

Come tradurre le tecnologie leggere nel quotidiano: esempi nel campo professionale

L'IA che non pesa sul pianeta

Small Language Model, strumenti locali e sovranità del dato

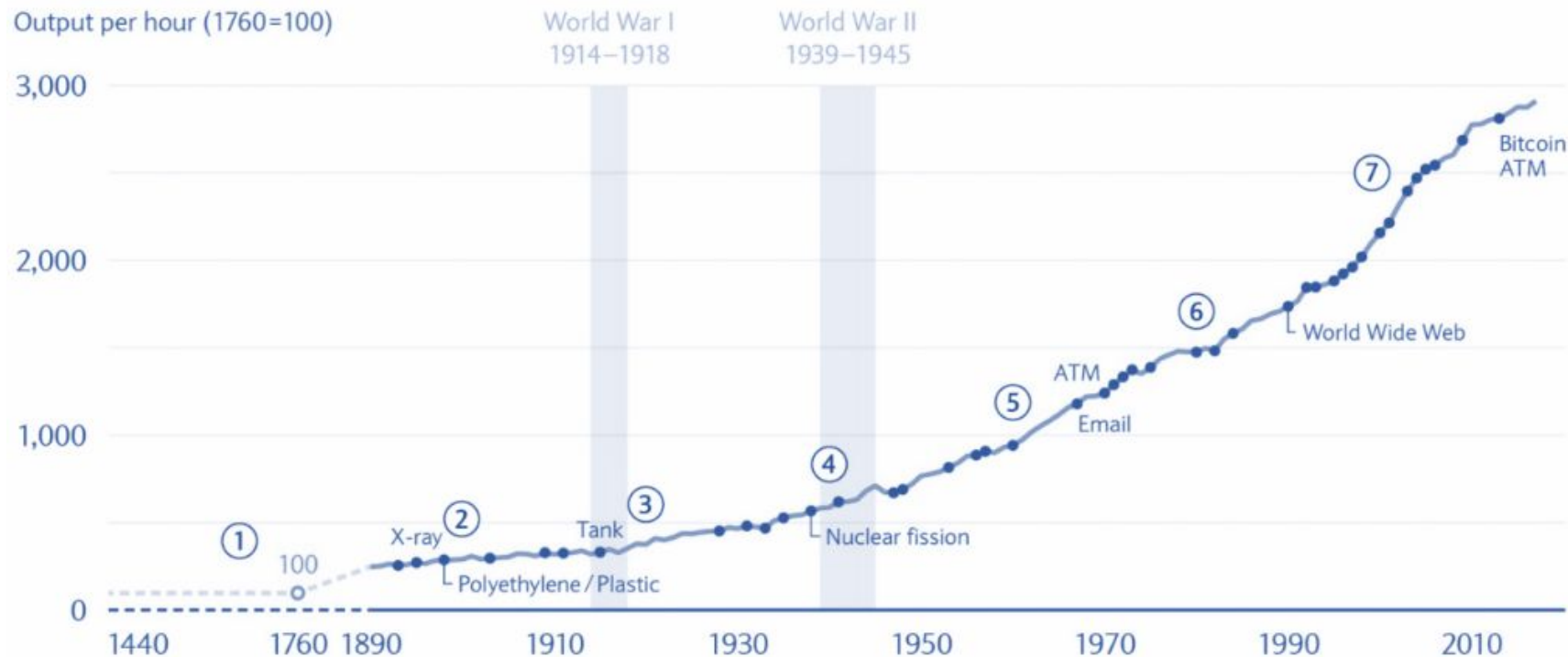
Ing. Massimiliano De Rose

Evento: Il paradosso dell'IA: potenza computazionale o sostenibilità reale?

2 luglio 2026 — Centro Congressi FAST, Milano

Se la conservazione dell'energia è dovuta al fatto che le leggi della fisica non cambiano nel corso del tempo, potrebbe essere che in particolari, rarissime circostanze, la simmetria sia spezzata? L'eventuale rottura della simmetria, in luoghi esotici e inattesi, costituisce la possibilità che il principio di conservazione venga violato su scala cosmica.

[Michio Kaku – La fisica dell'impossibile]



KEY BREAKTHROUGHS

FIRST WAVE

During the Industrial Revolution, the first factory emerged—a cotton mill in Britain.



