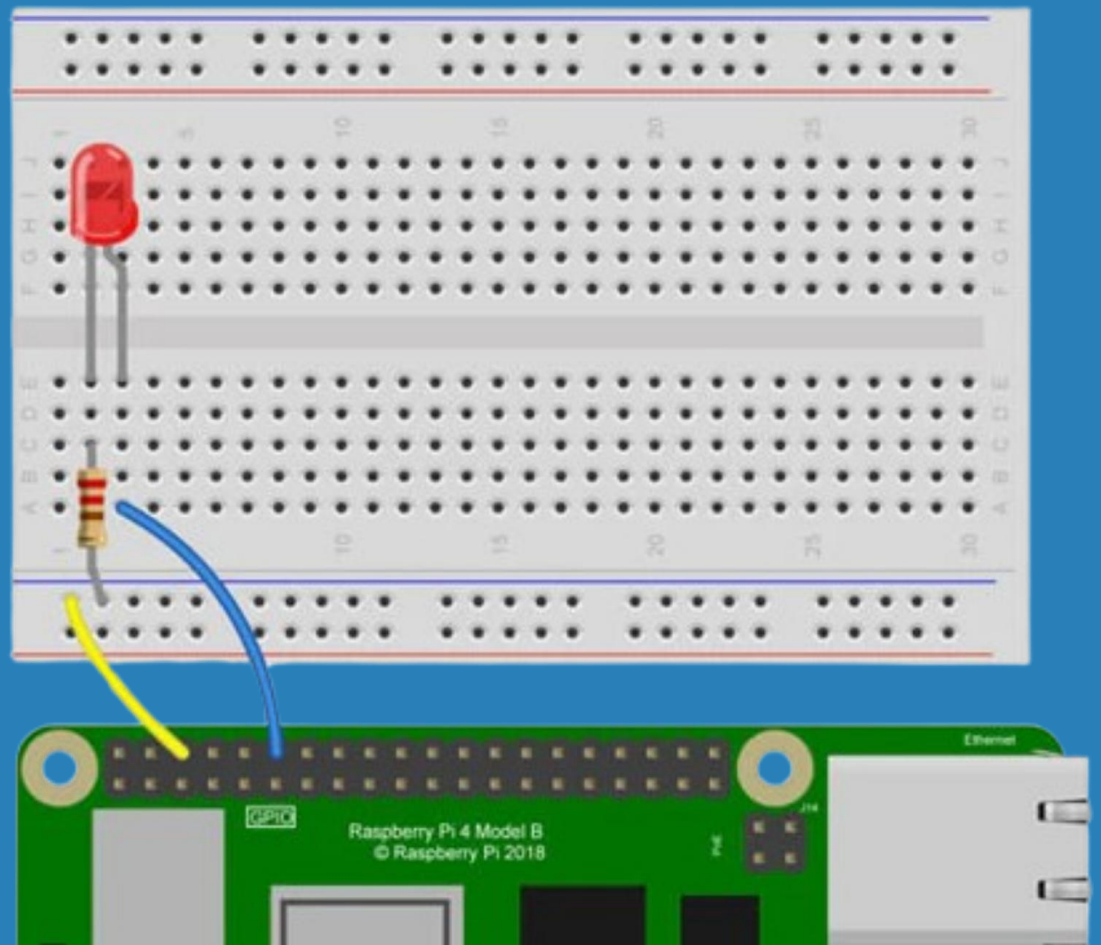


Quindi che modello scelgo?

