



Ollama: Local AI for my Raspberry Pi

Nicola Paro

SPONSOR



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»

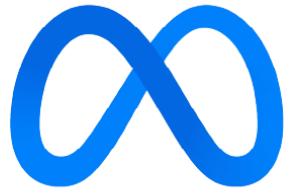
Perché voglio utilizzare un LLM Locale?

E' davvero per la privacy

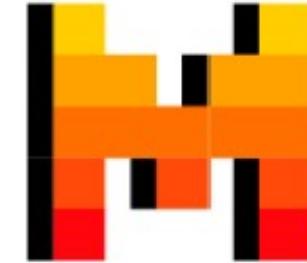
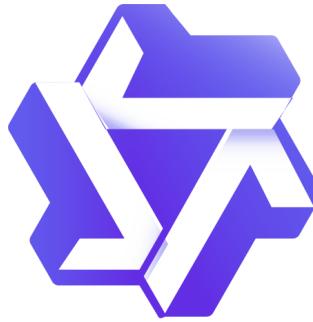
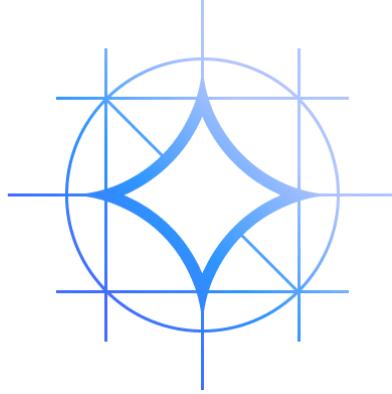
I miei dati sono su un server interno.
I miei clienti usano server interni dedicati anche loro.

Lavoro in scenari disconnessi dal cloud

Che modelli Open Source posso usare in locale?



Llama Meta



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

Aggiornamenti frequenti,
documentazione chiara.

API REST semplici
(localhost:11434).

Community enorme (soprattutto
su macOS, ora anche Linux e
Windows).

LM Studio

Interfaccia desktop molto popolare
per chi vuole un'alternativa
“ChatGPT offline”.

API OpenAI-compatible, quindi
integrabile ovunque senza
cambiare librerie.

Ampio uso in community e progetti
personalì/produttivi.

Foundry Local

Nuovo, ma in rapida crescita: non
ancora “di massa” come Ollama,
ma più solido in ottica enterprise.

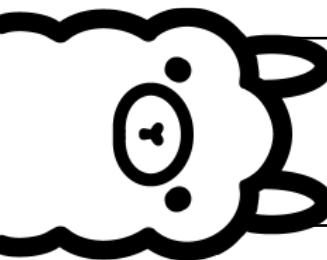
API REST e SDK OpenAI-
compatible.

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**.



[Discord](#) [GitHub](#) [Models](#)

Search models

[Sign in](#)

[Download](#)



Get up and running with large language models.

Run [DeepSeek-R1](#), [Qwen 3](#), [Llama 3.3](#), [Qwen 2.5-VL](#), [Gemma 3](#), and other models, locally.

[Download ↓](#)

Available for macOS, Linux, and Windows

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  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.

Raspberry Pi 5 8GB Quad-Core ARMA76 (64 Bits - 2,4 GHz)
4,7 ★★★★★ (1842)
100+ acquistati nel mese scorso
95⁵⁷ €
✓ 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)
149⁹⁹ €
Pagamento a rate disponibile
Paghi 142,49 € con coupon
✓ prime
Consegna senza costi aggiuntivi **mer, 10 set**
[Aggiungi al carrello](#)

Hardware

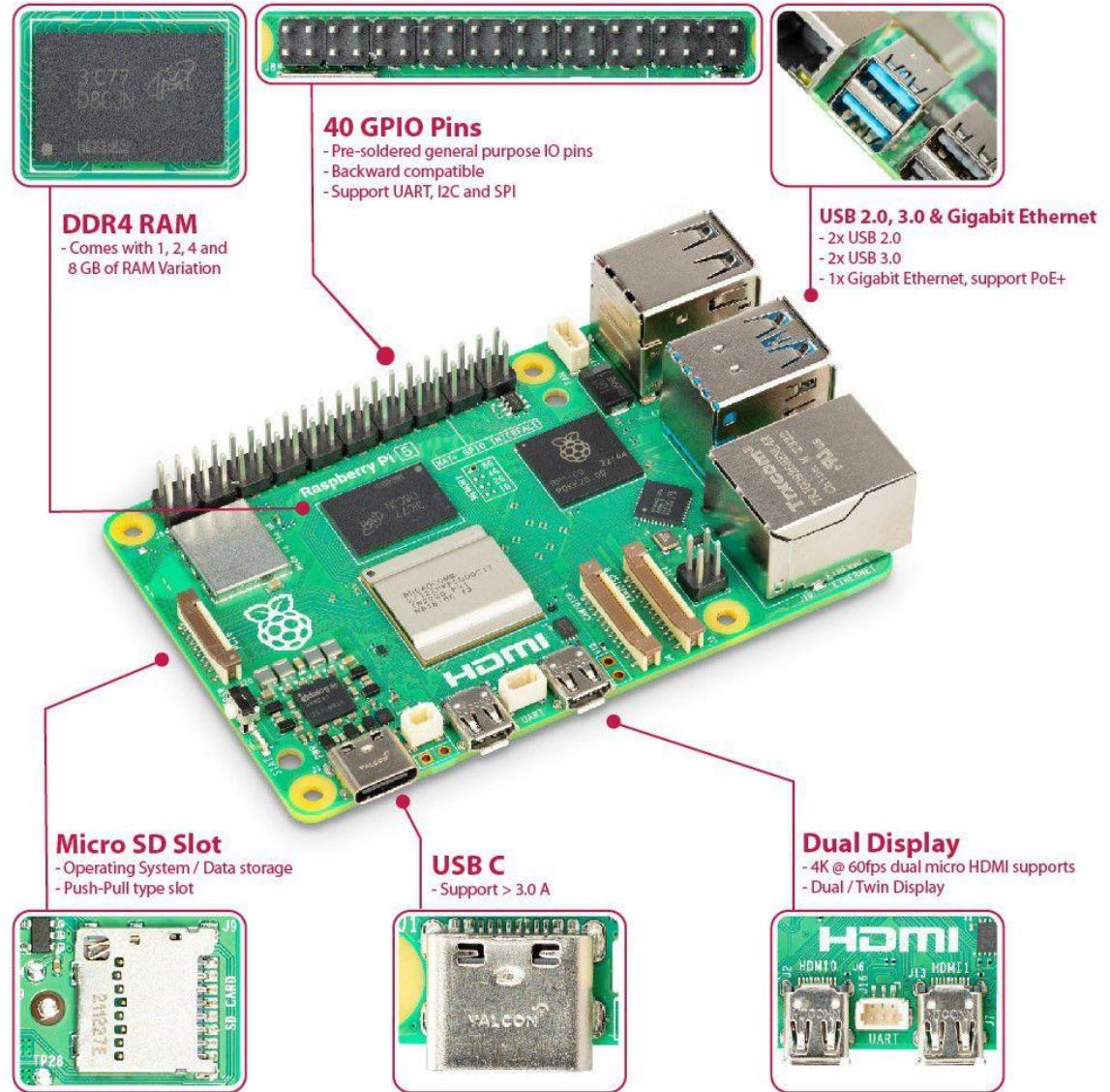
- Raspberry PI 5

Praticamente è un mini pc, ma con le GPIO

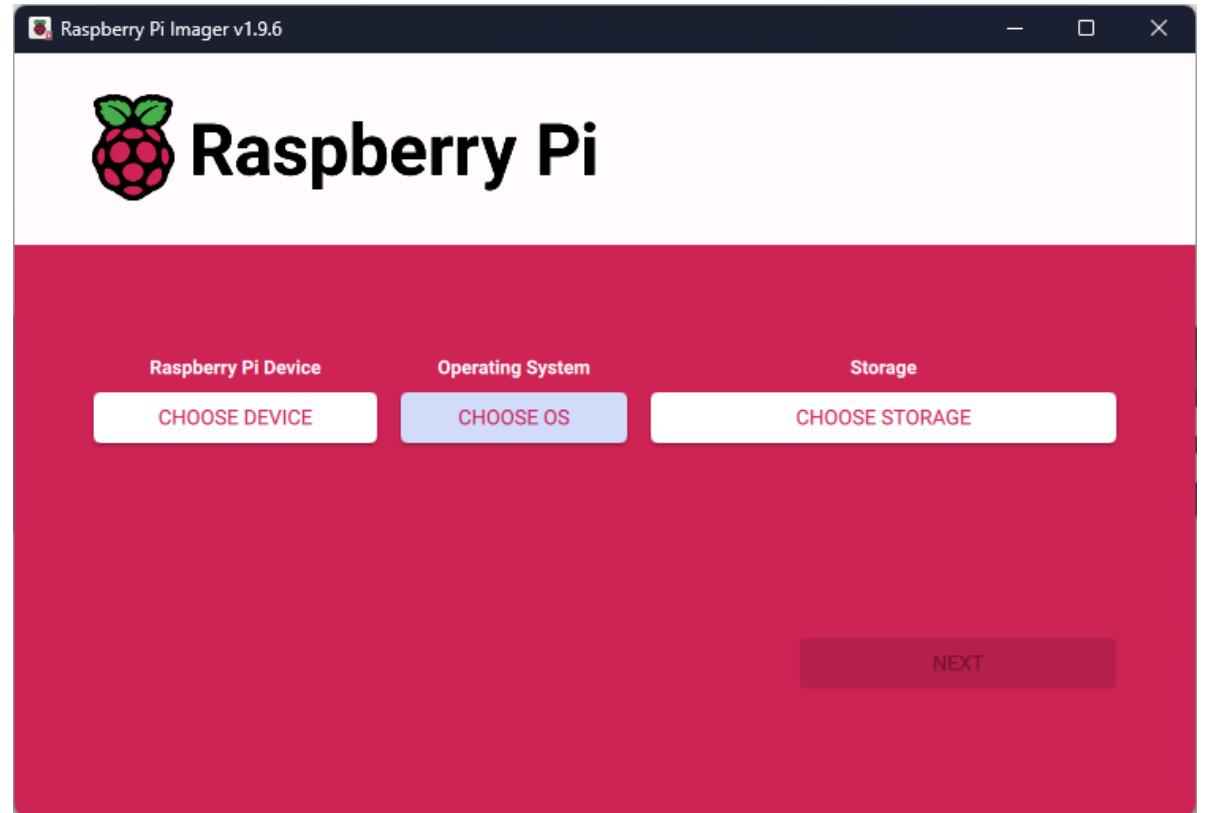
- 64bit ARM CPU
- 16GB RAM

- 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':