SECOND WAVE

As railways proliferated, their networks strongly influenced urban growth.

Source: Nacima Baron, HAL



THIRD WAVE

Henry Ford's Model T introduced the assembly line, revolutionizing the automotive industry.



FOURTH WAVE

Aviation gains mass adoption on a global scale, providing a lever to economic integration.

Source: OECD



FIFTH WAVE

In 1990, 2.3M used the internet—by 2016 this reached 3.4B.

Source: World Bank



SIXTH WAVE

As climate challenges intensify, clean tech may reshape business models and consumption patterns.

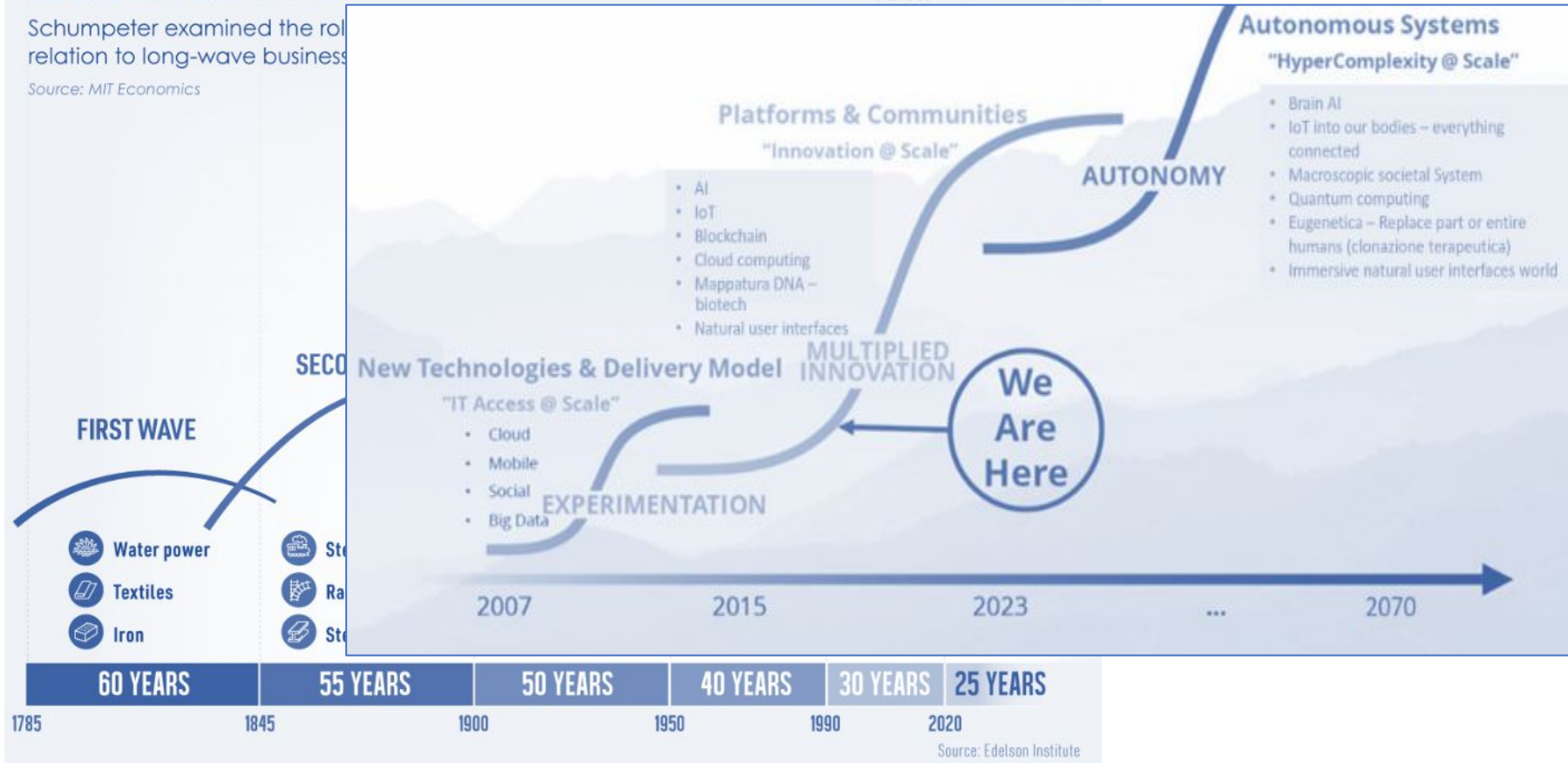


LONG WAVES OF INNOVATION

The theory of innovation cycles was developed by economist **Joseph Schumpeter** who coined the term 'creative destruction' in 1942.

