

Quanto costa l'AI

RIVOLUZIONE ONLINE – 28 GENNAIO 2026

Quanto costa l'AI?

Energia, Acqua, Ambiente. Guida pratica all'AI sostenibile

Usiamo l'intelligenza artificiale molte volte al giorno. Ma ci siamo mai chiesti quanto “pesa” ogni nostra domanda in termini di energia e ambiente?

Quando chiediamo a ChatGPT di scrivere un'email, a Claude di riassumere un documento o a Gemini di generare un'immagine, la **percezione** è quella **di un'operazione puramente virtuale**. In realtà, dietro ogni risposta c'è un'infrastruttura fisica mastodontica: **data center che consumano energia reale, acqua** per il raffreddamento e **producono emissioni** concrete.

La buona notizia? Con una strategia di **consumo consapevole** possiamo continuare a sfruttare il potenziale rivoluzionario dell'AI riducendone drasticamente l'impatto.

Come sottolinea anche il Parlamento Europeo nel suo briefing [AI and the energy sector](#), un uso consapevole e mirato degli strumenti AI non solo riduce i costi energetici, ma contribuisce a un ecosistema digitale più sostenibile.

Consumo di energia

Nel 2026, i data center sono destinati a raddoppiare il loro **consumo elettrico** rispetto al 2022, portando la quota del settore **dal 1,5% a oltre il 3%** del totale globale. Questo balzo è trainato principalmente dall'espansione dell'AI.

Un singolo data center di grandi dimensioni consuma annualmente quanto 100.000 famiglie.

A livello globale l'AI rappresenta il 15% di questo consumo, parliamo di uno 0,225% dell'energia mondiale.

Può sembrare una percentuale minuscola a livello globale, sebbene il ritmo di crescita sia notevole. Ma le statistiche aggregate rischiano di mascherare un grande problema: **l'impatto si concentra su territori specifici**, creando pressioni insostenibili sulle reti elettriche e idriche locali. In **Irlanda**, ad esempio, i data center assorbono già oltre il **20%** dell'elettricità nazionale.

Quanto 'beve' l'AI?

L'energia è solo metà della storia. L'altra metà è l'acqua. I server dei data center si surriscaldano e devono essere costantemente raffreddati. Il metodo più efficiente per farlo è con l'acqua. Tanta acqua.

Un modello efficiente come Gemini consuma circa 0,26 millilitri d'acqua per query (l'equivalente di cinque gocce). Se ti sembra poco, moltiplicalo per miliardi di query e arrivi a milioni di litri. Ma il vero problema è il training: addestrare un modello di grandi dimensioni può richiedere **milioni di litri d'acqua, spesso prelevati da aree già in stress idrico**.

Google riporta che l'86% della sua acqua proviene da fonti a basso rischio, altri provider sono più opachi. Alcuni operatori europei hanno di recente firmato il [**Climate Neutral Data Centre Pact**](#) con cui si impegnano a riusare il calore dei server per il teleriscaldamento urbano.

La competizione per le risorse idriche per altro può avere un costo sociale molto pesante,

quindi fare pressione affinché si trovino soluzioni strutturali e una pianificazione intelligente al problema è fondamentale.

Efficienza globale e salute delle reti locali

Parlare solo in termini globali rende il problema astratto. L'AI non è un'entità eterea, ma un carico elettrico che grava su reti specifiche. Scegliere provider che dichiarano trasparenza sui dati di [**PUE \(Power Usage Effectiveness\)**](#) e che localizzano le infrastrutture dove le reti sono meno sature (e le rinnovabili più presenti) è il primo passo per una sostenibilità tecnologica reale.

Un dato dovrebbe farci riflettere più di tutti: **una query a ChatGPT consuma circa 10 volte più energia di una ricerca Google**: 0,3 watt-ora contro 0,03.

Può sembrare poco, ma moltiplicate per 700 milioni di query al giorno (solo GPT-4o) e otteniamo l'equivalente del fabbisogno elettrico di 35.000 abitazioni americane. Ogni singolo giorno.

Consigli pratici per un'AI a basso impatto

Non dobbiamo smettere di usare l'IA, ma dobbiamo imparare a governarla.

La buona notizia è che non serve diventare esperti di machine learning per fare la differenza. Bastano alcune scelte consapevoli.

1.

Il consumo dell'AI non è uniforme. Secondo l'International Energy Agency, la (report [Energy and AI](#)) la **scelta del modello giusto** infatti **può ridurre i consumi del 50-90%** senza sacrificare i risultati.