RUN SSH-KEYGEN **ADD SSH KEY** **DELETE KEY**

CANCEL **SAVE**

OS Customisation

General Services Options

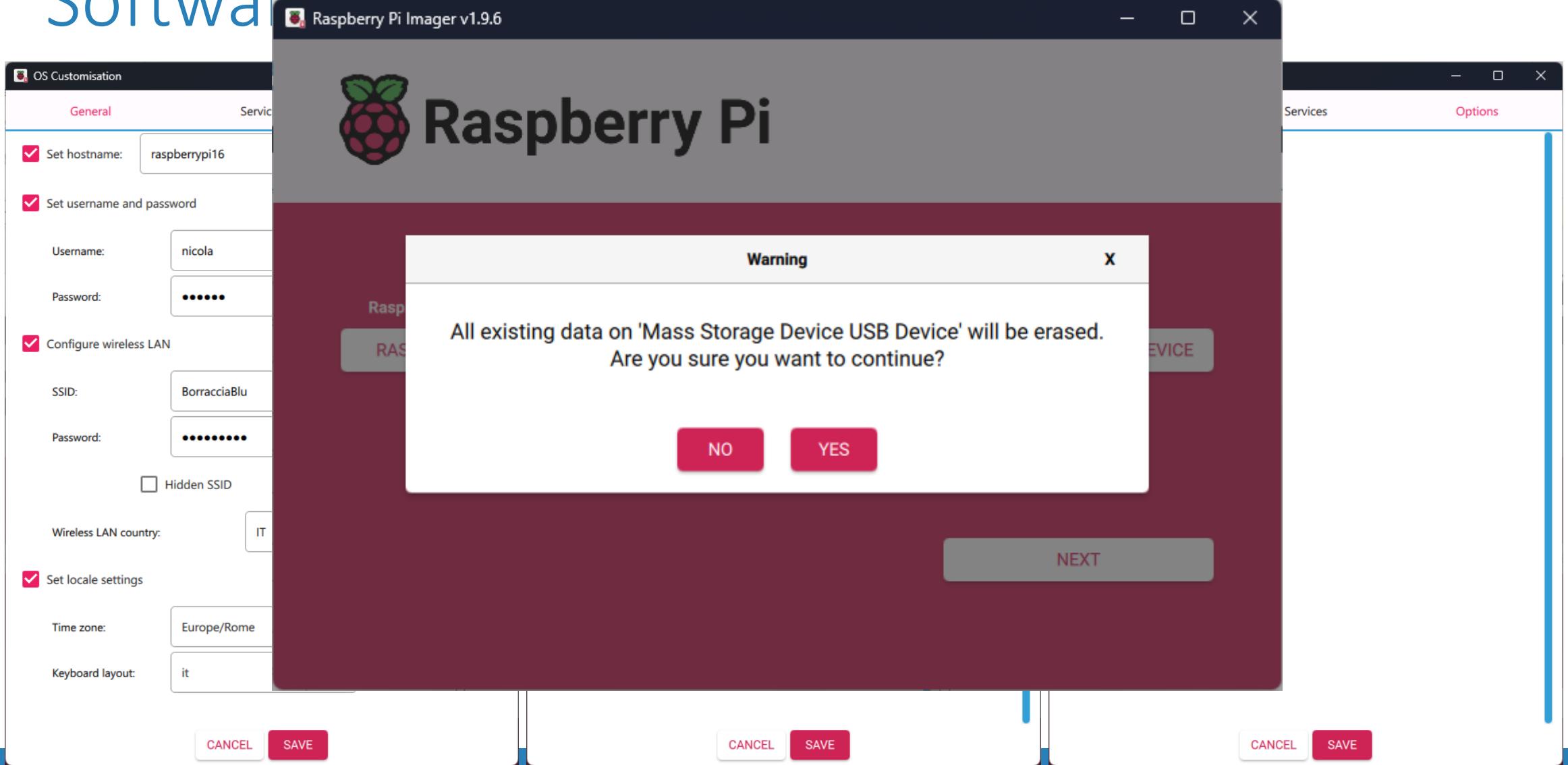
Play sound when finished

Eject media when finished

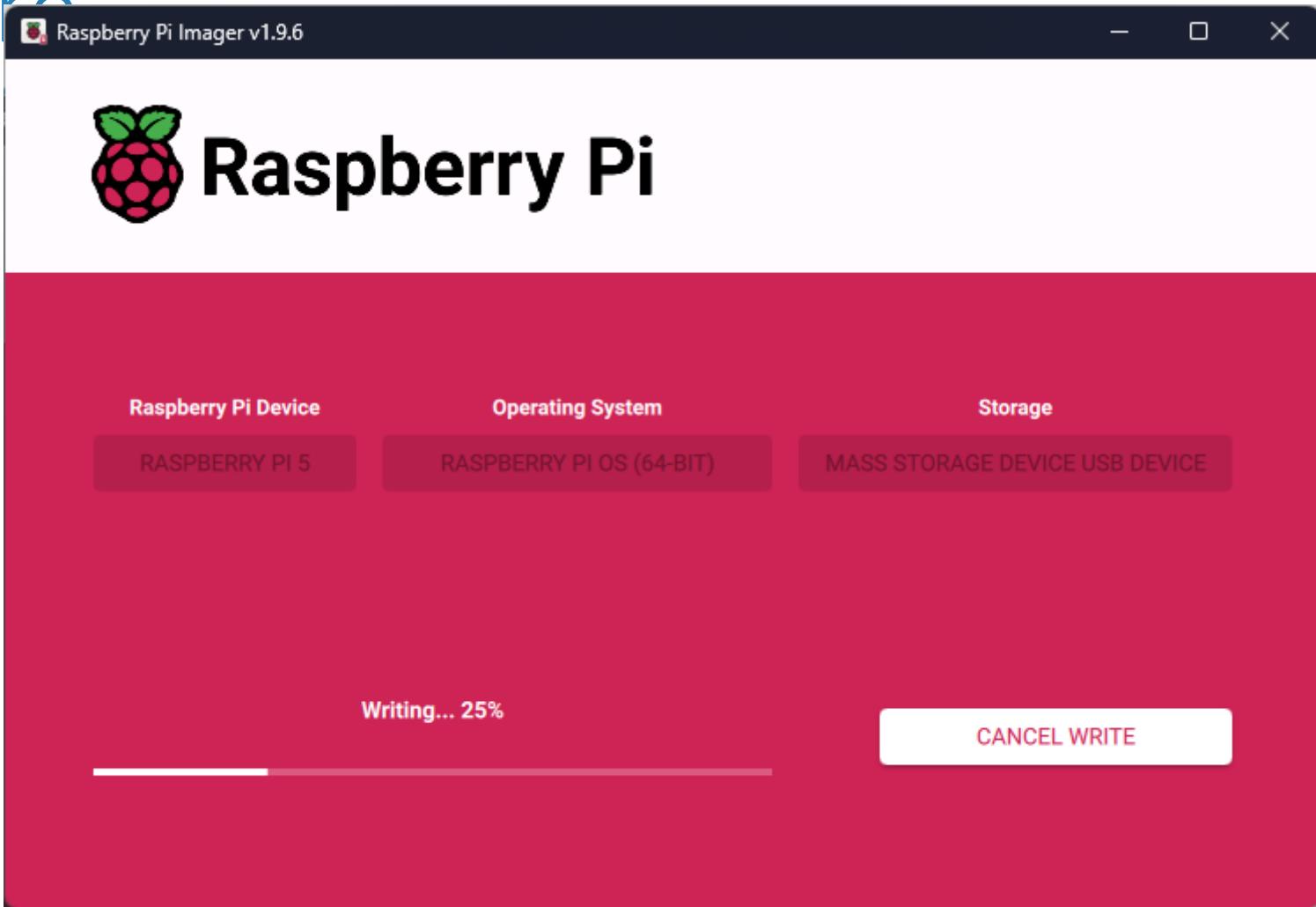
Enable telemetry

CANCEL **SAVE**

Software

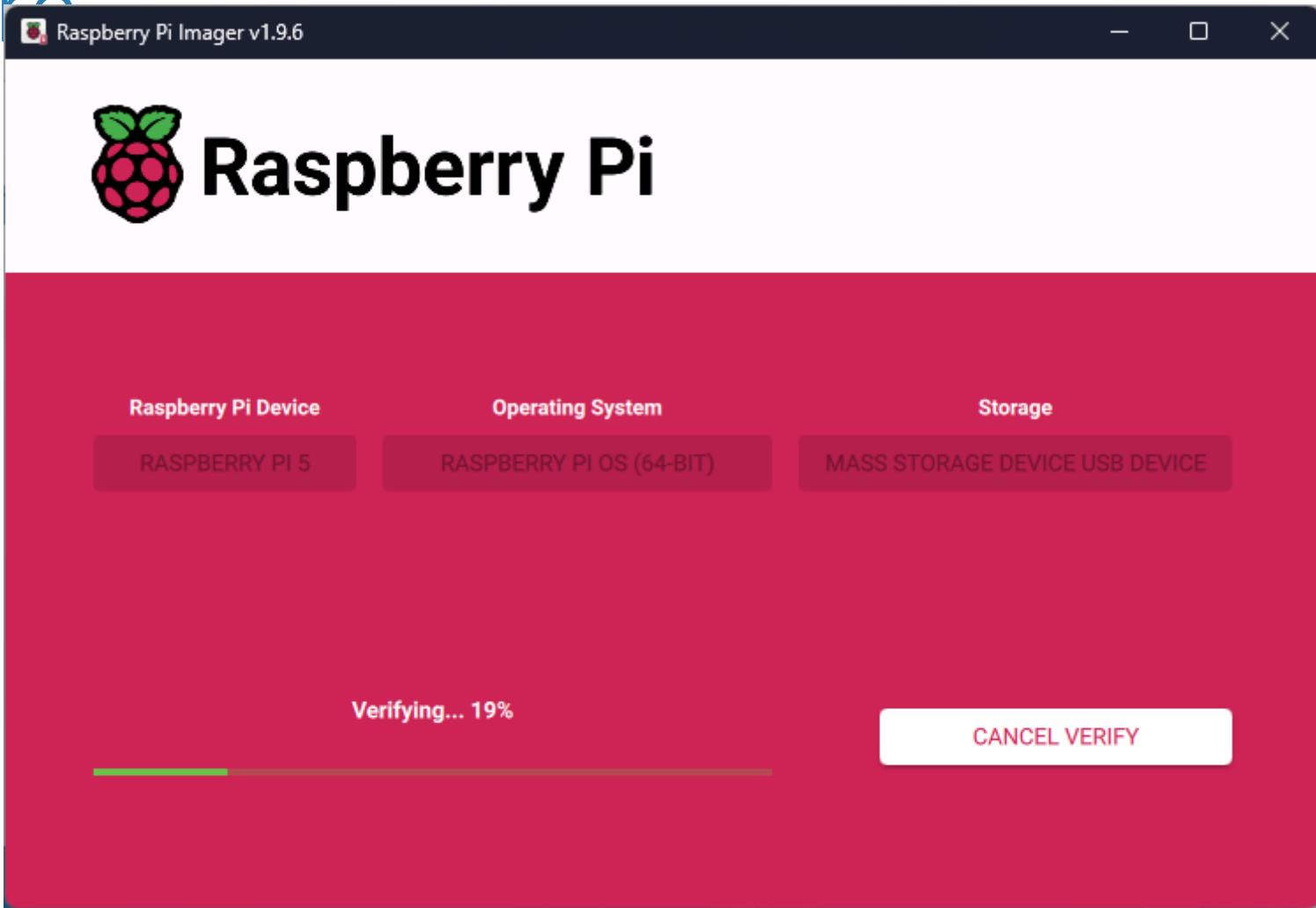


Software



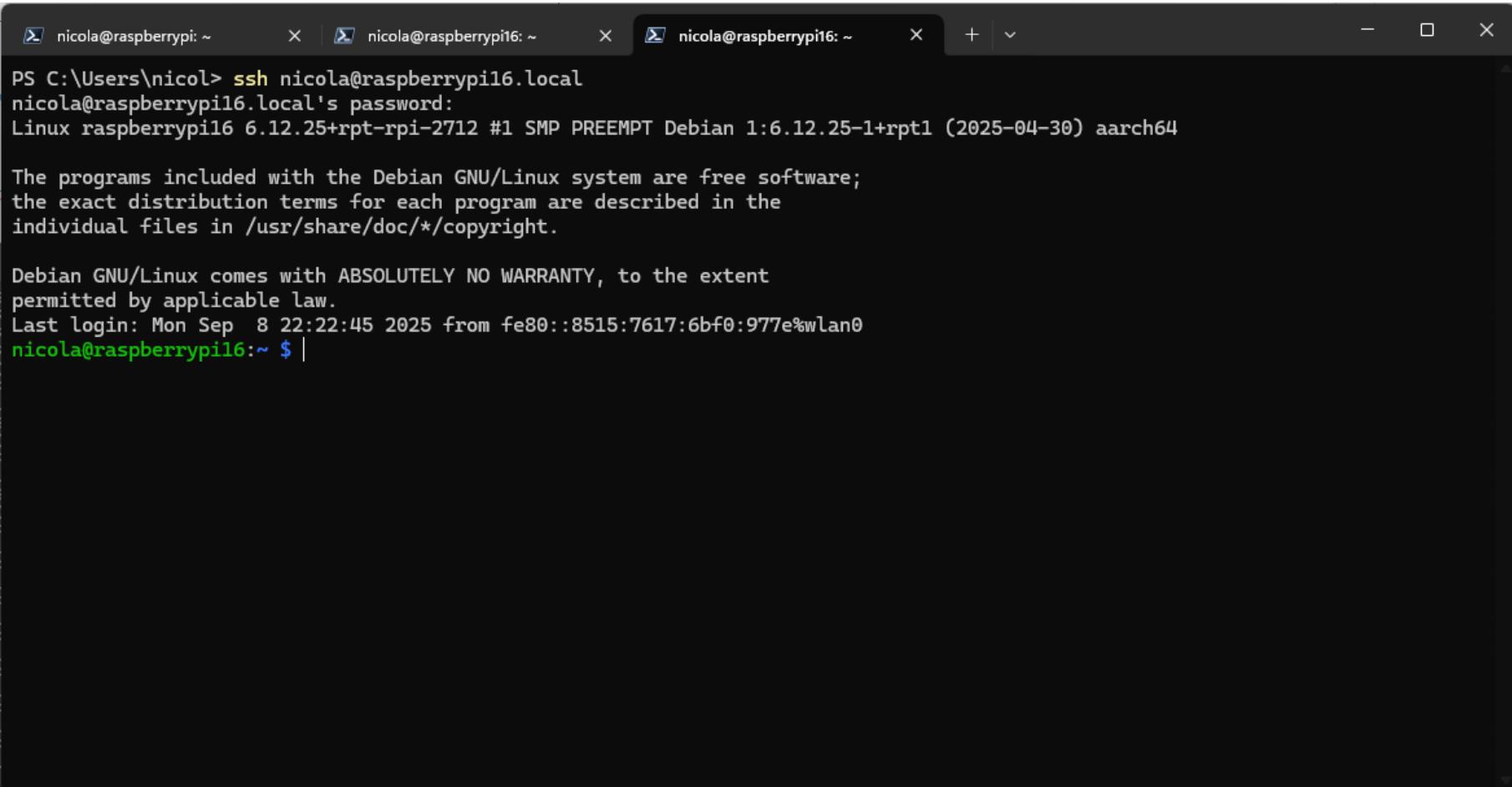
A FEW
MOMENTS LATER

Software



Connect via ssh

ssh <user>@<hostname>

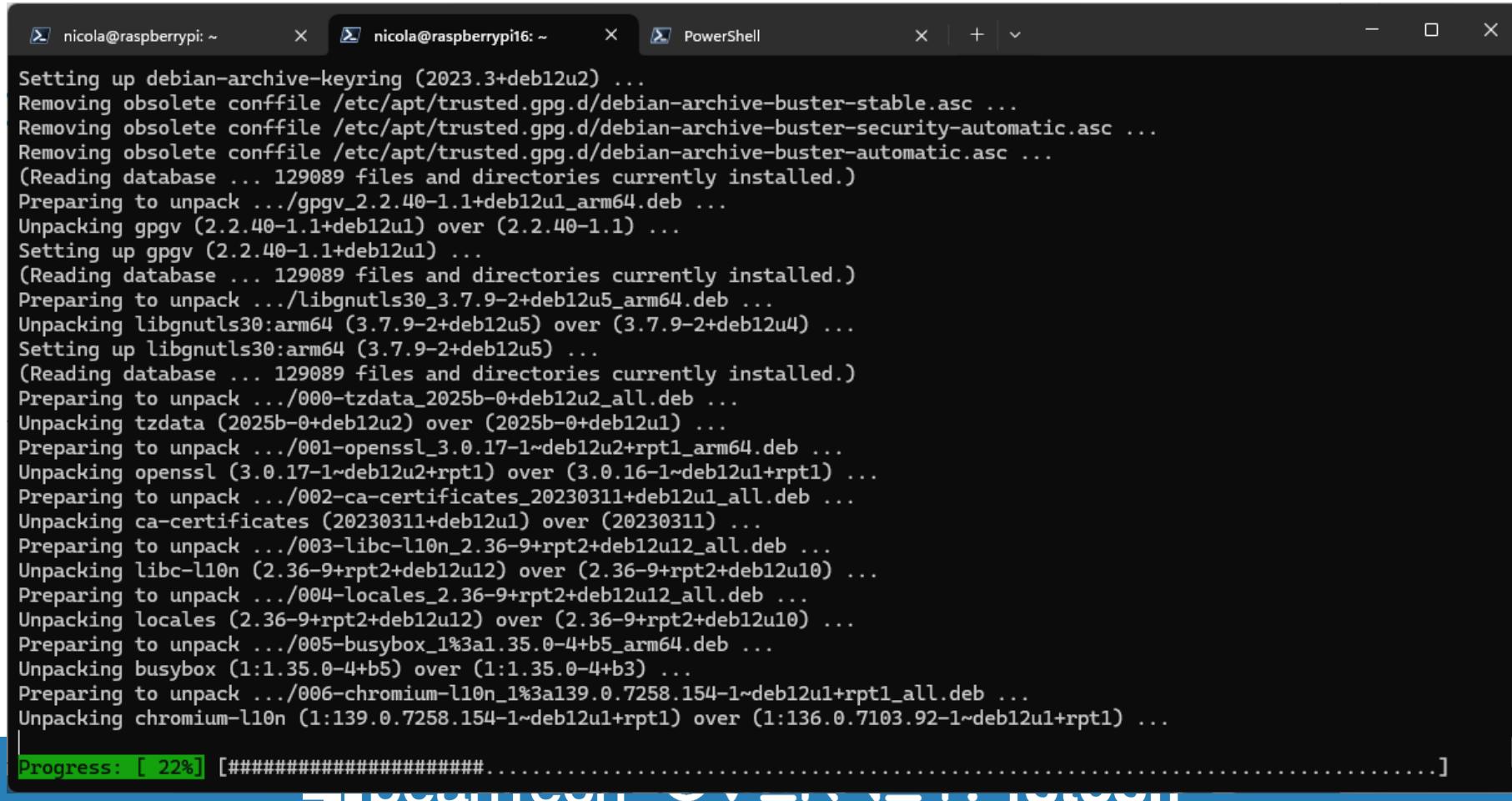


The screenshot shows a terminal window with three tabs open, each displaying a different command-line interface. The active tab is titled "nicola@raspberrypi16: ~". The session details are as follows:

- Windows Host:** PS C:\Users\nicol> ssh nicola@raspberrypi16.local
- Raspberry Pi 16 (Debian):** nicola@raspberrypi16.local's password: (password entered)
- Raspberry Pi 16 (Debian):** Linux raspberrypi16 6.12.25+rpt-rpi-2712 #1 SMP PREEMPT Debian 1:6.12.25-1+rpt1 (2025-04-30) aarch64
- Raspberry Pi 16 (Debian):** The programs included with the Debian GNU/Linux system are free software; the exact distribution terms for each program are described in the individual files in /usr/share/doc/*copyright.
