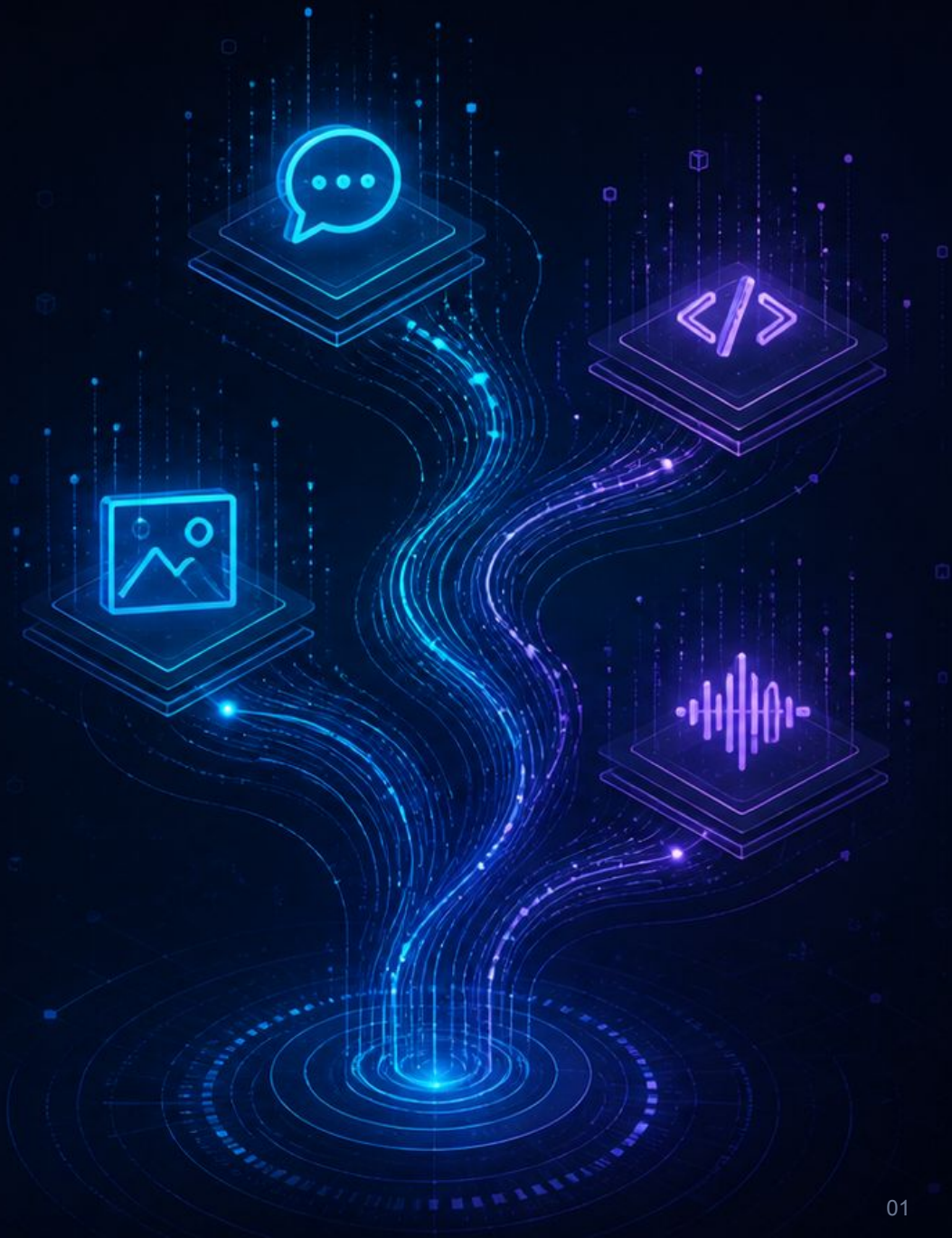
Local AI oggi

modelli, hardware e compromessi

Alessio Soldano

AI Italy • Meetup #14
12 Maggio 2026



AGENDA

Ha senso parlare di AI in locale nel 2026?

01 **Casi d’uso**
Chatbot, coding, immagini, etc: cosa conviene davvero

02 **Consigli pratici**
Scelta del modello LLM

03 **Software**
Model serving, UI, agenti, quantizzazione, …

04 **Hardware**
GPU, RAM/VRAM: cosa conta davvero

05 **Trade-off**
Qualità, privacy, costo, prestazioni, affidabilità

Tesi: locale e cloud non si escludono. Il futuro pratico è ibrido.

01 CASI D'USO

Dove il locale dà valore oggi

Perché dovremmo pensare di fare inferenza in locale invece di affidarci ad abbonamenti a servizi in cloud? Quali pro e contro?

- Differenti tipi di spesa (CAPEX vs OPEX) con diverse implicazioni contabili
- Controllo su aggiornamenti e disponibilità servizi
- Differenti skillset richiesti (platform / admin engineers, …)
- Differenti prestazioni (velocità e accuratezza/qualità)

Esistono casi d'uso per i quali i vantaggi possono bilanciare gli svantaggi.

Takeaway: partire dal caso d'uso, dal problema o opportunità.