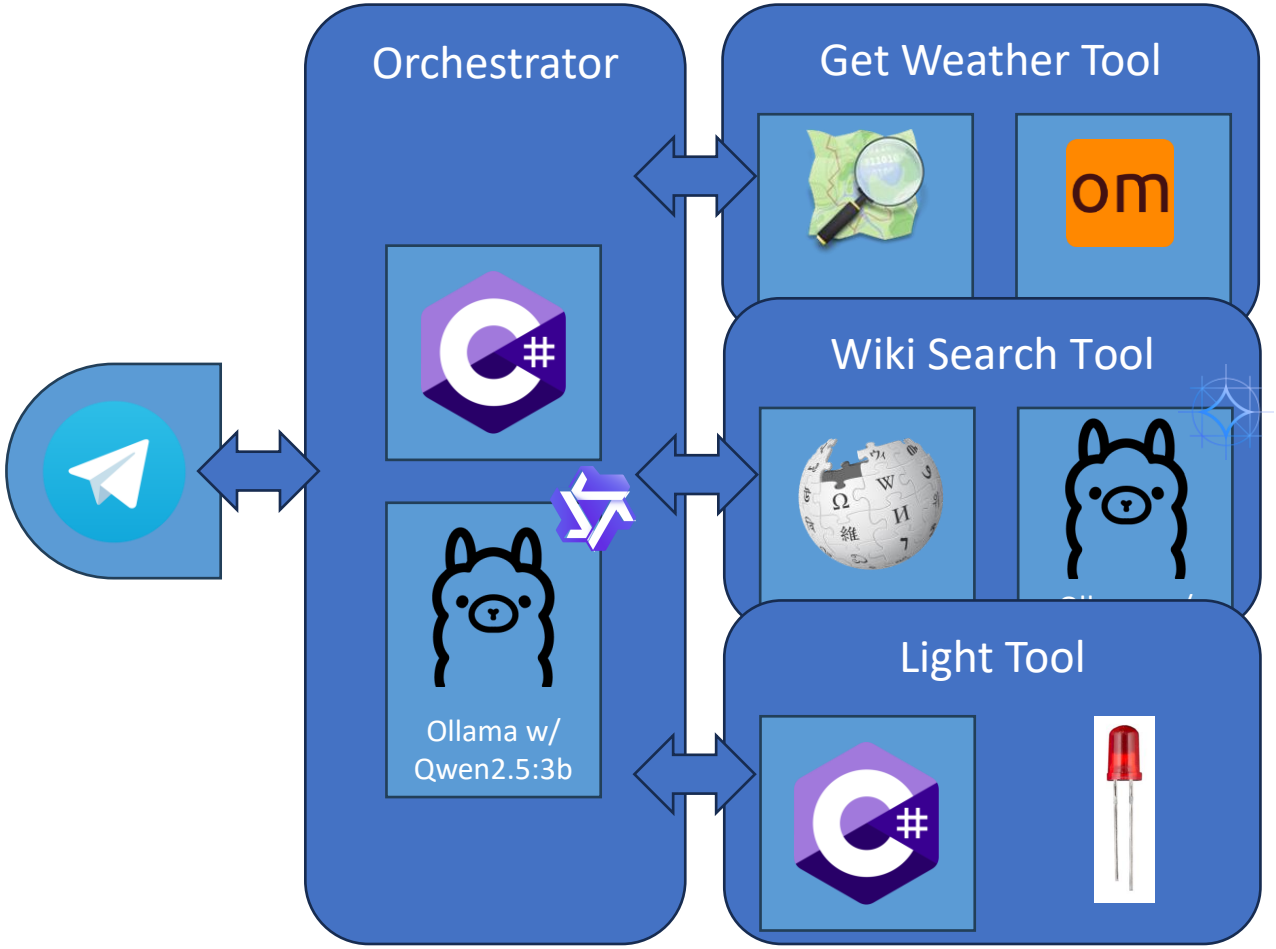
- Llama3.1 o 3.2 ❌
  - Dalla mia personale esperienza, i modelli llama3.1 e llama3.2 hanno un forte bias nei confronti dei tool, preferendone l'invocazione anche se non sono realmente richiesti.
- Gemma 3 ❌
  - Qualità nella risposta, ma non supporta i tools
- Qwen2.5:3b ✔
  - Modello leggero e abbastanza preciso per le sue dimensioni
  - Le invocazioni dei tool avvengono tendenzialmente quando sono realmente necessarie



DEMO



La soluzione implementata



Conclusioni

Conclusioni

- Il «fai da te» da sempre le sue soddisfazioni
  - Ho imparato ad avere a che fare con un ambiente più «ostile» rispetto al mio use case solito: risorse limitate, surriscaldamento della scheda
- Anche se sono un informatico, le lezioni di elettronica del liceo mi sono servite
- Mettere un LLM in un ambiente così limitato, come la Raspberry PI, non ha molto senso

Piuttosto…

GMKtec EVO-X2 Mini PC AI AMD Ryzen AI Max+ 395 (fino a 5,1GHz) Mini PC Gaming, 128GB LPDDR5X 8000MHz (16GB*8), SSD PCIe 4.0 da 2TB, Display 8K a Quattro Schermi, WiFi 7&USB4, Lettore di Schede SD 4.0

4,4 ★★★★½ (251)

2.799[96] €

prime

Resi GRATUITI

I prezzi degli articoli in vendita su Amazon includono l'IVA. In base all'indirizzo di spedizione, l'IVA potrebbe variare durante il processo di acquisto. Per maggiori informazioni clicca qui.

Applica | Risparmio Compra 1 per €2099.96 Termini

Paga a rate a **tasso zero** con Cofidis. Fino al 30/09/2025.
Scopri di più

Taglia: **EVO X2-AI MAX 390-128+2T**

EVO X2-AI MAX 390-...	K8 Plus 8845HS...	K8 Plus-8845HS-...	K11-8945HS-...
2.799,96€	629,96€	799,96€	889,96€
Consegna GRATIS lunedì	Consegna GRATIS giovedì	Consegna GRATIS martedì	Consegna GRATIS martedì

Fai clic per visualizzare la vista completa

Chiedi a Rufus

Qual è la capacità di archiviazione?

Che tipo di processore ha? | Supporta il Wi-Fi 7?