Schumpeter examined the role of innovation in the relation to long-wave business cycles.

Source: MIT Economics



Cos'hanno in comune queste invenzioni ?





MATEMATICA

STATISTICA

DATI

POTENZA DI CALCOLO

Intelligenza frugale: fare di più con meno

Large Language Model (LLM)

Modelli come GPT-4 o Claude possiedono centinaia di miliardi di parametri. Richiedono infrastrutture cloud enormi, connessione costante e comportano alti costi energetici ed economici.

Small Language Model (SLM)

Modelli come Llama 3, Phi-3, Mistral o Gemma hanno tra 1 e 13 miliardi di parametri. Sono ottimizzati per girare su un **normale laptop** con GPU integrata.

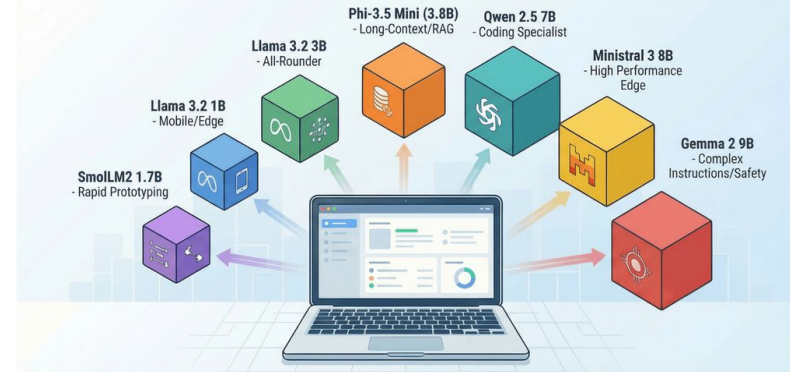
Nessun dato in rete

Latenza quasi nulla

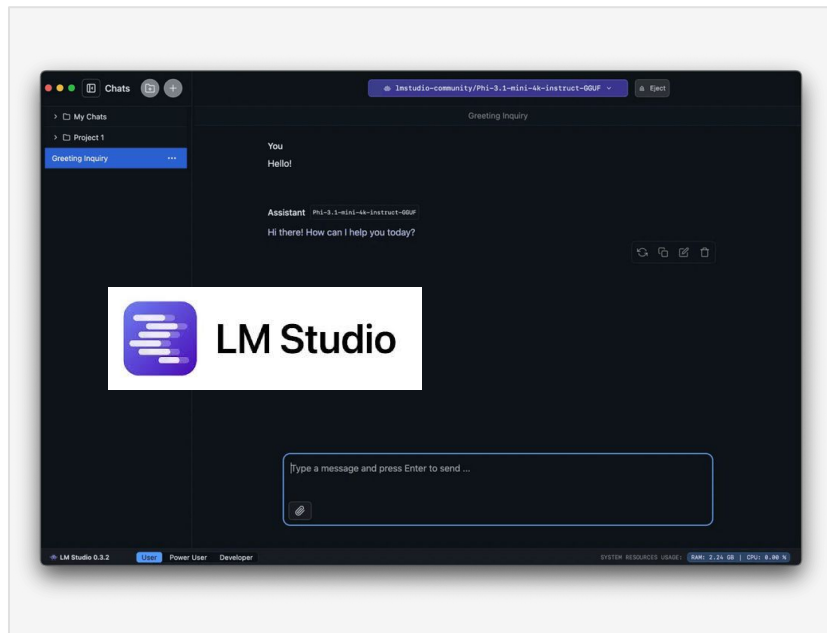
Funzionamento offline

Costo operativo zero

TOP 7 SMALL LANGUAGE MODELS YOU CAN RUN ON A LAPTOP

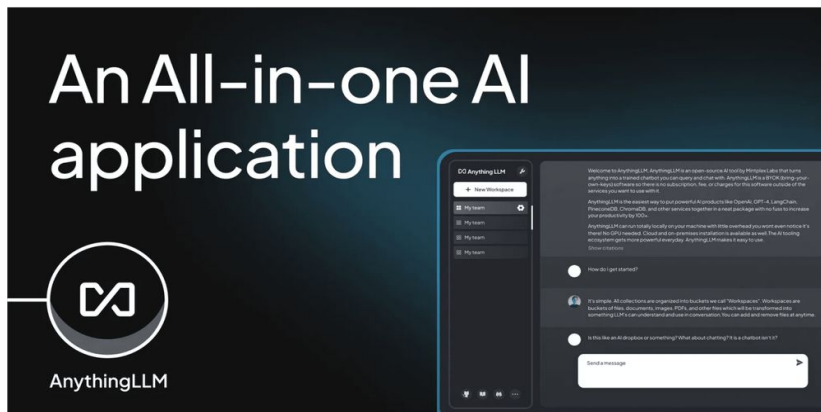


LM Studio: scarica, esegui, conversa tutto in locale



- **LM Studio** è un'applicazione desktop (Windows, macOS, Linux) che permette di scaricare e far girare modelli open-source direttamente sul proprio computer.
- Interfaccia grafica intuitiva, **nessuna riga di codice** necessaria.
- Supporta modelli in formato GGUF: Llama, Mistral, Phi, DeepSeek e molti altri.
- Espone un **server locale compatibile con le API OpenAI**: integrazione immediata con qualsiasi applicazione esistente.
- **Ideale per**: prototipazione rapida, test di modelli, uso professionale senza dipendenza da cloud.