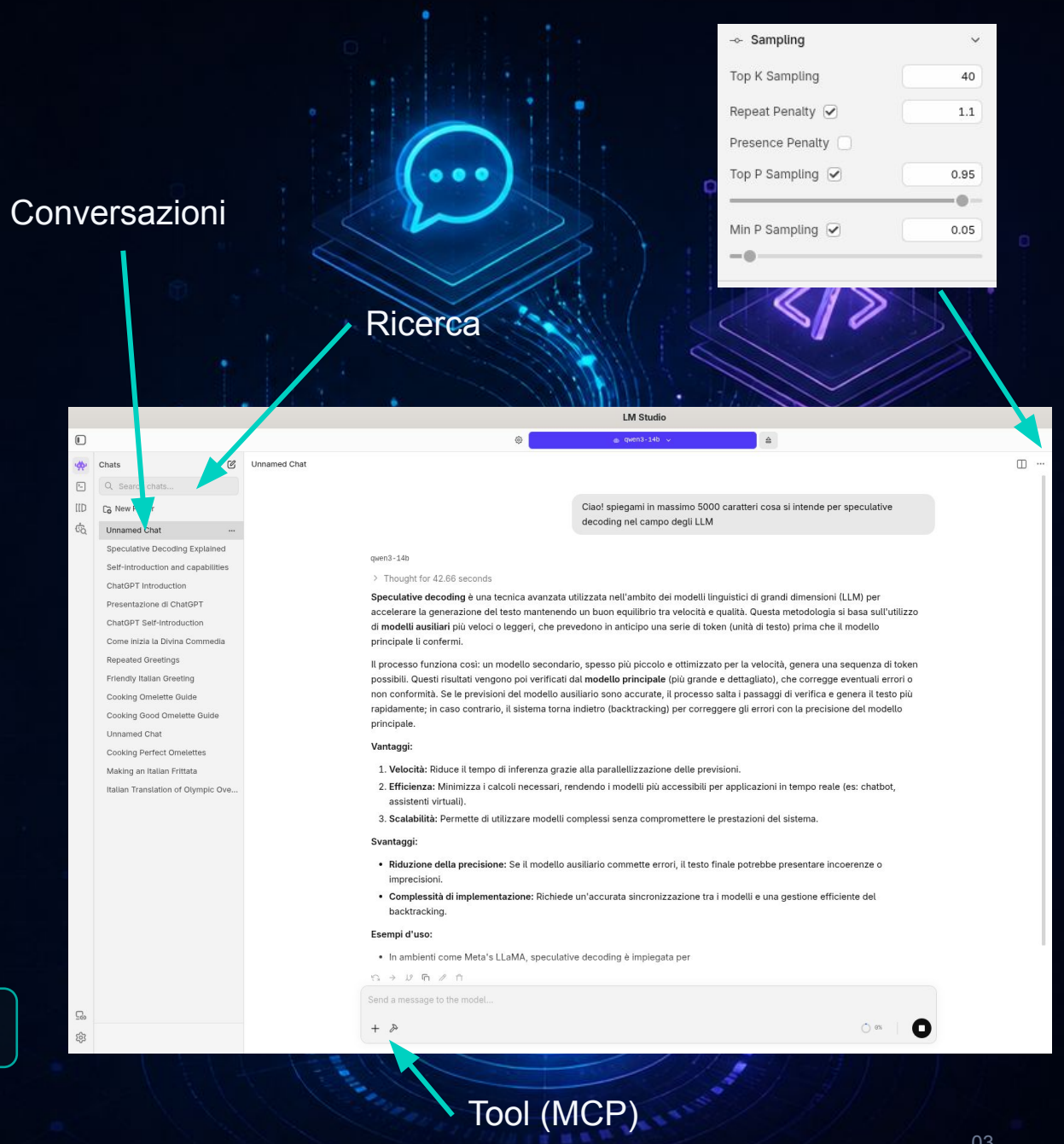
01 CASI D'USO

Chatbot

Conversazione con un LLM eventualmente supportato da tool (ricerca web, basi dati, …) e dotato di memoria

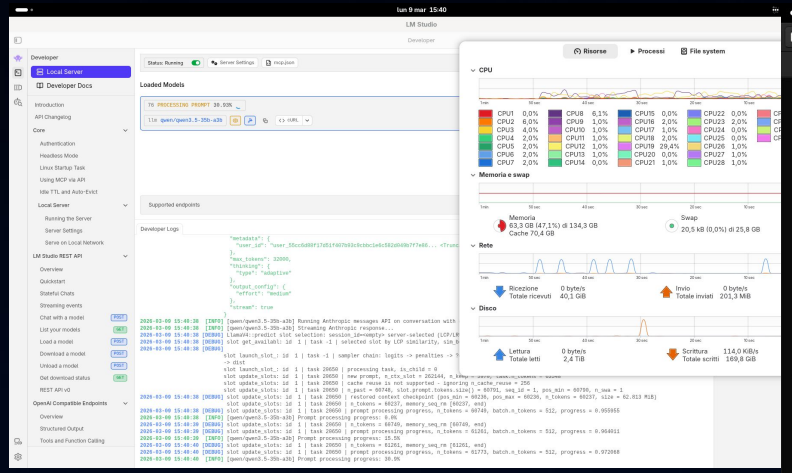
- LM Studio, Open WebUI, Ollama, …
- Perché in locale?
  - Discussioni "private"
  - Discussioni trivial / non meritevoli di spesa token (domande mi-sento-fortunato, traduzioni, …)
  - Gestione personalizzata di profilazione, memoria, utenti
  - Controllo della configurazione dell'inferenza del bot

