L’AI sta divorando elettricità, chip e capitali: dentro la macchina industriale dei prompt

Introduzione

Ogni volta che qualcuno scrive “fammi un riassunto”, “crea questa immagine” oppure “genera un video realistico”, da qualche parte nel mondo si accendono migliaia di GPU.L’intelligenza artificiale generativa appare immateriale: una finestra di chat, una risposta in pochi secondi, un’immagine prodotta quasi istantaneamente. In realtà è una delle infrastrutture industriali più energivore e capital intensive mai costruite dall’industria tecnologica.Dietro un prompt si muove una catena gigantesca: data center hyperscale; reti elettriche; sistemi di raffreddamento a liquido; fabbriche di semiconduttori avanzati; miniere di rame e terre rare; investimenti da centinaia di miliardi di dollari.La corsa all’AI non è più soltanto una competizione software. È una guerra per:
energia; chip; capacità produttiva; infrastrutture; accesso ai dati.Ed è probabilmente la più grande espansione computazionale dai tempi della nascita di Internet.

Il numero reale di prompt: la nuova misura del pianeta digitale

Nessuna azienda pubblica dati completi e verificabili sul numero reale di prompt elaborati ogni giorno. I motivi sono evidenti:segreto industriale; concorrenza; metriche interne differenti; enorme variabilità tra utenti consumer, enterprise e API.Le stime vengono quindi ricostruite indirettamente attraverso:
utenti attivi; traffico web; capacità dei cluster GPU; token elaborati; investimenti infrastrutturali; dichiarazioni pubbliche. Il quadro resta imperfetto, ma gli ordini di grandezza sono ormai chiari: l’AI generativa sta entrando nella dimensione delle infrastrutture planetarie.

ChatGPT e il traffico AI globale

Le stime più prudenti suggeriscono che ChatGPT gestisca ormai centinaia di milioni di prompt al giorno. Le valutazioni più aggressive parlano addirittura di oltre un miliardo di richieste quotidiane considerando:
utenti diretti; integrazioni API; utilizzi enterprise; agenti automatici; strumenti di coding. Nel frattempo Gemini sfrutta l’enorme ecosistema Google, Copilot è integrato in Windows e Microsoft 365, Meta AI può potenzialmente raggiungere miliardi di utenti attraverso Facebook, Instagram e WhatsApp. Claude cresce rapidamente nel settore enterprise. Perplexity tenta di ridefinire il motore di ricerca. Grok sfrutta la piattaforma X come vettore di distribuzione. La vera rivoluzione però è invisibile: l’AI sta smettendo di essere un’applicazione separata e sta diventando una funzione incorporata in tutto il software.

Prompt, token e utenti: tre cose completamente diverse

Uno degli errori più comuni nel dibattito pubblico consiste nel confondere:

Termine — Significato
Utente attivo — Persona che usa il servizio
Prompt — Singola richiesta inviata
Token — Unità elementare di testo elaborata dal modello

Quando si scrive una domanda, il modello non “vede” parole nel senso umano del termine. Vede token.

Una frase semplice può richiedere pochi token. Un contratto di 40 pagine può richiederne decine di migliaia.

E ogni token ha un costo:computazionale; energetico; economico. È questo il motivo per cui l’economia reale dell’AI non si misura tanto in utenti, quanto nella quantità totale di token elaborati.

L'AI consuma davvero così tanta energia?

La risposta breve è sì. Ma non nel modo semplicistico spesso raccontato.

Quanta elettricità c'è dentro un prompt?

Non esiste un valore universale. Il consumo dipende da:
dimensione del modello; numero di token; lunghezza del contesto; tipo di hardware; efficienza del data center; tipologia della richiesta. Una semplice domanda testuale può consumare relativamente poco. Ma: richieste lunghe; immagini; video; agenti autonomi; conversazioni vocali in tempo reale possono moltiplicare il costo computazionale.

Il salto più drammatico arriva con il video generativo. Generare immagini è già molto più pesante rispetto a una chat testuale. Generare video è un'altra categoria industriale.

Ogni sequenza richiede: coerenza temporale; memoria enorme; elaborazione frame per frame; bandwidth interna elevatissima. In prospettiva, il video AI potrebbe diventare uno dei principali motori della domanda energetica globale dei data center.

Dentro i data center dell'AI

La vera protagonista della rivoluzione AI non è ChatGPT. È la GPU.

NVIDIA e il dominio dell'infrastruttura

La H100 di NVIDIA è diventata il simbolo della nuova economia computazionale. Un singolo acceleratore: contiene decine di miliardi di transistor; utilizza memoria HBM ad altissima banda; consuma centinaia di watt; lavora in cluster da migliaia di unità. Le nuove architetture Blackwell spingono ancora più in alto: densità computazionale; consumo energetico; requisiti termici; necessità di raffreddamento. I rack AI moderni stanno raggiungendo densità energetiche considerate estreme solo pochi anni fa. Questo obbliga l'industria ad abbandonare il semplice raffreddamento ad aria.