- Raspberry Pi 16 (Debian):** Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent permitted by applicable law.
- Raspberry Pi 16 (Debian):** Last login: Mon Sep 8 22:22:45 2025 from fe80::8515:7617:6bf0:977e%wlan0
- Raspberry Pi 16 (Debian):** nicola@raspberrypi16:~ \$ |

Update

`sudo apt update && sudo apt upgrade -Y`

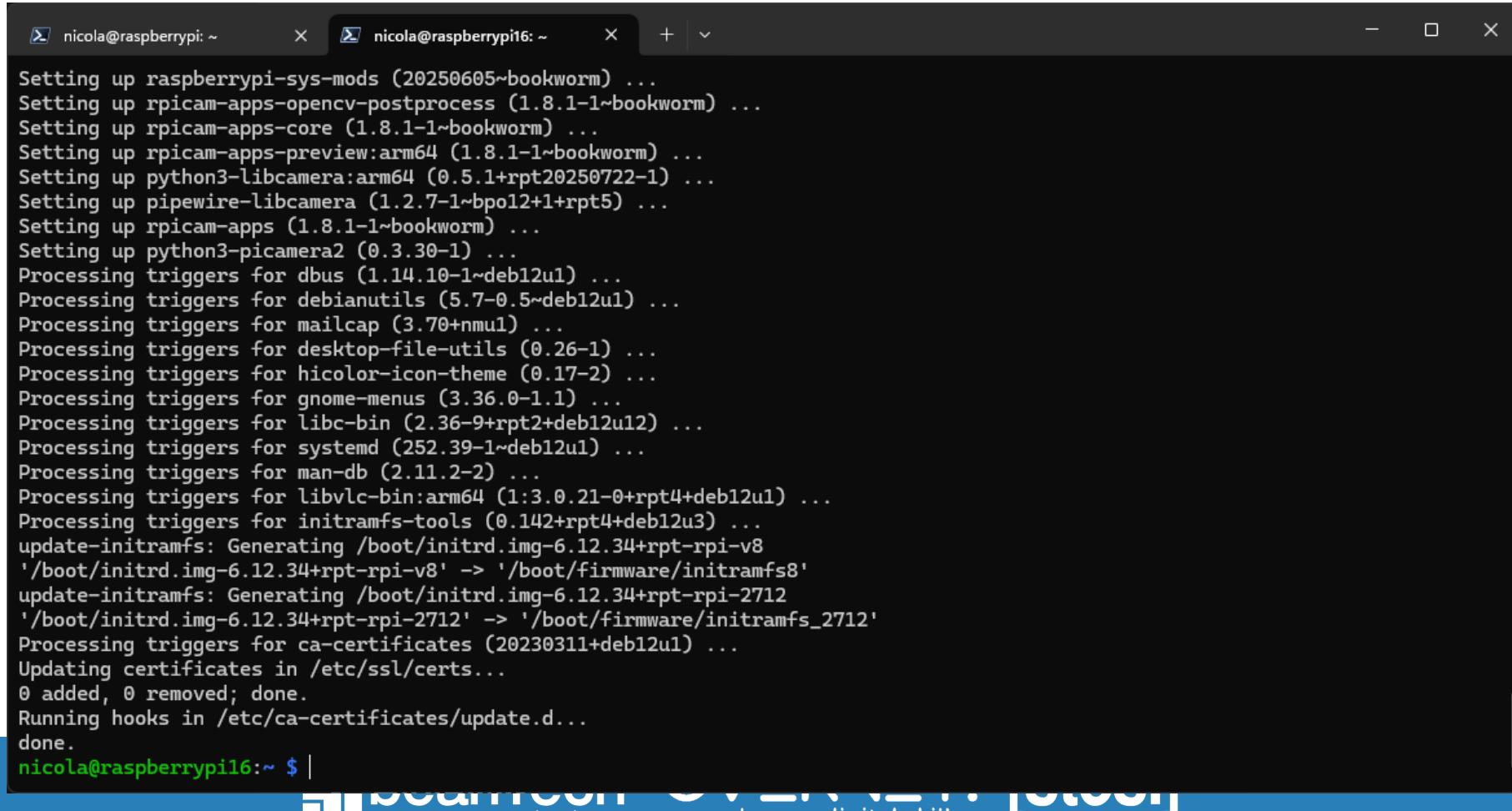


```
Setting up debian-archive-keyring (2023.3+deb12u2) ...
Removing obsolete conffile /etc/apt/trusted.gpg.d/debian-archive-buster-stable.asc ...
Removing obsolete conffile /etc/apt/trusted.gpg.d/debian-archive-buster-security-automatic.asc ...
Removing obsolete conffile /etc/apt/trusted.gpg.d/debian-archive-buster-automatic.asc ...
(Reading database ... 129089 files and directories currently installed.)
Preparing to unpack .../gpgv_2.2.40-1.1+deb12u1_arm64.deb ...
Unpacking gpgv (2.2.40-1.1+deb12u1) over (2.2.40-1.1) ...
Setting up gpgv (2.2.40-1.1+deb12u1) ...
(Reading database ... 129089 files and directories currently installed.)
Preparing to unpack .../libgnutls30_3.7.9-2+deb12u5_arm64.deb ...
Unpacking libgnutls30:arm64 (3.7.9-2+deb12u5) over (3.7.9-2+deb12u4) ...