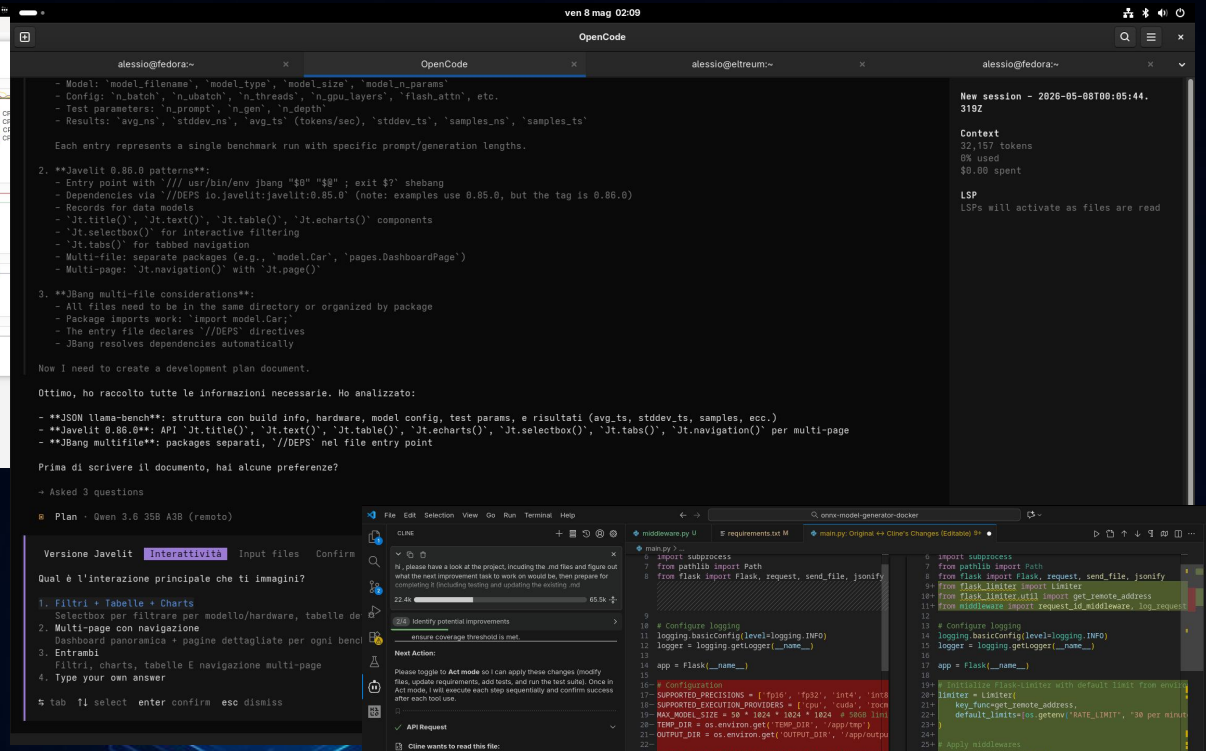
"Alternativa" a ChatGPT / Gemini / …



01 CASI D'USO

Coding





Code completion e agentic coding a complessità bassa o media

- Opencode, LM Studio, llama.cpp
- Claude Code, KiloCode, Cline, Cursor, …
- Sandbox -> lince.sh, SDD -> backlog.md

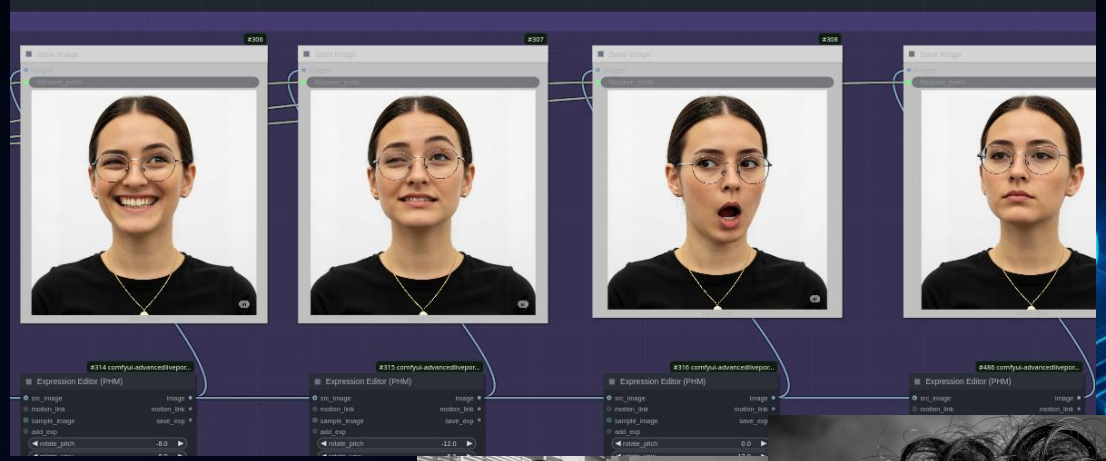
- Quando in locale?
  - Sviluppo su codebase proprietarie e/o ridotte
  - Task non critici (test, operazioni riptetitive, …)
  - Task non meritevoli di utilizzo modelli SOTA

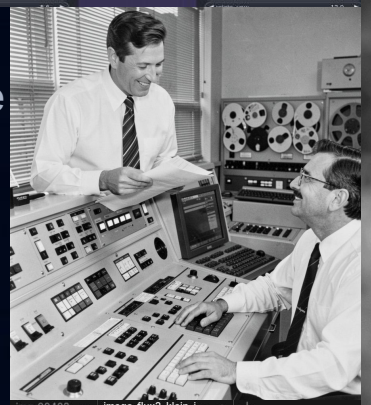
★ Running your coding agents locally: Lessons from a Real-World Experiment (Zurich, Mar 2026)
★ Aladino Digitale

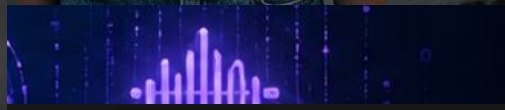
Profilo di utilizzo aggiuntivo per Claude Code / Opencode / Cline / Cursor …