AnythingLLM: il tuo archivio documentale con AI incorporata

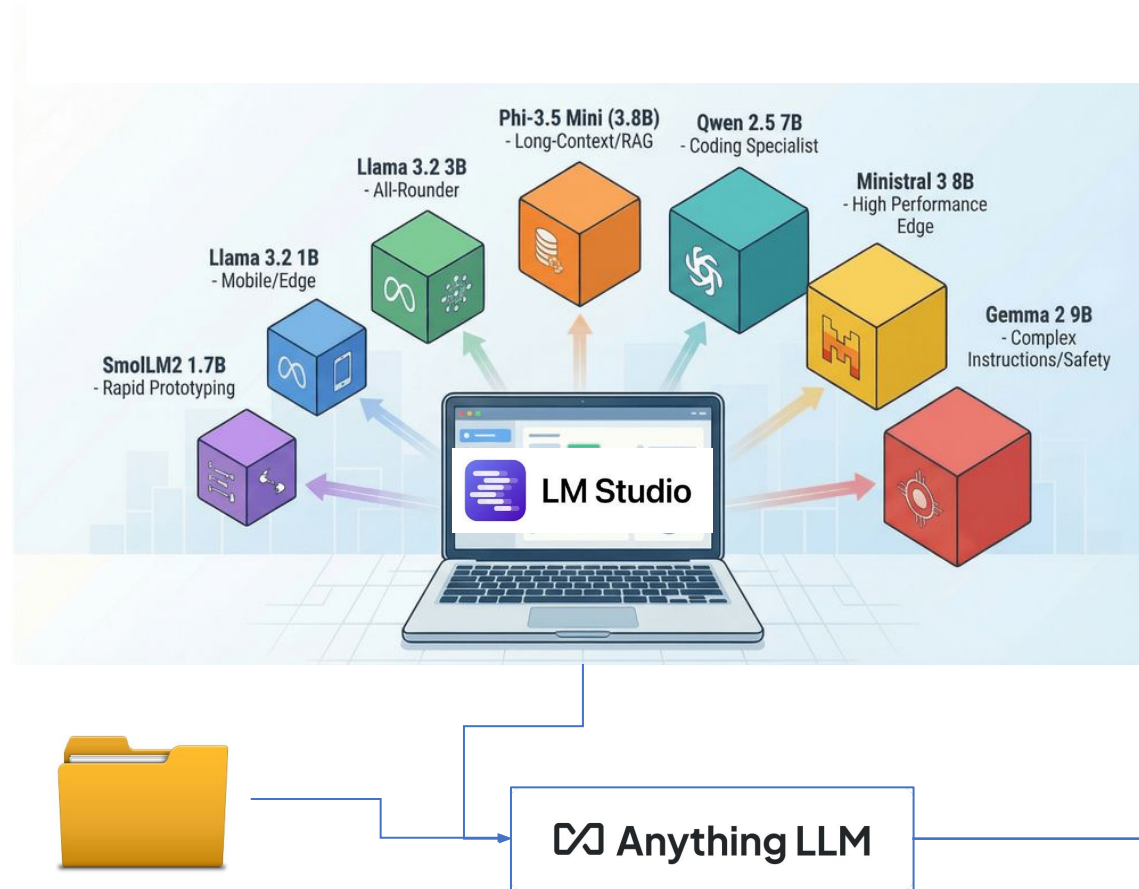


Tecnica RAG (Retrieval-Augmented Generation)

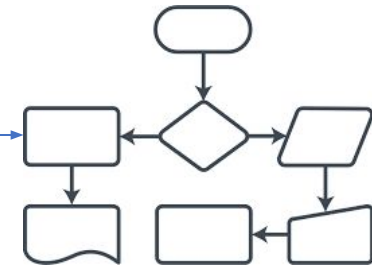
L'AI risponde attingendo esclusivamente ai tuoi documenti, garantendo risposte contestuali e non basate su dati generici di internet.

- **Funziona interamente in locale:** PDF, Word, pagine web, note — tutto rimane sul tuo dispositivo.
- Si collega a **LM Studio** (o Ollama) come motore di inferenza locale.
- **Casi d'uso:** interrogare archivi progettuali, normative tecniche, documentazione interna, contratti.
- **Zero data leakage:** i dati aziendali non escono mai dal perimetro aziendale.

Un sistema “domestico”



 **python**[™]



Agent

Cloud o locale? La scelta dipende dal contesto

Aspetto	AI Cloud (GPT, Claude...)	AI Locale (SLM + LM Studio)
Privacy dei dati	Dati inviati al provider	Dati mai in rete
Costo	Abbonamento / token	Gratuito dopo setup
Qualità su task generici	Eccellente	Buona/ottima
Qualità su dati propri (RAG)	Buona	Eccellente
Consumo energetico	Alto (data center)	Basso (laptop)
Disponibilità offline	No	Sì
Aggiornamento modelli	Automatico	Manuale

Manus: dall'assistente all'agente che agisce

- **Manus** è un agente AI autonomo di nuova generazione: non risponde solo a domande, ma pianifica ed esegue task complessi in modo autonomo (navigazione web, scrittura codice, analisi dati, creazione documenti).

CICLO OPERATIVO

Analizza l'obiettivo → Pianifica i passi → Usa strumenti → Verifica → Consegna

- Disponibile come servizio cloud, ma anche **self-hostable** su server privato o on-premise.