Setting up libgnutls30:arm64 (3.7.9-2+deb12u5) ...
(Reading database ... 129089 files and directories currently installed.)
Preparing to unpack .../000-tzdata_2025b-0+deb12u2_all.deb ...
Unpacking tzdata (2025b-0+deb12u2) over (2025b-0+deb12u1) ...
Preparing to unpack .../001-openssl_3.0.17-1~deb12u2+rpt1_arm64.deb ...
Unpacking openssl (3.0.17-1~deb12u2+rpt1) over (3.0.16-1~deb12u1+rpt1) ...
Preparing to unpack .../002-ca-certificates_20230311+deb12u1_all.deb ...
Unpacking ca-certificates (20230311+deb12u1) over (20230311) ...
Preparing to unpack .../003-libc-l10n_2.36-9+rpt2+deb12u12_all.deb ...
Unpacking libc-l10n (2.36-9+rpt2+deb12u12) over (2.36-9+rpt2+deb12u10) ...
Preparing to unpack .../004-locales_2.36-9+rpt2+deb12u12_all.deb ...
Unpacking locales (2.36-9+rpt2+deb12u12) over (2.36-9+rpt2+deb12u10) ...
Preparing to unpack .../005-busybox_1%3a1.35.0-4+b5_arm64.deb ...
Unpacking busybox (1:1.35.0-4+b5) over (1:1.35.0-4+b3) ...
Preparing to unpack .../006-chromium-l10n_1%3a139.0.7258.154-1~deb12u1+rpt1_all.deb ...
Unpacking chromium-l10n (1:139.0.7258.154-1~deb12u1+rpt1) over (1:136.0.7103.92-1~deb12u1+rpt1) ...
|
Progress: [ 22%] [#####.....]
```