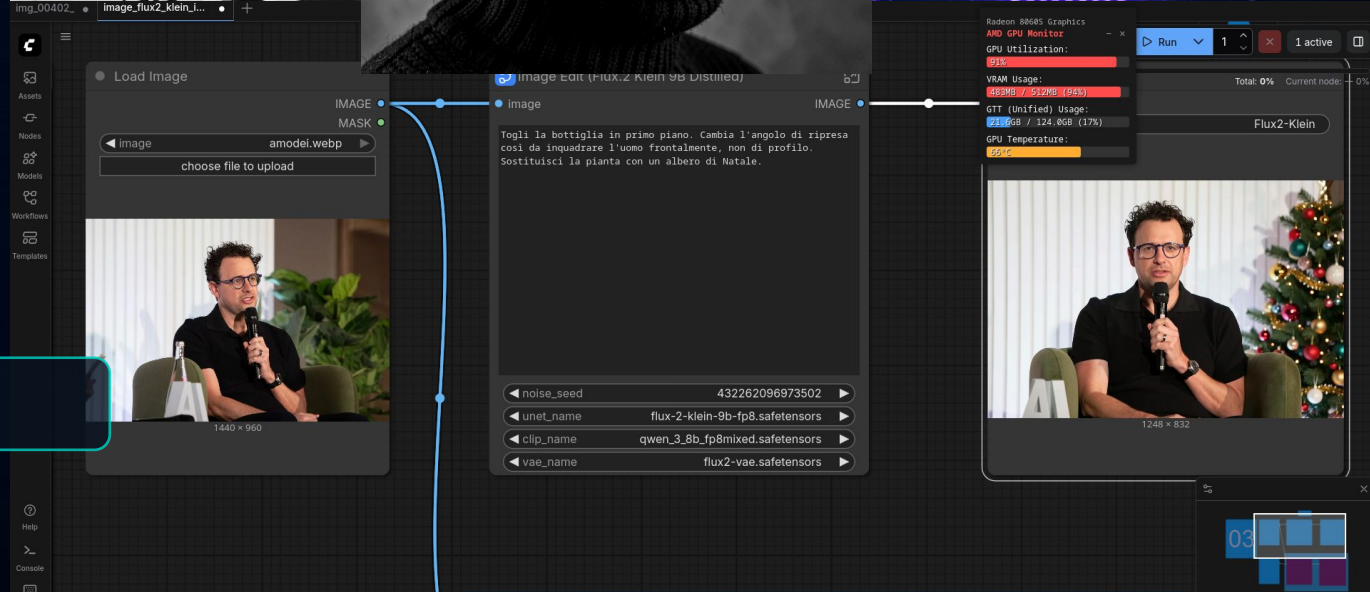
01 CASI D'USO

Immagini









Fotoritocco / editing immagini personali, categorizzazione e raggruppamento foto, creazione materiale grafico …

- Quando in locale?
  - Foto personali / "private" o con diritti d'uso a pagamento
  - Lavorazioni rifiutate da tool cloud
  - LORA, Consistent characters e risparmio
  - Generazione immagini ai vertici delle classifiche non richiesto
  - Risparmio su piattaforme dedicate