- In modalità self-hosted: nessun dato esce dal dominio aziendale, garantendo il **controllo totale sull'infrastruttura**.



manus
The general AI agent

Manus sul tuo server: potenza agentiva senza rinunciare alla privacy

- **Installazione on-premise:** deploy su server Linux aziendale o VPS privata tramite Docker.
- **Integrazione locale:** si collega direttamente a modelli locali (Ollama, LM Studio) o ad API private.
- **Zero esposizione:** accesso via browser dalla rete interna, nessuna connessione a internet necessaria per i dati sensibili.
- **Scenari applicativi:** analisi automatizzata di documenti tecnici, redazione di relazioni, ricerca normativa, supporto alla progettazione.

Il professionista rimane supervisore del processo: l'agente esegue, l'esperto valida.

The Agent Loop: How Manus AI Solves Tasks



Manus mimics human reasoning:
perception → planning → action → feedback → delivery.

L'AI sostenibile non è un compromesso, è una strategia

- Usare SLM in locale non significa rinunciare alla qualità: significa **scegliere lo strumento giusto per il task giusto**.
- La sostenibilità dell'AI non è solo energetica: è anche **sovranità del dato, indipendenza infrastrutturale** e **controllo professionale**.
- Il professionista del futuro non è chi usa l'AI più potente, ma chi sa **scegliere quella più appropriata**.
- Gli strumenti esistono già, sono gratuiti e open-source: **LM Studio, AnythingLLM, Manus** — pronti all'uso oggi.

**Sperimenta in locale. Valuta cosa serve davvero.
Scegli con consapevolezza.**



[Massimiliano De Rose]