**ONE
HOUR LATER**

Update

`sudo apt update && sudo apt upgrade -Y`



```
Setting up raspberrypi-sys-mods (20250605~bookworm) ...
Setting up rpicam-apps-opencv-postprocess (1.8.1-1~bookworm) ...
Setting up rpicam-apps-core (1.8.1-1~bookworm) ...
Setting up rpicam-apps-preview:arm64 (1.8.1-1~bookworm) ...
Setting up python3-libcamera:arm64 (0.5.1+rpt20250722-1) ...
Setting up pipewire-libcamera (1.2.7-1~bpo12+1+rpt5) ...
Setting up rpicam-apps (1.8.1-1~bookworm) ...
Setting up python3-picamera2 (0.3.30-1) ...
Processing triggers for dbus (1.14.10-1~deb12u1) ...
Processing triggers for debianutils (5.7-0.5~deb12u1) ...
Processing triggers for mailcap (3.70+nmu1) ...
Processing triggers for desktop-file-utils (0.26-1) ...
Processing triggers for hicolor-icon-theme (0.17-2) ...
Processing triggers for gnome-menus (3.36.0-1.1) ...
Processing triggers for libc-bin (2.36-9+rpt2+deb12u12) ...
Processing triggers for systemd (252.39-1~deb12u1) ...
Processing triggers for man-db (2.11.2-2) ...
Processing triggers for libvlc-bin:arm64 (1:3.0.21-0+rpt4+deb12u1) ...
Processing triggers for initramfs-tools (0.142+rpt4+deb12u3) ...
update-initramfs: Generating /boot/initrd.img-6.12.34+rpt-rpi-v8
'/boot/initrd.img-6.12.34+rpt-rpi-v8' -> '/boot/firmware/initramfs8'
update-initramfs: Generating /boot/initrd.img-6.12.34+rpt-rpi-2712
'/boot/initrd.img-6.12.34+rpt-rpi-2712' -> '/boot/firmware/initramfs_2712'
Processing triggers for ca-certificates (20230311+deb12u1) ...
Updating certificates in /etc/ssl/certs...
0 added, 0 removed; done.
Running hooks in /etc/ca-certificates/update.d...
done.
nicola@raspberrypi16:~ $ |
```



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

The screenshot shows a terminal window with two tabs. The left tab is titled 'nicola@raspberrypi: ~' and the right tab is titled 'nicola@raspberrypi16: ~'. Both tabs have an icon of a terminal window. The terminal content is as follows:

```
nicola@raspberrypi16:~ $ rpi-connect on
✓ Raspberry Pi Connect started
nicola@raspberrypi16:~ $ rpi-connect signin
Complete sign in by visiting https://connect.raspberrypi.com/verify/R2VZ-C32Z

✓ Signed in
nicola@raspberrypi16:~ $ |
```

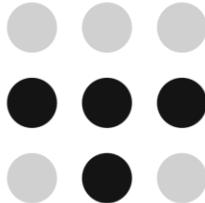


Raspberry Pi Connect

The screenshot shows the Raspberry Pi Connect web interface at <https://connect.raspberrypi.com/devices>. The interface has a header with a back button, refresh button, and a search bar. The main area is titled "Raspberry Pi Connect" and shows a "Personal" section. A "Devices" tab is selected, and there is a "Add device" button. Two devices are listed:

Device	Client version	Last seen	Action
raspberrypi Screen sharing Remote shell	2.5.2	● Online	Connect via ▾
raspberrypi16 Screen sharing Remote shell	2.5.2	● Online	Connect via ▾

VPN?

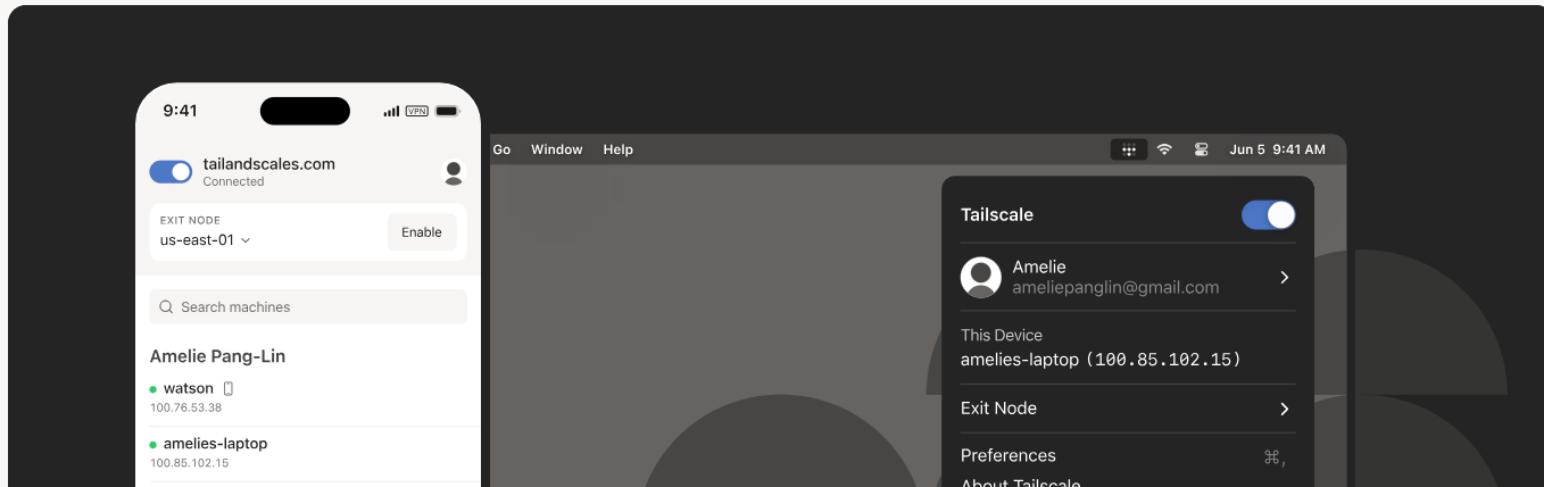


tailscale

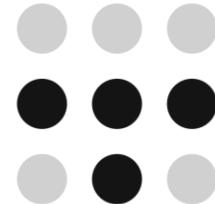
Your legacy VPN belongs in the past

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

[Start connecting devices →](#) [Contact sales](#)



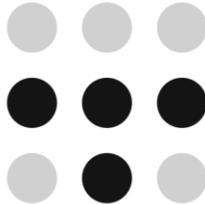
VPN?



tailscale

```
nicola@raspberrypi:~ $ curl -fsSL https://tailscale.com/install.sh | sh
Installing Tailscale for debian bookworm, using method apt
+ sudo mkdir -p --mode=0755 /usr/share/keyrings
+ curl -fsSL https://pkgs.tailscale.com/stable/debian/bookworm.noarmor.gpg
+ sudo tee /usr/share/keyrings/tailscale-archive-keyring.gpg
+ sudo chmod 0644 /usr/share/keyrings/tailscale-archive-keyring.gpg
+ curl -fsSL https://pkgs.tailscale.com/stable/debian/bookworm.tailscale-keyring.list+
sudo tee /etc/apt/sources.list.d/tailscale.list
# Tailscale packages for debian bookworm
deb [signed-by=/usr/share/keyrings/tailscale-archive-keyring.gpg] https://pkgs.tailscale.com/stable/debi
an bookworm main
+ sudo chmod 0644 /etc/apt/sources.list.d/tailscale.list
+ sudo apt-get update
Hit:1 http://deb.debian.org/debian bookworm InRelease
Hit:2 http://deb.debian.org/debian-security bookworm-security InRelease
Hit:3 http://deb.debian.org/debian bookworm-updates InRelease
```

VPN?



tailscale

```
PS C:\Users\nicol> ssh nicola@raspberrypi16
The authenticity of host 'raspberrypi16 (100.103.91.128)' can't be established.
ED25519 key fingerprint is SHA256:g5FwUywJIrZbMksgEEk4KJLZYvQjU9xrm7ahCQIvT8I.
This host key is known by the following other names/addresses:
  C:\Users\nicol/.ssh/known_hosts:7: raspberrypi16.local
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'raspberrypi16' (ED25519) to the list of known hosts.
nicola@raspberrypi16's password:
Linux raspberrypi16 6.12.25+rpi-2712 #1 SMP PREEMPT Debian 1:6.12.25-1+rpi1 (2025-04-30) aarch64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/*copyright.

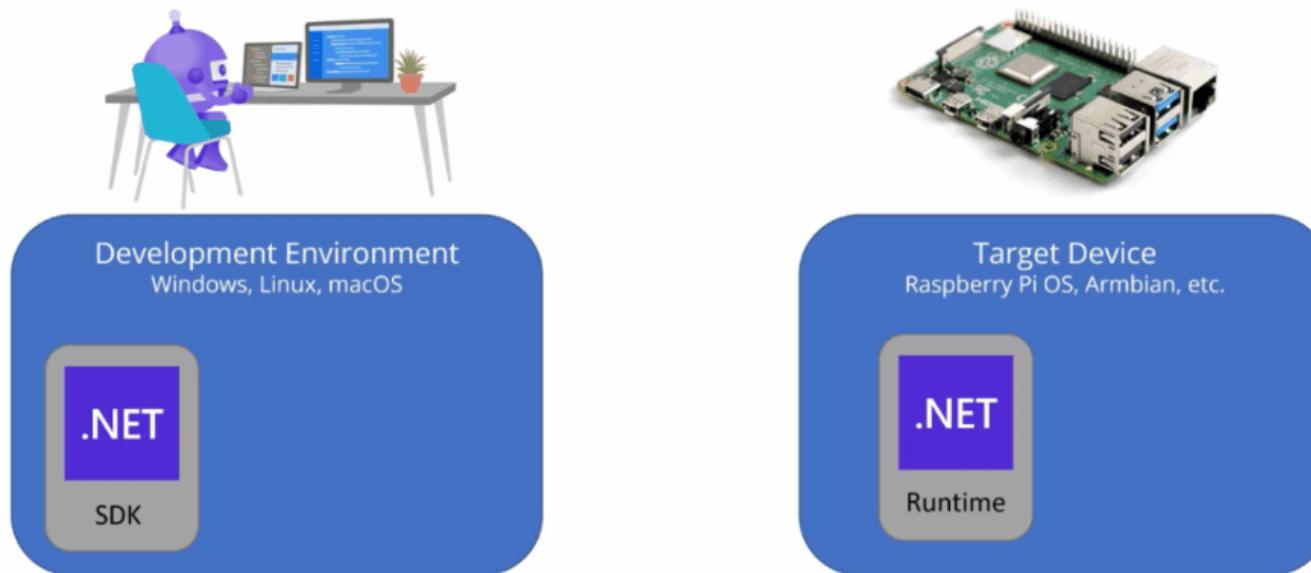
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Mon Sep  8 22:38:50 2025 from fe80::8515:7617:6bf0:977e%eth0
nicola@raspberrypi16:~ $ |
```

