

AI senza permessi

L'ascesa dei modelli Open Weight

Stefano Driussi

Open Source: come sfruttare il lavoro altrui per massimizzare i profitti degli shareholder

Piano strategico FY2027 · Offsite del Consiglio di Amministrazione

Fork & rebrand

Prendi il progetto, cambia il logo, chiudi il codice. Tanto nessuno legge le licenze.

Relicensing tattico

Quando la community ha fatto il grosso del lavoro, passa a una licenza “source available”.

Upstream? No, grazie

Le patch restano interne: perché regalare valore ai concorrenti?

Di cosa parliamo

01	Cos'è l'AI oggi	Transformer e Diffusion
02	Aperto quanto?	Pesi, sorgenti e dati: open weight contro open source
03	L'ascesa	Da un leak del 2023 ai modelli da migliaia di miliardi di parametri
04	In pratica	Licenze, strumenti per Linux, hardware e zone d'ombra

Reti neurali enormi che imparano da esempi, non da regole scritte a mano.

[TESTO](#) · [CODICE](#) · [AGENTI](#)

Transformer

Prevede il prossimo pezzo di testo, un token alla volta. È il motore di chatbot, assistenti di programmazione e agenti.

[IMMAGINI](#) · [VIDEO](#) · [AUDIO](#)

Diffusion

Parte da puro rumore e lo ripulisce passo dopo passo, finché emerge ciò che hai descritto.

In entrambi i casi l'addestramento produce un file: **i pesi**. Miliardi di numeri che, di fatto, sono il modello.

Transformer: indovina la parola successiva

01 →

Tokenizzazione

Il testo diventa una sequenza di pezzi: i token.

[il][ping][uino]

02 →

Embedding

Ogni token diventa un vettore di migliaia di numeri.

[0.12, -0.8, ...]

03 →

Attenzione

Ogni token “guarda” gli altri e pesa cosa conta nel contesto.

× decine di strati

04

Previsione

Probabilità sul token successivo: ne sceglie uno, lo aggiunge, ricomincia.

è 62% · sta 21%

Mixture-of-Experts: i modelli di punta hanno migliaia di miliardi di parametri, ma per ogni token ne attivano solo una frazione. Kimi K3: 2.800 miliardi totali, 104 miliardi attivi.



01 / UN TOKEN ALLA VOLTA

Cosa mettiamo sulla pizza?

La domanda entra a destra, attraversa embedding, decine di strati di transformer e unembedding, ed esce come lista di probabilità.

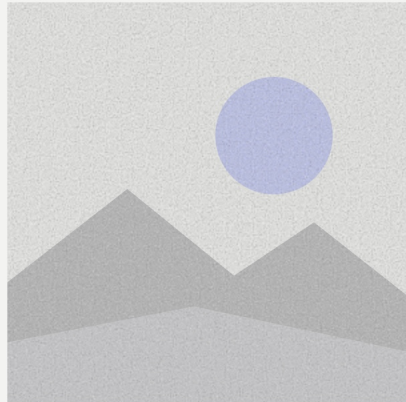
Pepperoni era al **46%**, eppure è uscito **pineapple** (22%).

Il modello non prende sempre il più probabile: estrae a sorte, pesando le probabilità. La “temperatura” regola quanto rischiare.

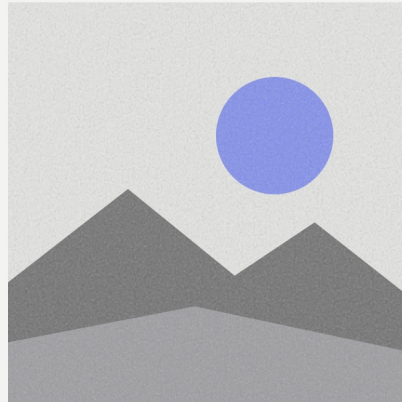
Diffusion: dal rumore all'immagine



$t = 0$ · rumore



passo 10



passo 25



passo 40



immagine

Il modello impara a togliere un po' di rumore alla volta. Il prompt, trasformato in vettori da un encoder di testo, guida ogni passaggio.

Cosa c'è dentro un “modello”

Componente	Che cos'è	Nel mondo del software
Pesi	File da GB a TB: miliardi di numeri appresi	Il binario compilato
Codice di inferenza	Carica i pesi e genera le risposte	Il runtime, l'interprete
Codice di training	La ricetta che produce i pesi	Makefile e toolchain
Dati di training	Migliaia di miliardi di token: testo, codice, immagini	Il codice sorgente

Avere solo i pesi è come avere un `.deb` senza il pacchetto sorgente.

Chiuso, open weight, open source

Cosa ottieni	Chiuso (API)	Open weight	Open source
Usarlo	Sì, alle loro condizioni	Sì	Sì
Scaricare i pesi	No	Sì	Sì
Eseguirlo offline	No	Sì	Sì
Fine-tuning	Limitato, sul loro cloud	Sì	Sì
Codice di training	No	Raramente	Sì
Dati di training	No	Quasi mai	Sì, o descritti in dettaglio
Esempi	Claude, GPT, Gemini	Llama, Qwen, DeepSeek, Kimi, GLM	OLMo, Pythia

OSI ha scritto la definizione: OSAID 1.0

Ottobre 2024. Un sistema di AI è open source se garantisce quattro libertà:

Usare

Studiare

Modificare

Condividere

Servono pesi, codice e informazioni sui dati sufficienti a ricostruire un sistema equivalente.