Il ritorno dell'acqua nell'era digitale

Molti data center AI stanno tornando a utilizzare sistemi di raffreddamento liquido. Il motivo è brutale: l'aria da sola non basta più. Le GPU moderne concentrano enormi quantità di calore in spazi ridottissimi. Raffreddarle è diventato un problema di fisica industriale prima ancora che informatico.

Training vs inference: il grande equivoco

Nel dibattito pubblico si parla spesso dell'addestramento dei modelli, ma il vero nodo economico potrebbe diventare l'inference.

Fase — Descrizione
Training — Addestramento iniziale del modello
Inference — Utilizzo quotidiano da parte degli utenti

Il training richiede enormi quantità di GPU per settimane o mesi. Ma l'inference è continua. Se miliardi di persone iniziano a utilizzare AI integrate ovunque — browser, smartphone, office automation, motori di ricerca, sistemi operativi — il costo operativo quotidiano può superare quello dell'addestramento. Ed è qui che nasce il vero problema economico.

Perché molte piattaforme AI perdono denaro

L'AI moderna richiede investimenti giganteschi: GPU; elettricità; reti; storage; raffreddamento; data center; personale altamente specializzato. Molte piattaforme stanno adottando una strategia classica della Silicon Valley: "growth before profit". Prima conquistare il mercato. Poi cercare un modello economico sostenibile.

Il problema è che l'AI generativa non scala come i social network tradizionali. Ogni utente aggiuntivo produce un costo computazionale reale. Nel cloud classico il software veniva replicato quasi gratuitamente. Nell'AI ogni nuova richiesta attiva GPU costosissime. Questo rende la sostenibilità economica molto più complessa.

La geopolitica dell'AI passa da Taiwan

Dietro quasi tutta l'AI moderna esiste un collo di bottiglia fondamentale: TSMC. La produzione dei chip più avanzati è concentrata in modo impressionante. Questo crea: dipendenze geopolitiche; rischi sistemici; vulnerabilità industriali; tensioni strategiche.

Il dominio NVIDIA dipende non solo dal design dei chip, ma anche dalla capacità produttiva avanzata di Taiwan. Per questo motivo l'AI non è più soltanto una questione tecnologica. È diventata: politica industriale; sicurezza nazionale; strategia energetica; equilibrio geopolitico.

AI e consumo elettrico: il nuovo shock infrastrutturale

Per anni il mondo tecnologico ha parlato di software come qualcosa di leggero e immateriale. L'AI generativa sta riportando brutalmente al centro: centrali elettriche; trasformatori; linee ad alta tensione; acqua; rame; semiconduttori. Le big tech stanno investendo direttamente in: nucleare; small modular reactors; innovabili; contratti energetici dedicati. La ragione è semplice: senza energia non esiste AI.

Il rischio bolla esiste davvero?

Sì. Ma il termine "bolla" va usato con precisione. Esiste sicuramente: hype mediatico; marketing aggressivo; inflazione delle aspettative; corsa speculativa agli investimenti. Molti benchmark vengono presentati in modo quasi pubblicitario. La parola "AGI" viene spesso usata senza definizioni rigorose. Tuttavia sarebbe un errore considerare l'AI soltanto una moda. Anche nel caso di ridimensionamento finanziario, l'infrastruttura costruita rimarrà. Esattamente come accadde: con Internet; con il cloud; con le reti mobili. Molte aziende fallirono durante la bolla dot-com. Ma Internet trasformò comunque il pianeta.

2026–2035: cosa potrebbe accadere davvero

Gli scenari futuri dipendono da tre variabili: costo dell'energia; disponibilità di chip; efficienza algoritmica.

Scenario conservativo

L'AI continua a crescere ma rallenta: limiti energetici; costi troppo elevati; saturazione del mercato.

Scenario realistico

L'AI diventa infrastruttura standard: integrata nei sistemi operativi; presente nei motori di ricerca; incorporata nei software aziendali; diffusa negli smartphone.

Scenario espansivo

Esplodono: agenti autonomi; video generativo; robotica; AI multimodale continua. In questo caso la domanda energetica dei data center potrebbe crescere in modo estremamente aggressivo.

La vera domanda non è “quanto è intelligente l’AI”

La vera domanda è: quanto possiamo permetterci di alimentarla?

Perché l’intelligenza artificiale moderna non è soltanto software.

È una gigantesca macchina industriale fatta di energia; silicio; acqua; rame; capitale; geopolitica.

E il mondo ha appena iniziato a misurare le conseguenze di questa trasformazione.