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

Framework-Dependent

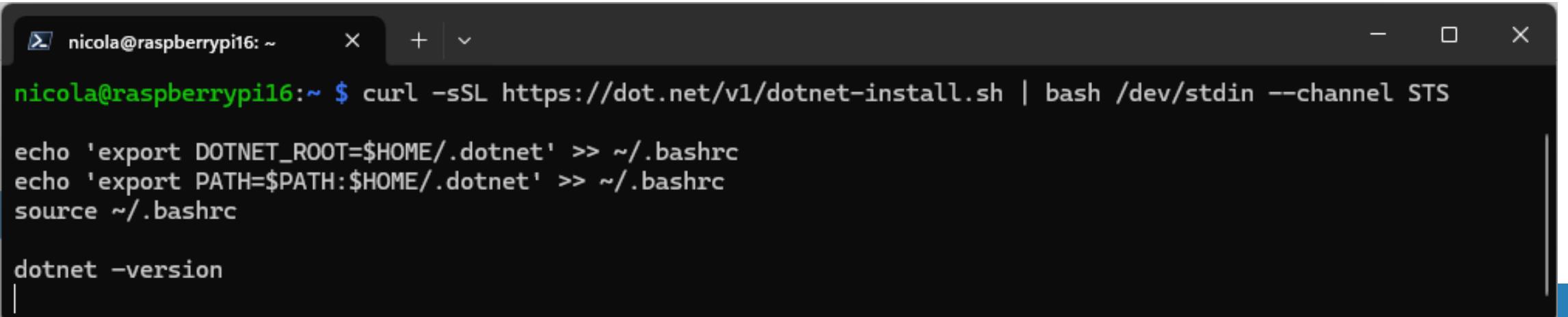


Installiamo .NET

```
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
```



The screenshot shows a terminal window titled "nicola@raspberrypi16: ~". The window contains the following command and its execution:

```
nicola@raspberrypi16:~ $ 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
```

Below the terminal window, there is a horizontal bar with several logos:

- hnova community
- pearlcom
- overvalo
- plus
- SORZIO
UNIVERSITARIO
DI PORDENONE
MOLTIPLICATORE DI VALORE

Installiamo .NET

```
nicola@raspberrypi16: ~      X  nicola@raspberrypi16: ~      X + - ▾
t.com/dotnet/Sdk/9.0.304/dotnet-sdk-9.0.304-linux-arm64.tar.gz
dotnet-install: Remote file https://builds.dotnet.microsoft.com/dotnet/Sdk/9.0.304/dotn
et-sdk-9.0.304-linux-arm64.tar.gz size is 213015907 bytes.
dotnet-install: Extracting archive from https://builds.dotnet.microsoft.com/dotnet/Sdk/
9.0.304/dotnet-sdk-9.0.304-linux-arm64.tar.gz
dotnet-install: Downloaded file size is 213015907 bytes.
dotnet-install: The remote and local file sizes are equal.
dotnet-install: Installed version is 9.0.304
dotnet-install: Adding to current process PATH: `/home/nicola/.dotnet`. Note: This chan
ge will be visible only when sourcing script.
dotnet-install: Note that the script does not resolve dependencies during installation.
dotnet-install: To check the list of dependencies, go to https://learn.microsoft.com/do
tnet/core/install, select your operating system and check the "Dependencies" section.
dotnet-install: Installation finished successfully.

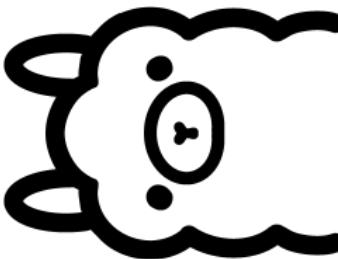
Welcome to .NET 9.0!
-----
SDK Version: 9.0.304

Telemetry
-----
The .NET tools collect usage data in order to help us improve your experience. It is co
llected by Microsoft and shared with the community. You can opt-out of telemetry by set
```

Installiamo Ollama



Download Ollama



macOS



Linux



Windows

Install with one command:

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



[View script source](#) • [Manual install instructions](#)

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

```
nicola@raspberrypi16:~ $ curl -fsSL https://ollama.com/install.sh | sh
>>> Cleaning up old version at /usr/local/lib/ollama
>>> Installing ollama to /usr/local
>>> Downloading Linux arm64 bundle
#####
# 100.0%
>>> Creating ollama user...
>>> Adding ollama user to render group...
>>> Adding ollama user to video group...
>>> Adding current user to ollama group...
>>> Creating ollama systemd service...
>>> Enabling and starting ollama service...
Created symlink /etc/systemd/system/default.target.wants/ollama.service → /etc/systemd/
system/ollama.service.
>>> The Ollama API is now available at 127.0.0.1:11434.
>>> Install complete. Run "ollama" from the command line.
WARNING: No NVIDIA/AMD GPU detected. Ollama will run in CPU-only mode.
```



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

Filosofia

Plugin modulari + pianificatori

Memoria

Embedding + contesto persistente

Integrazione

Forte con Azure e Microsoft

Target

Aziende, sviluppatori enterprise

Use case ideali

Copilot aziendali, workflow complessi



Python, JavaScript

Catene di chiamate + agenti

Memoria conversazionale

Ampia con strumenti esterni

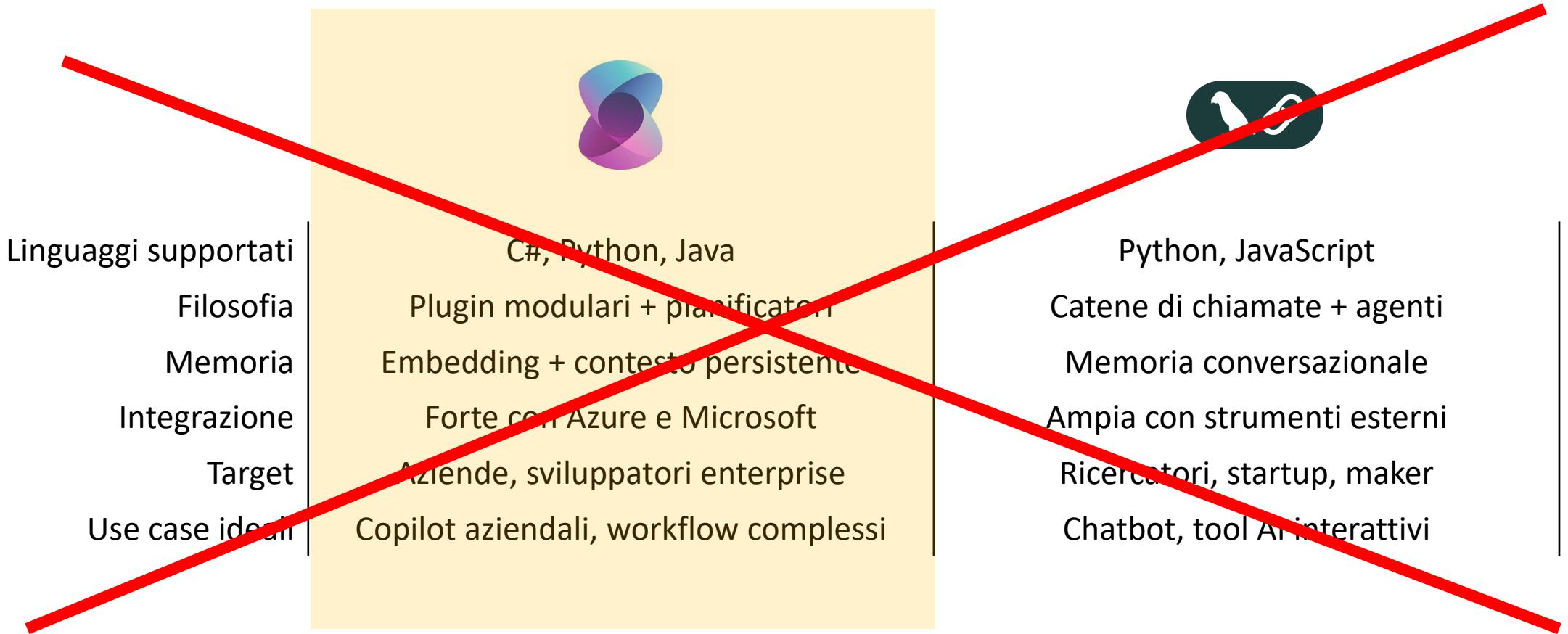
Ricercatori, startup, maker

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



La soluzione?