Chiedi qualcos'altro

MINISFORUM AI X1 Pro Mini PC, AMD Ryzen AI 9 HX370 (12 C/24 T, fino a 5,1 GHz) e AMD Radeon 890M, 96 GB DDR5, 2 TB PCIe 4.0 SSD, display quadruplo 4K, doppia LAN 2.5/WiFi 7/BT 5.4/Oculink

Visita lo Store di MINISFORUM

4,5 ★★★★½ (20)

Scelta Amazon

1.255[00] €

prime Un giorno

Resi GRATUITI

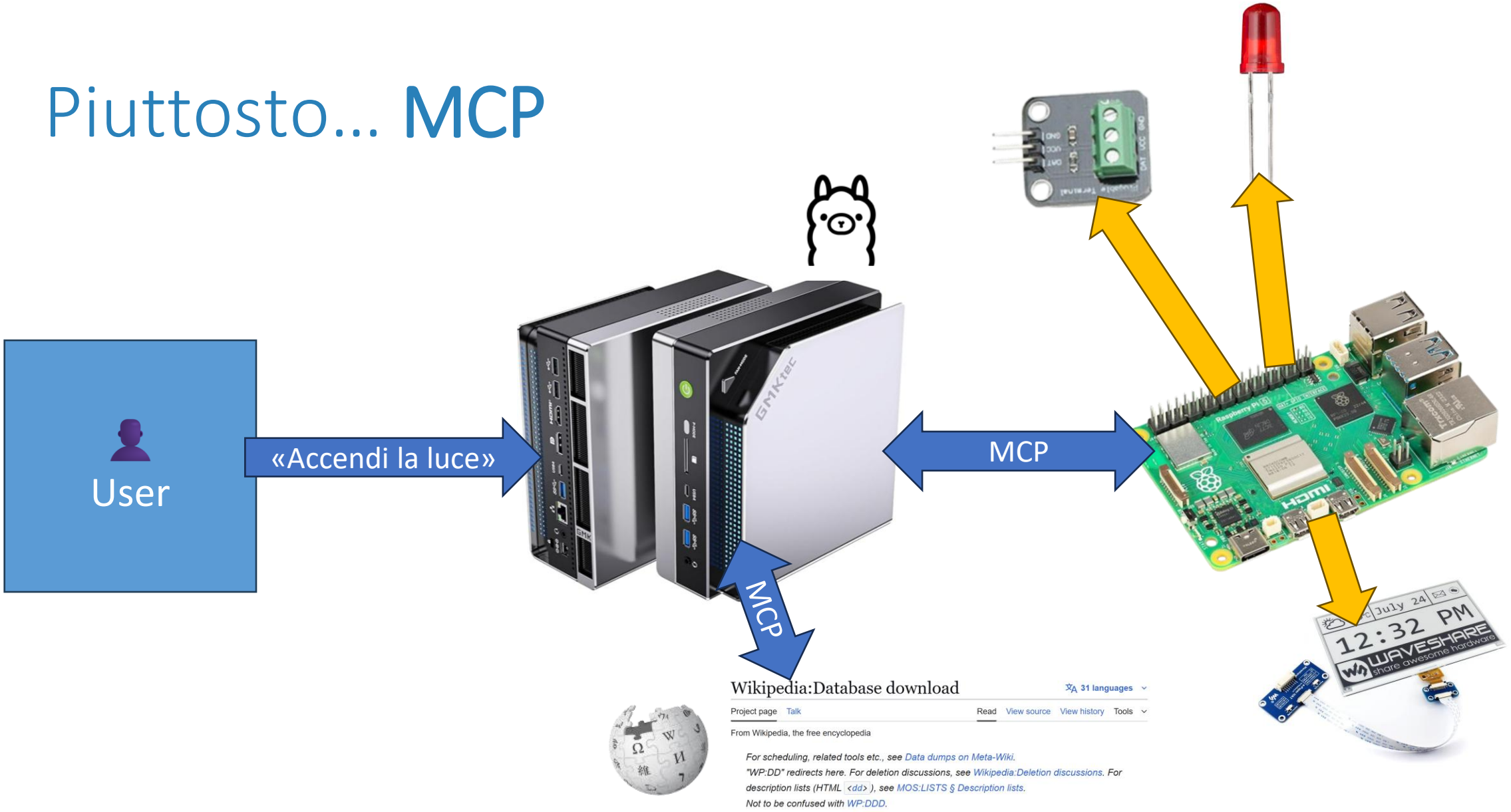
I prezzi degli articoli in vendita su Amazon includono l'IVA. In base all'indirizzo di spedizione, l'IVA potrebbe variare durante il processo di acquisto. Per maggiori informazioni clicca qui.

Paga a rate a **tasso zero** con Cofidis. Fino al 30/09/2025.
Scopri di più

Taglia: **X1 Pro-370 96/2TB**

M1Pro-125H 0/0 GB	M1Pro-125H 32/1TB	M1 Pro-285H...	X1 Pro-370 32/1TB
415,00€	583,00€	1.119,99€ ~~1.399,99€~~	999,00€
	Consegna GRATIS		Consegna GRATIS

Piuttosto… MCP



GRAZIE!

About me

Nicola Paro

Solutions Architect @ beanTech

.NET & Azure Meetup Štajerska Community Lead

codice

https://www.linkedin.com/in/nicolaparo/