- ComfyUI, raccolte di prompt e workflow già fatti

"Alternativa" a GPT Image, Nano Banana, … e servizi tipo Openart.ai

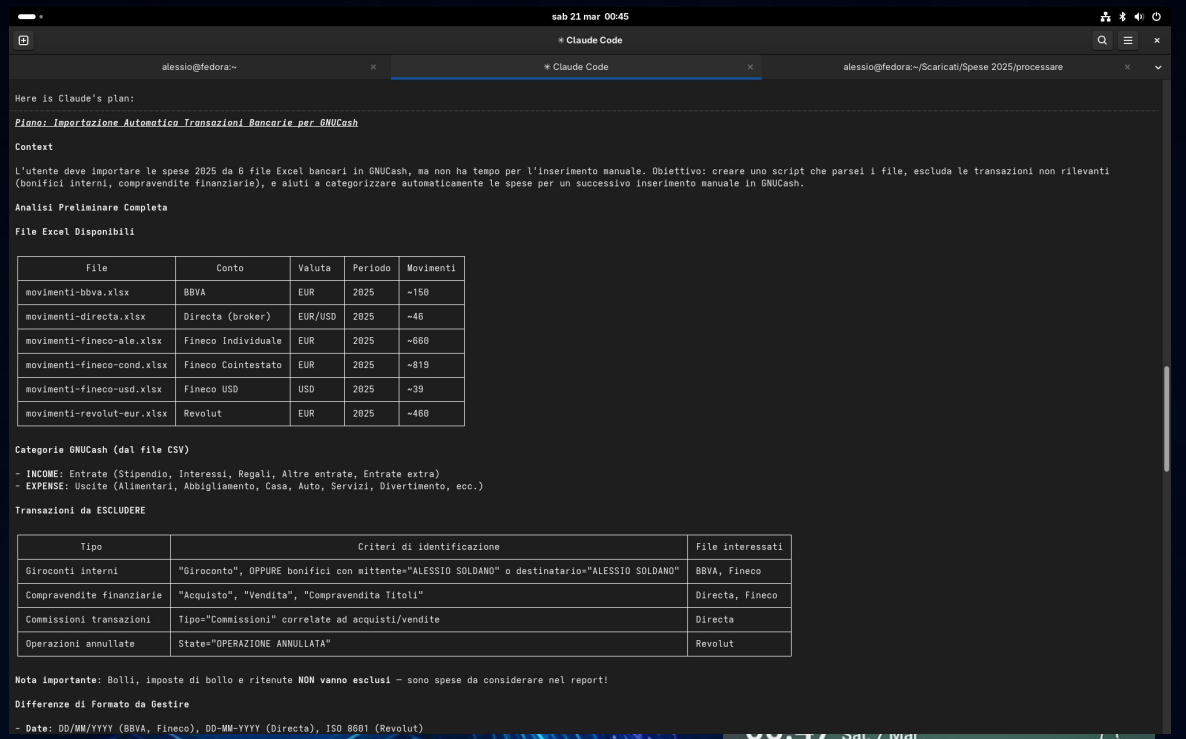
01 CASI D'USO

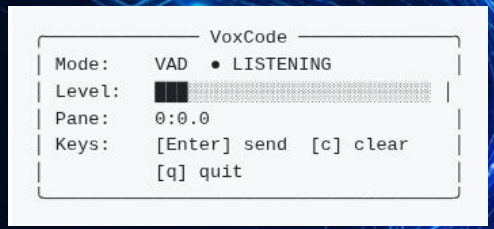
Task vari e custom application

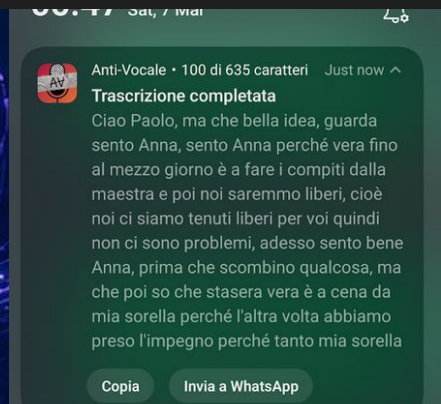
Elaborazione dati sensibili (email, status reports, strategy docs, bank transactions…), migliorie AI in applicazioni, …

- Skill produttività personale
- Applicazioni AI-infused
  - Categorizzazione e incrocio transazioni finanziarie
  - Applicazioni con speech-to-text (Anti-Vocale, VoxCode)
  - Assistenti con supporto NL (es. WildFly Model Graph Tools)
- Analisi immagini (modelli vision)
- Hermes, OpenClaw, …

- Perché in locale?
  - Dati sensibili (privacy, vincoli legali, etc.)
  - Risparmio token dove taks non meritano uso cloud







02 CONSIGLI PRATICI

Da dove partire, cosa evitare

Procedi con metodo: **Scenario -> Modello -> Verifica**

1. Definisci lo scenario

Da dove partire, cosa evitare

Procedi con metodo: **Scenario -> Modello -> Verifica**

1. Definisci lo scenario
2. Identifica modelli soddisfacenti (accuratezza)

Artificial Analysis Intelligence Index

Artificial Analysis Agentic Index
Represents the average of **agentic capabilities benchmarks** in the Artificial Analysis Intelligence Index (**GDPval-AA, τ^2-Bench Telecom**)

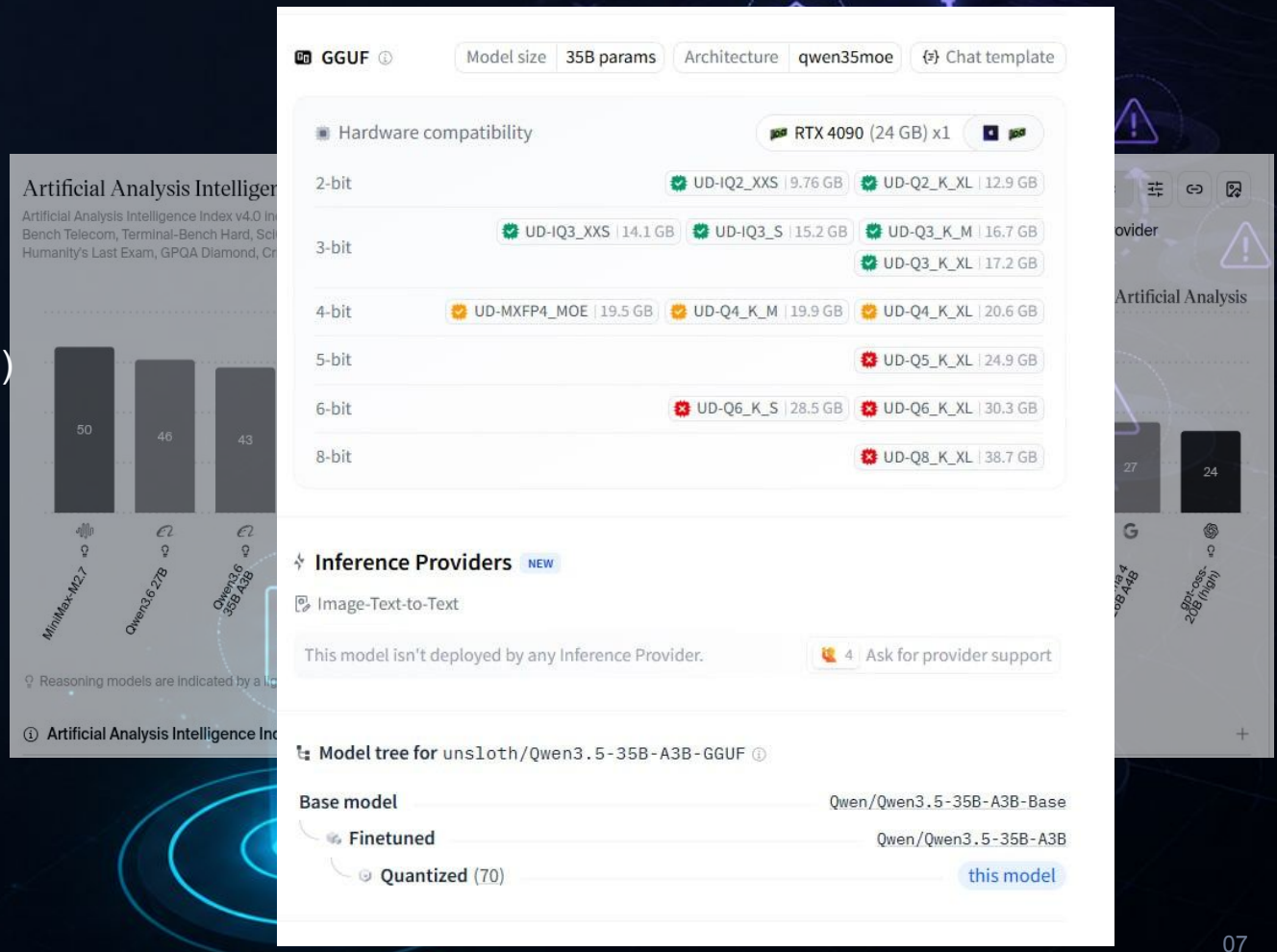
Artificial Analysis Coding Index
Represents the weighted average of **coding benchmarks** in the Artificial Analysis Intelligence Index (**Terminal-Bench Hard, SciCode**)