SUPERANO LA VALIDAZIONE

OLMo (AI2) · Pythia (EleutherAI) · T5 (Google) ·
CrystalCoder (LLM360)

NON LA SUPERANO

Llama 2 (Meta) · Grok (xAI) · Phi-2 (Microsoft) ·
Mixtral (Mistral)

Come ci siamo arrivati

2022

Stable Diffusion

Pesi pubblici: la generazione di immagini esce dai laboratori.

2023

Il leak di LLaMA

I pesi di Meta finiscono online. Nasce llama.cpp: un LLM gira su un laptop.

2023

Mistral 7B

Apache 2.0, dall'Europa: piccolo, veloce, davvero libero.

2025

DeepSeek R1

Ragionamento di frontiera con licenza MIT. I mercati tremano.

2025

gpt-oss

Anche OpenAI torna a pubblicare pesi, in Apache 2.0.

2026

L'era dei giganti

DeepSeek V4 (MIT), Kimi K3, GLM-5: migliaia di miliardi di parametri, scaricabili.

In tre anni: da “pesi trapelati” a **strategia industriale**.

Ottobre 2026: quanto manca ai modelli chiusi?

22%

dei 50 migliori modelli nella classifica
BenchLM è open weight

-12

punti su 100 tra il miglior modello chiuso
e il miglior aperto

104

modelli open weight in classifica,
aggiornata al 20 ottobre

IN VETTA TRA GLI APERTI

MiMo-V2.6-Pro

Kimi K3

GLM-5.3

Qwen3.8

DeepSeek V4.1

Open weight sì, ma a quali condizioni?

Licenza	Cosa permette	Esempi
Apache 2.0	Uso commerciale libero, con concessione di brevetti	gpt-oss, Qwen3.6, Mistral Small 4, Gemma 4, Wan 2.2, Z-Image
MIT	Quasi tutto: basta mantenere l'avviso di copyright	DeepSeek V4, GLM-5
“Community”	Soglie di fatturato, attribuzione obbligatoria, esclusioni geografiche	Llama 4 (multimodale non concesso nell'UE), Muse Spark Open, Kimi K3
Non commerciale	Studio e uso personale; il resto si negozia	FLUX.2 [dev]

Regola pratica: leggete la licenza dei pesi come leggereste quella di una libreria prima di linkarla.

Perché può essere interessante?

01

Privacy

I dati non lasciano mai la tua macchina.

02

Sovranità

Nessuno può cambiarti prezzi o condizioni da un giorno all'altro.

03

Offline

Funziona senza rete: in aereo, in laboratorio, in reti isolate.

04

Studio

Puoi misurarlo, smontarlo, criticarlo e pubblicare i risultati.

05

Su misura

Fine-tuning e LoRA sui tuoi dati, per i tuoi casi d'uso.

06

Costi prevedibili

Paghi corrente e hardware, non i token.

Lo stack libero: tutto gira su Linux

TESTO E CODICE

llama.cpp · Ollama · vLLM · Open WebUI

IMMAGINI E VIDEO

stable-diffusion.cpp · ComfyUI · Diffusers

FORMATI

GGUF · safetensors · quantizzazione a 4 bit

CHE HARDWARE SERVE?

Macchina	Modelli realistici
Una GPU consumer o un Mac	Qwen3.6-27B, Gemma 4, Mistral Small 4
Una GPU da 80 GB	gpt-oss-120b
Un server multi-GPU	Kimi, GLM-5, DeepSeek V4: centinaia di GB anche quantizzati

Aperti a tutti, alla portata di pochi

2,8 T

parametri di Kimi K3, 104 B attivi

1,56 TB

solo i pesi, già a 4 bit

~2 TB

di memoria GPU per servirlo

Scenario	Hardware	Costo	Consumo
Kimi K3 in casa	2 server 8× NVIDIA B200	oltre 1 M\$	~29 kW, ~250 MWh/anno (≈ 90 famiglie)
Kimi K3 a noleggio	16 B200 in cloud	~110 \$/ora, ~80.000 \$/mese	a carico del provider
Kimi K2 (1 T) da appassionati	4 Mac Studio M3 Ultra 512 GB	~40.000 \$	~450 W di picco (cluster), ~25–30 token/s

Fonti: yottalabs.ai, morphllm.com (ott. 2026), Jeff Geerling, Creative Strategies, Apple. Server: stime a pieno carico.

Le zone d'ombra

Dati opachi

Quasi nessuno pubblica i dati di training: bias, copyright e benchmark “contaminati” restano difficili da verificare.

La barriera hardware

I modelli più forti chiedono centinaia di GB di memoria: aperti, ma non per tutti.

Open-washing

“Open source” nel comunicato stampa, licenza restrittiva nei fatti.

Pochi possono addestrare

Scaricare è gratis; addestrare un modello di frontiera costa milioni. Il potere resta concentrato.

E la sicurezza: i pesi pubblici non si possono ritirare. [Nel bene e nel male.](#)

Il software libero ci ha insegnato a pretendere il sorgente.

Con l'AI, i pesi sono solo l'inizio.

Grazie. Domande?

05 / Feedback

Com'è andata?

Inquadra il QR code e lascia il tuo parere: bastano un paio di minuti.

forms.lugud.it/435964