Metodo Coleman

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



Scelta dei modelli

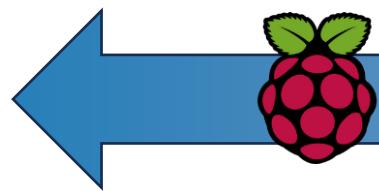
Scelta dei modelli

<https://ollama.com/search>

The screenshot shows the Ollama search interface. At the top, there is a search bar with the query "gwen", a sidebar icon, and a menu icon. Below the search bar are filter buttons for "Embedding", "Vision", "Tools", and "Thinking", with "Popular" selected. The main results section displays the "gwen" model card. The card features a large blue box containing the text "Capabilities" and "Numero di Parametri". Arrows point from these labels to specific sections of the card. The "Capabilities" section describes the model as offering a comprehensive suite of dense models. The "Numero di Parametri" section describes it as being part of the Qwen series, using a mixture-of-experts (MoE) architecture. Below this is a horizontal bar with buttons for "tools" and "thinking", and a list of parameter counts: 0.6b, 1.7b, 4b, 8b, 14b, 30b, 32b, and 235b. At the bottom of the card, there are metrics: 8M Pulls, 56 Tags, and an update timestamp of "Updated 1 month ago".

Scelta dei modelli – Numero di Parametri

Pochi Parametri



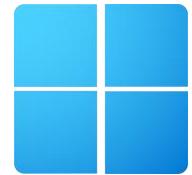
- Velocità di risposta
- Allucinazioni

Tanti Parametri

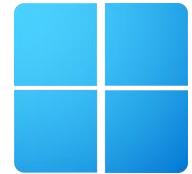


- Accuratezza della risposta
- Consumo di energia
- Consumo di memoria

Scelta dei modelli – Numero di Parametri



Ollama, Ollama



Eating RAM?



Telling lies?



Open your mouth



Yes, papa



No, papa



No, papa



Processes			CPU	Memory	Disk	Network
Name	S...	46%				
> ollama.exe (3)		46.2%	5,900.5 MB	0.1 MB/s	0 Mbps	

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

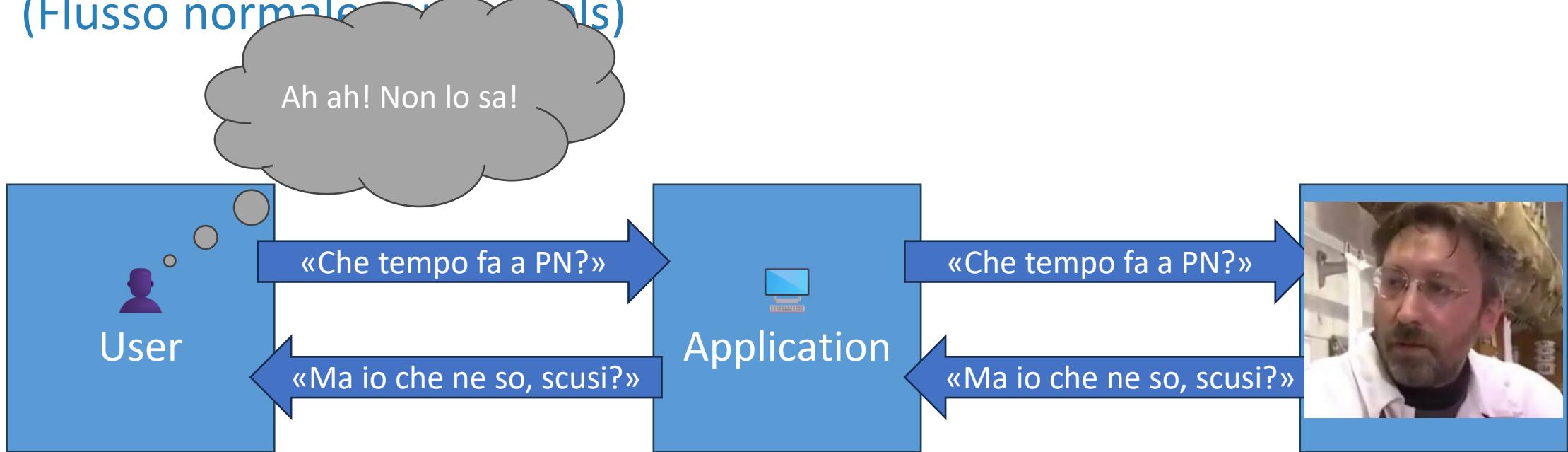
Scelta dei modelli – Capabilities

- 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?

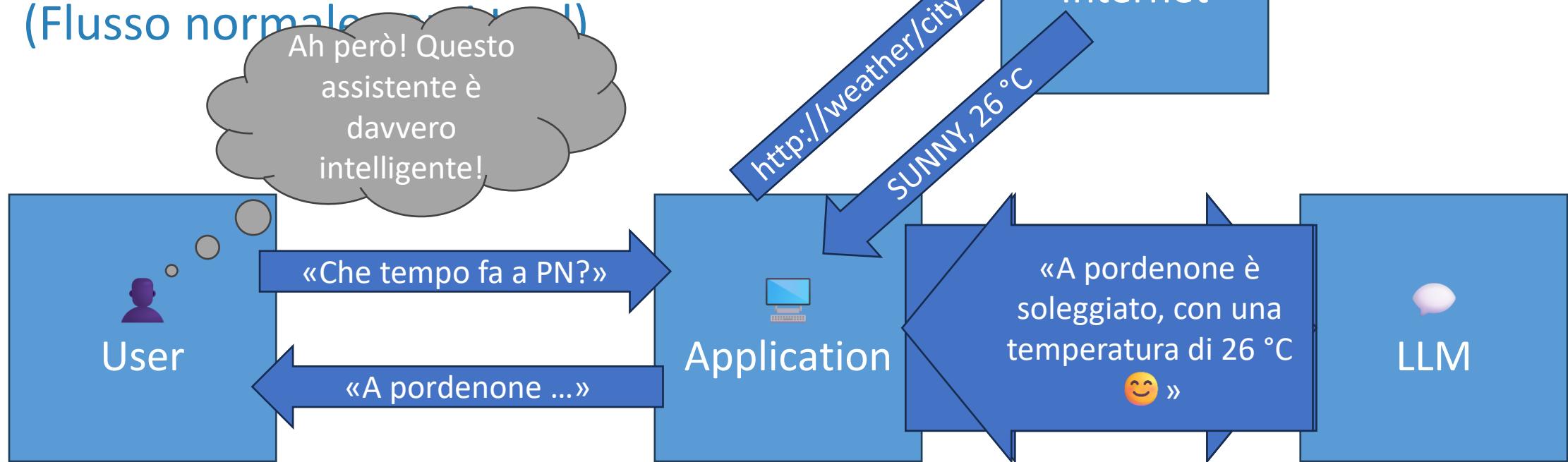
- Un tool è una funzionalità esterna che il modello può attivare per svolgere compiti specifici che vanno oltre la semplice generazione di testo.
- I tools permettono al modello di interagire con il mondo esterno o di eseguire operazioni complesse che richiedono capacità specifiche.

Come funziona un tool? (Flusso normale - normal flow)



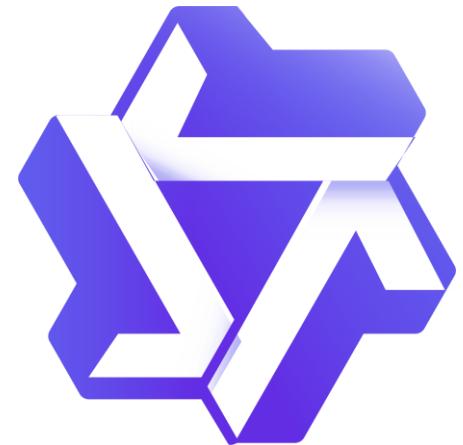
Come funziona un tool?

(Flusso normale - ChatGPT)



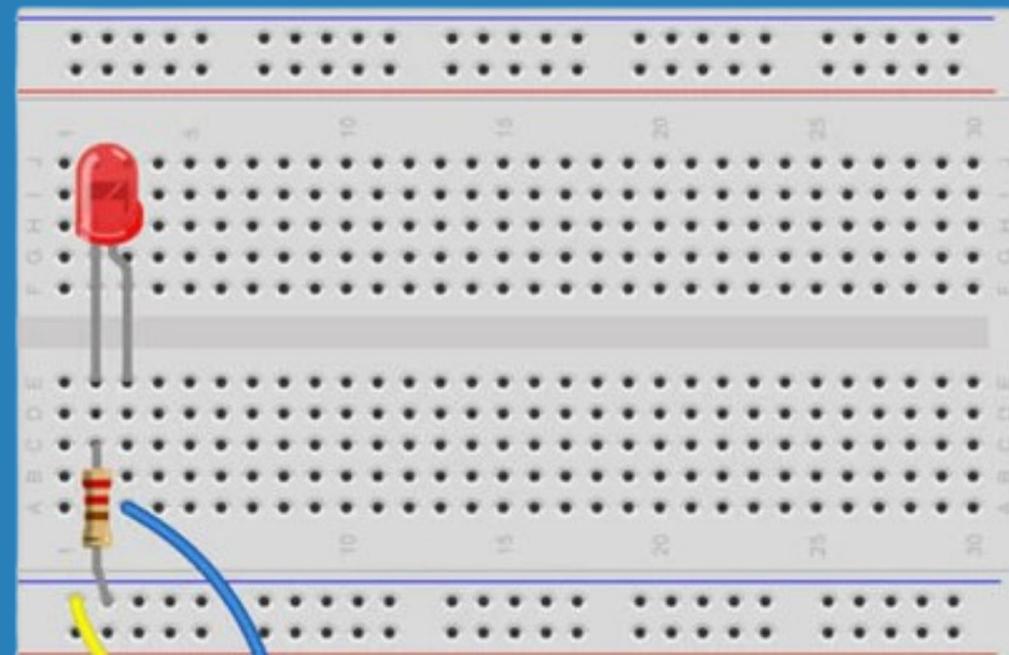
Quindi che modello scelgo?