02 CONSIGLI PRATICI

Da dove partire, cosa evitare

Procedi con metodo: **Scenario -> Modello -> Verifica**

1. Definisci lo scenario
2. Identifica modelli soddisfacenti (accuratezza)
3. Filtra in base alla memoria disponibile (quantizzazione?)

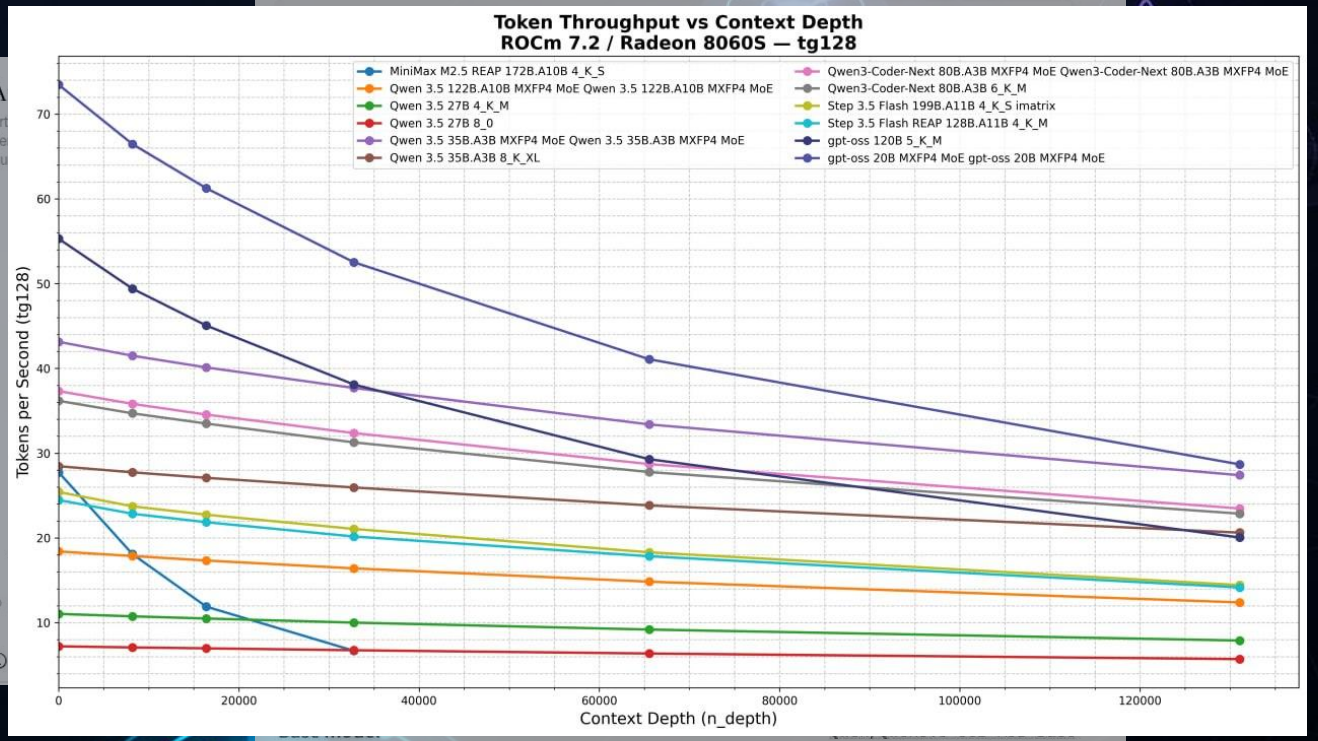


02 CONSIGLI PRATICI

Da dove partire, cosa evitare

Procedi con metodo: **Scenario -> Modello -> Verifica**

1. Definisci lo scenario
2. Identifica modelli soddisfacenti (accuratezza)
3. Filtra in base alla memoria disponibile (quantizzazione?)
4. Misura il tuo workflow, non solo il benchmark sintetico
   a. Generazione

