Ollama — panoramica rapida

Cos’è, come funziona, storia, Cloud e prospettive (18 ottobre 2025)

1) Cos’è Ollama

- Runtime leggero per eseguire modelli open in locale (LLM, embeddings, vision).
- Multipiattaforma: macOS, Linux, Windows. Interfacce: CLI + API REST.
- Obiettivi: semplicità d’installazione, privacy (dati in locale), integrazione veloce.
- App desktop ufficiale (macOS/Windows) per chat, drag&drop di file/PDF, impostazioni di contesto.

2) Cenni tecnici: piattaforme e GPU

- Supporto CPU e accelerazione GPU (MPS su Apple Silicon, CUDA su NVIDIA, ROCm dove disponibile).
- Funziona anche solo CPU (prestazioni inferiori).
- Servizio locale: porta 11434 (API). Sessione interattiva via CLI o client.

2) Cenni tecnici: modelli & Modelfile

- Formato modelli: GGUF (ecosistema llama.cpp).
- Gestione: `ollama pull <modello>`, `ollama run <modello>`, `ollama list`, `ollama show`.
- Personalizzazione con Modelfile: FROM, SYSTEM, TEMPLATE, PARAMETER, ADAPTER, MESSAGE, LICENSE.
- Build personalizzata: `ollama create mio-modello -f Modelfile`.

2) Cenni tecnici: API e integrazioni

- API REST nativa per chat/completions, embeddings, gestione modelli e streaming.
- Compatibilità (sperimentale) con API OpenAI per riuso di SDK e client esistenti.
- Ecosistema: integrazioni con framework (es. LangChain) e frontend come Open WebUI.

2) Cenni tecnici: scheduling & performance

- Scheduler di memoria/VRAM: stime più accurate, riduzione OOM, supporto multi-GPU.
- Parametri comuni: temperature, top-p, top-k, max tokens, repeat penalty, num_ctx.
- Best practice: pinna i modelli usati, verifica VRAM disponibile, non esporre 11434 su Internet.

3) Cenni storici: autori

- Fondatori: Jeffrey Morgan (CEO) e Michael Chiang.
- Y Combinator Winter 2021. Sede: Palo Alto (CA).
- Evoluzione: da tool CLI per LLM locali (2023) a piattaforma con app, librerie e orchestrazione (2024–2025).

3) Cenni storici: finanziamenti

- Seed iniziale contenuto; le cifre pubbliche variano (ordine centinaia di migliaia di USD).
- Investitori citati: Y Combinator, Essence VC; fonti indicano anche Rogue Capital e Sunflower Capital.
- Modello product-led con forte trazione open-source.

4) Versione Cloud (preview)

- Cloud models su infrastruttura remota con la stessa UX/API della versione locale.
- Attivazione: `ollama signin`; modelli con suffisso `:cloud`. Nessun cambio codice nelle integrazioni.
- Use-case: modelli più grandi/veloci senza dGPU; ibrido locale+cloud per carichi variabili.

5) Prospettive & alternative

- Direzione: ibrido locale+cloud, scheduler più efficiente, focus su privacy e sicurezza.
- Alternative/adiacenti: llama.cpp (puro), LM Studio (GUI), Open WebUI (frontend per Ollama), altri runtime locali.
- Per il talk: layout in orizzontale, font grandi, margini ampi per proiezione.