- Llama3.1 o 3.2 X
 - 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 X
 - 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



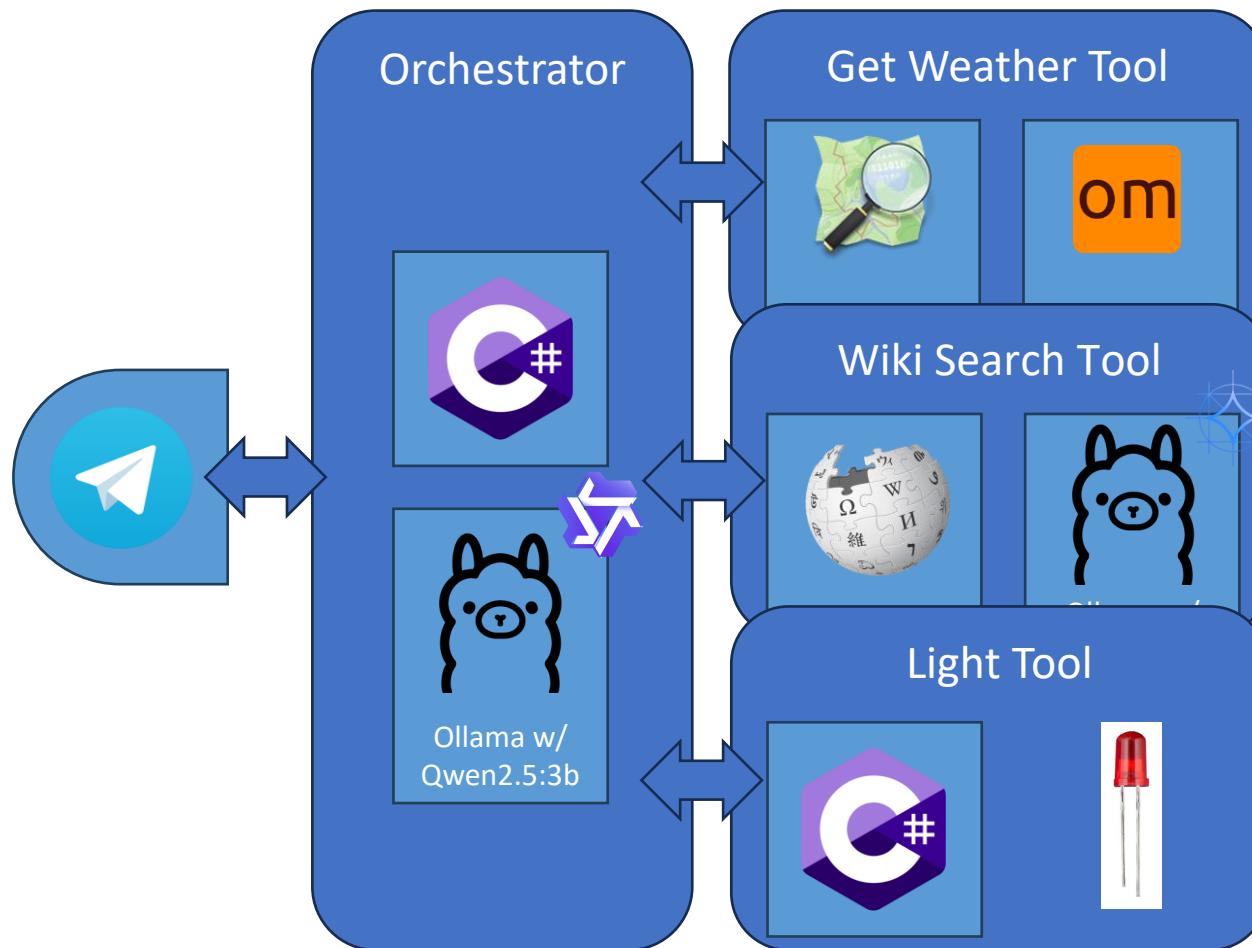


DEMO



CloudGen
ognidi verso l'evoluzione

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 ★★★★☆ (251)

2.799⁹⁶ €

✓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-...
2.799,96€
Consegna GRATIS
lunedì

K8 Plus
8845HS...
629,96€
Consegna GRATIS
giovedì

K8 Plus-
8845HS-...
799,96€
Consegna GRATIS
martedì

K11-
8945HS-...
889,96€
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 ★★★★☆ (20)

[Scelta Amazon](#)

1.255⁰⁰ €

✓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

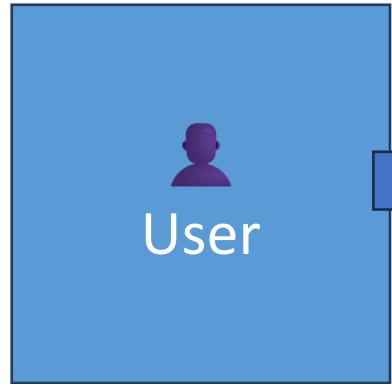
M1Pro-125H
0/0 GB
415,00€

M1Pro-125H
32/1TB
583,00€

M1 Pro-
285H...
1.119,99€

X1 Pro-370
32/1TB
999,00€
Consegna GRATIS

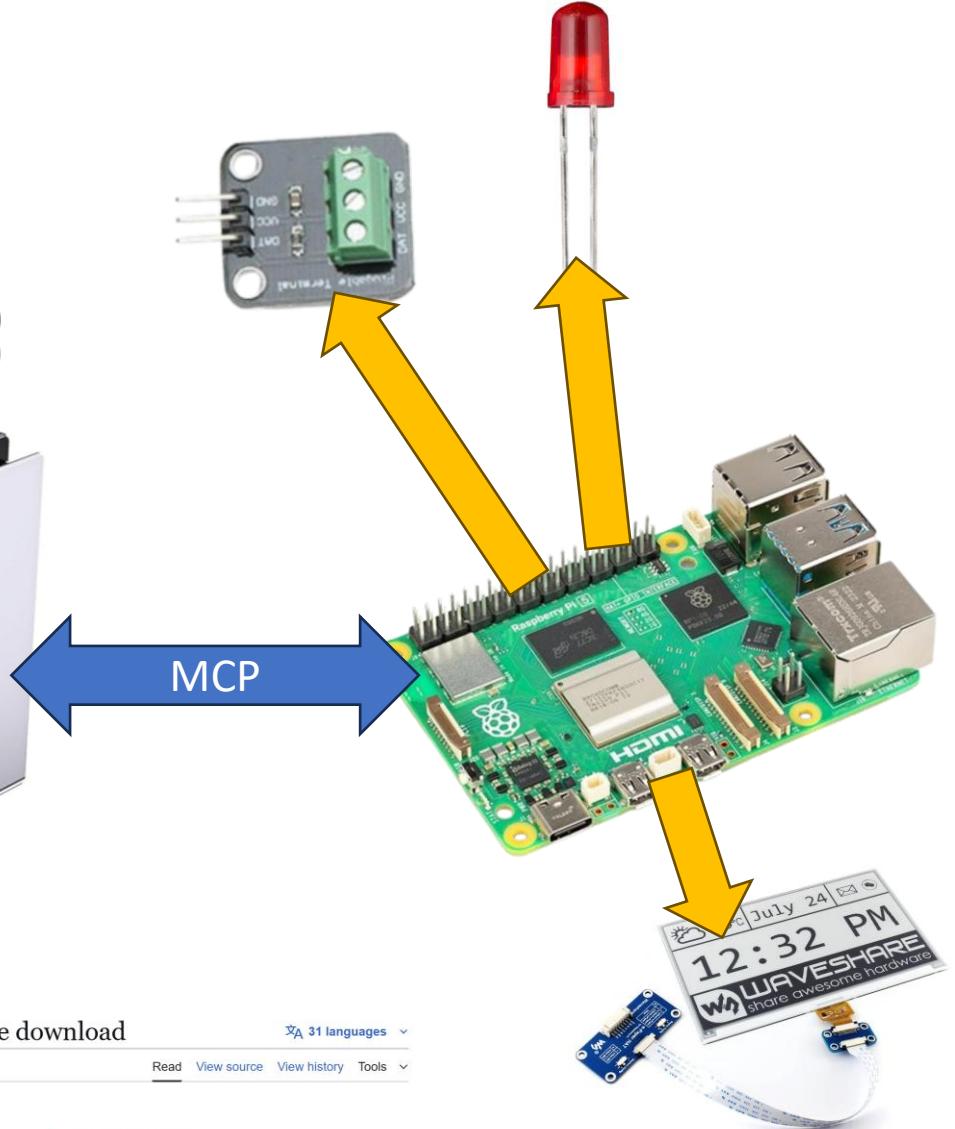
Piuttosto... MCP

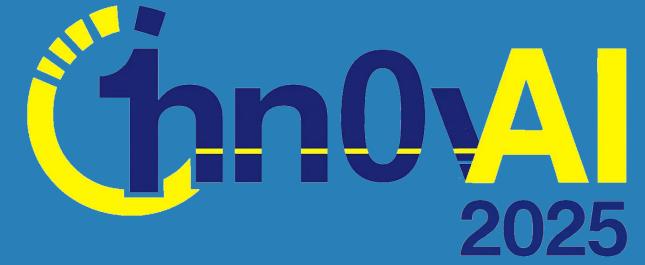


«Accendi la luce»



MCP





GRAZIE!



About me

Nicola Paro

Solutions Architect @ beanTech

.NET & Azure Meetup Štajerska Community Lead



codice



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