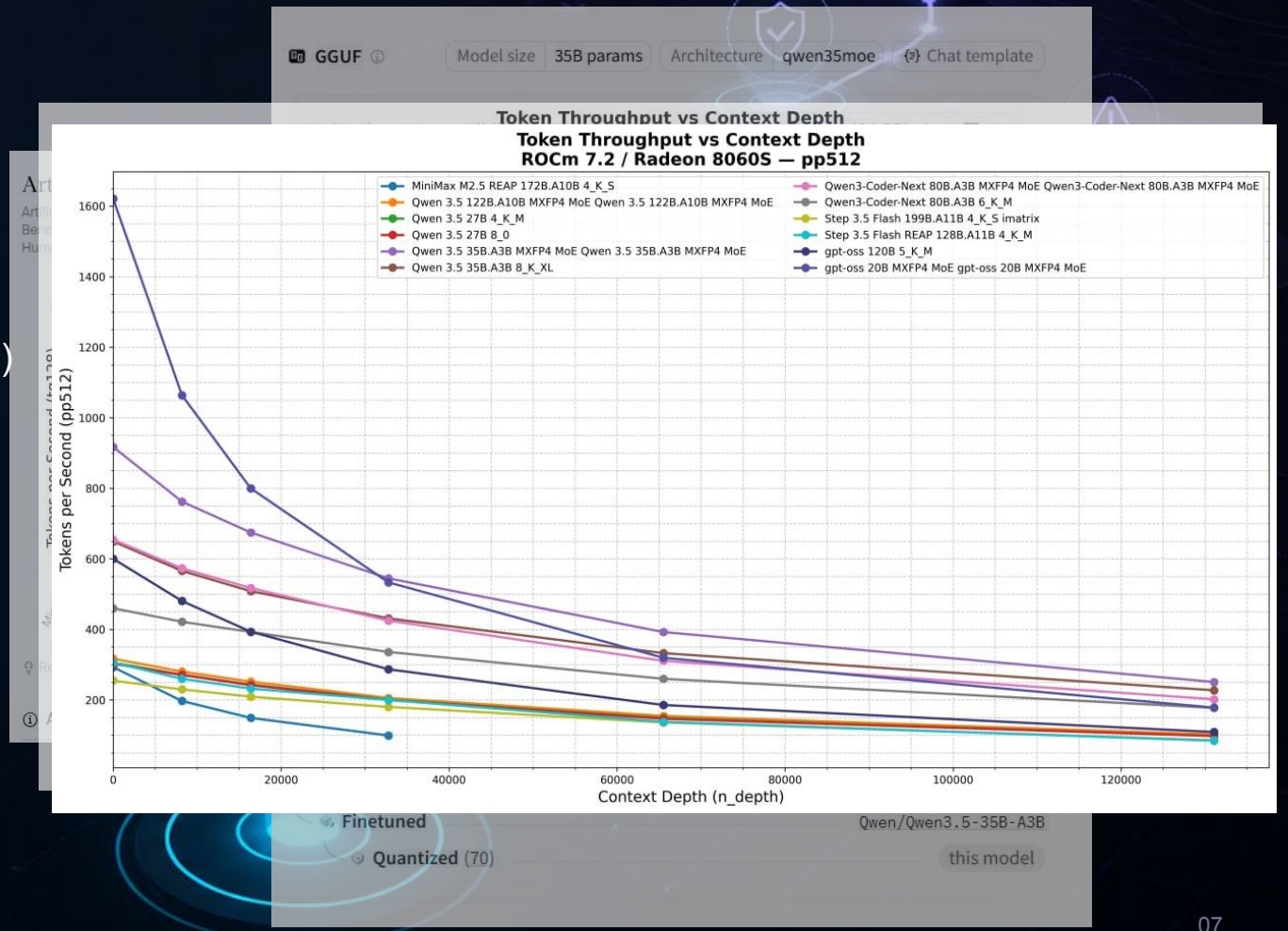


02 CONSIGLI PRATICI

Da dove partire, cosa evitare

Procedi con metodo: **Scenario -> Modello -> Verifica**

1. Definisci lo scenario
2. Identifica modelli soddisfacenti (accuratezza)
3. Filtra in base alla memoria disponibile (quantizzazione?)
4. Misura il tuo workflow, non solo il benchmark sintetico
   a. Generazione
   b. Prefill



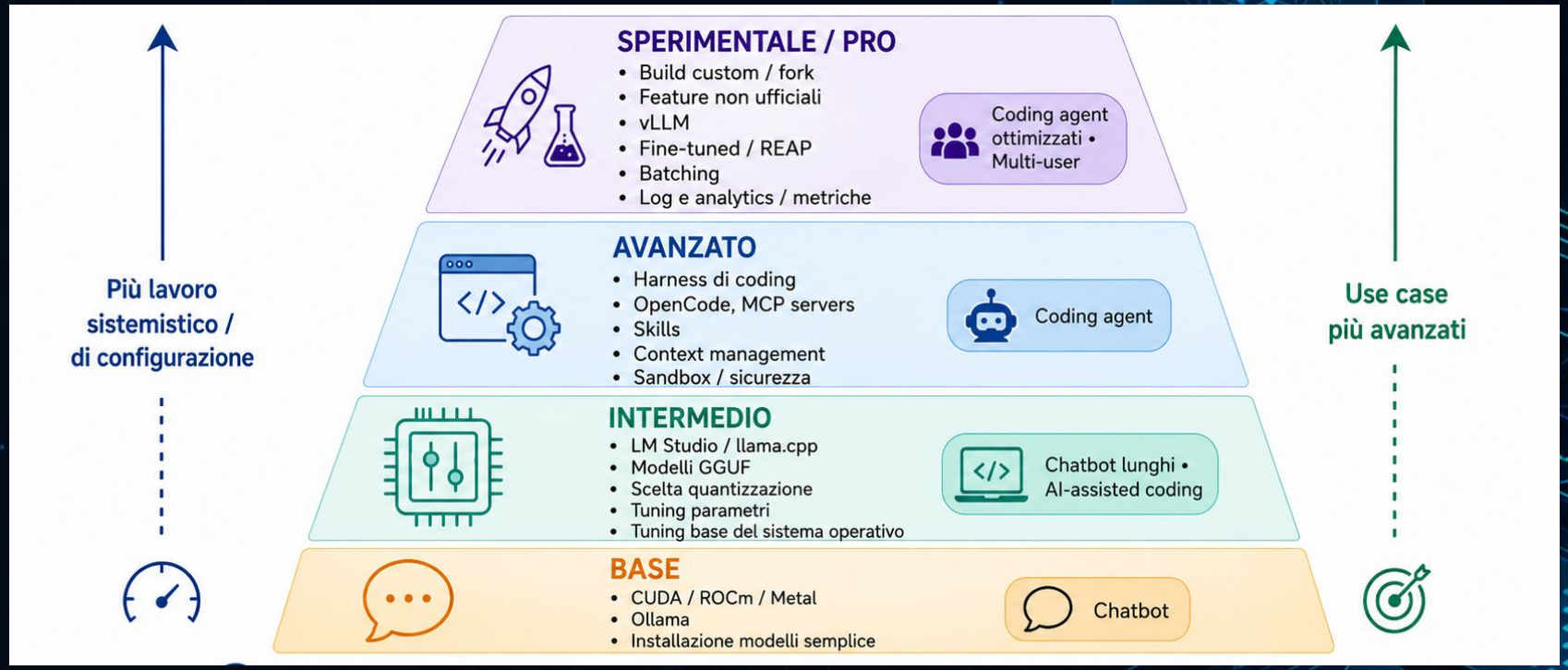
02 CONSIGLI PRATICI

Da dove partire, cosa evitare

Procedi con metodo: **Scenario -> Modello -> Verifica**

1. Definisci lo scenario
2. Identifica modelli soddisfacenti (accuratezza)
3. Filtra in base alla memoria disponibile (quantizzazione?)
4. Misura il tuo workflow, non solo il benchmark sintetico
   a. Generazione
   b. Prefill
5. Scegli in funzione delle performance minime richieste

Lo stack: quello che non vedi coi modelli in cloud



Takeaway: il modello è solo un pezzo del sistema.

04
HARDWARE

Che macchina serve davvero?

La RAM decide cosa puoi caricare; bandwidth, backend e ottimizzazioni decidono quanto sarà piacevole usarlo.

RAM / VRAM

Dimensione del modello, contesto e KV cache.

GPU

Velocità (token/s) prefill e generazione

Architettura & driver

NVIDIA, AMD, Intel, Apple
CUDA, ROCm, OpenVino, MLX, …

Carico

Scenari di utilizzo, numero utenti / concorrenza, …

04
HARDWARE

Che macchina serve davvero? (a casa, utente singolo…)

La RAM decide cosa puoi caricare; bandwidth, backend e ottimizzazioni decidono quanto sarà piacevole usarlo.

Categoria	Prefill	Generazione	Memoria utile
1. Laptop high-end RTX, 8–16 GB VRAM + 32 GB RAM	Medio-alto	Medio-alto	Bassa
2. Desktop gaming/high-end RTX, 16–24 GB VRAM, o 32 GB con RTX 5090	Molto alto	Molto alto	Medio-bassa
3. AMD AI Max+ 395 / Strix Halo, 128 GB condivisi	Medio	Medio-basso/medio	Alta
4. Apple M4 Max / M3 Ultra, 96–128 GB+ condivisi	Medio-alto	Medio	Alta
5. NVIDIA DGX Spark, 128 GB	Alto	Alto	Alta

RAM / VRAM

Dimensione del modello, contesto e KV cache.

GPU

Velocità (token/s) prefill e generazione

Architettura & driver

NVIDIA, AMD, Intel, Apple
CUDA, ROCm, OpenVino, MLX, …

Carico

Scenari di utilizzo, numero utenti / concorrenza, …

04
HARDWARE

Che **macchina** serve davvero? (a casa, utente singolo…)

La RAM decide cosa puoi caricare; bandwidth, backend e ottimizzazioni decidono quanto sarà piacevole usarlo.

Categoria	Prefill	Generazione	Memoria utile
1. Laptop high-end RTX, 8–16 GB VRAM + 32 GB RAM	Medio-alto	Medio-alto	Bassa
2. Desktop gaming/high-end RTX, 16–24 GB VRAM, o 32 GB con RTX 5090	Molto alto	Molto alto	Medio-bassa
3. AMD AI Max+ 395 / Strix Halo, 128 GB condivisi	Medio	Medio-basso/medio	Alta
4. Apple M4 Max / M3 Ultra, 96–128 GB+ condivisi	Medio-alto	Medio	Alta
5. NVIDIA DGX Spark, 128 GB	Alto	Alto	Alta

RAM / VRAM

Dimensione del modello, contesto e KV cache.

GPU

Velocità (token/s) prefill e generazione

Architettura & driver

NVIDIA, AMD, Intel, Apple
CUDA, ROCm, OpenVino, MLX, …

Carico

Scenari di utilizzo, numero utenti / concorrenza, …

05 TRADE-OFF

Locale vs cloud: il confine giusto

La scelta non è una questione di analisi e compromessi: si considerano qualità, costi, controllo, latenza, affidabilità…

- Locale: privacy, controllo, personalizzazione, investimento iniziale importante (specie per supporto multi utente)
- Cloud: qualità massima, comodità, contesto lungo e tool maturi
- Coding: locale utile, ma cloud resta riferimento per compiti difficili
- Architettura realistica: ibrida, non “tutto locale”

Qualità

Cloud spesso superiore; locale abbastanza buono per molti task.

Privacy

Locale forte quando dati e repository non devono uscire.

Costo

Locale costa prima; cloud costa a consumo o abbonamento.

Affidabilità

Nel locale sei anche l’ops team del tuo stack AI.

Takeaway: il locale vince quando controllo e privacy valgono più di comodità e prestazioni.