Scegli il modello giusto (Edge AI)

Non usare un "cannone per sparare a una mosca"!

Un esempio concreto. Se usi GPT-4o per correggere un errore grammaticale, stai consumando circa 0,30 Wh per query. Ma se usi GPT-4o-mini per lo stesso compito, scendi a 0,05 Wh. Un **risparmio energetico** dell' 83% per una correzione grammaticale identica.

I modelli più recenti come Claude 3.7 Sonnet e Gemini 2.5 si attestano intorno a 0,24-0,25 Wh per query, ottimizzati per bilanciare prestazioni ed efficienza. Sono perfetti per analisi complesse e scrittura creativa. Ma se ti serve solo un riassunto veloce o una traduzione base, i modelli "mini" fanno esattamente lo stesso lavoro consumando una frazione dell'energia.

E poi ci sono i modelli "reasoning" come DeepSeek R1 e o1. Questi mostri sono progettati per risolvere problemi matematici avanzati o debugging complesso, e consumano **30-50 volte più dei modelli premium**. Sono potentissimi, ma usarli per scrivere un post su LinkedIn è come prendere un SUV per andare in panetteria.

Usa il modello giusto per il tuo compito!

La regola d'oro è semplice: **inizia sempre dal modello più leggero adeguato al tuo caso d'uso**.

Hai bisogno di correggere una email? Usa un modello mini. Devi analizzare un contratto complesso? Allora sì, vale la pena usare un modello premium.

Nessuno prenderebbe un jet privato per andare al supermercato. Eppure facciamo esattamente questo quando usiamo GPT-4o per compiti che GPT-4o-mini farebbe benissimo.

Dopo più di 3 anni dall'esordio pubblico di ChatGpt, ora che l'AI è diventata familiare a molti, è il momento di sensibilizzarci e alfabetizzarci anche su questo.

2.

C'è una strada ancora più sostenibile: e sono i modelli che girano direttamente sul tuo computer o smartphone, senza mai toccare un data center. Si chiamano Small Language Models (SLM) e stanno diventando sempre più capaci.

Per micro-task, prediligi modelli **on-device** usando strumenti gratuiti come **LM Studio** o **Ollama**, che ti permettono di scaricare modelli (come Mistral NeMo o DeepSeek-R1 distillati) e usarli offline. Girano sul tuo hardware, niente server remoti, niente consumo di rete, niente emissioni di trasporto dati. E soprattutto: **privacy totale**. I tuoi dati non lasciano mai il dispositivo.

Certo, questi modelli sono meno potenti di GPT-4o. Ma per correzioni grammaticali, traduzioni base, o micro-task ripetitivi, fanno egregiamente il loro lavoro. E consumano solo l'energia del tuo computer, che comunque avresti acceso.

3.

Ogni parola che invii consuma energia. Un prompt scritto bene non è solo più efficace, ma risparmia token. Meno token significa meno calcoli, meno calore prodotto e meno energia consumata. Eliminare i fronzoli nei prompt può ridurre l'uso di token del 30-50%.

Un esempio: "Gentile intelligenza artificiale, potresti per favore aiutarmi a riassumere questo lungo testo che segue? Grazie mille per la tua disponibilità!" può diventare tranquillamente "Riassumi in 3 punti".

Stessa richiesta, metà dei token, metà dell'energia.

Inoltre, **dai indicazioni per limitare la lunghezza dell'output** a ciò che davvero ti serve: "Rispondi in massimo 100 parole (o rispondi in modo sintetico)". Risparmi energia evitando risposte verbose.

4.

Quando cambi argomento, inizia una nuova chat invece di continuare quella vecchia.

I modelli LLM (Large Language Models) funzionano tramite un meccanismo chiamato “Attention”. Quando invii un messaggio in una chat lunga, il sistema non legge solo la tua ultima frase, ma deve processare l'intera cronologia precedente (il “context window”) per generare una risposta coerente.

Dopo 15-20 scambi, stai inviando migliaia di token ripetuti ogni volta. Più la chat è lunga, più token (parole/frammenti) il server deve elaborare ad ogni singolo invio, più consumi.

I moderni sistemi di caching aiutano solo con contenuti identici tra conversazioni diverse, non all'interno della stessa chat.

5.

Il risparmio maggiore viene dal non usarla quando non serve.

Per un calcolo semplice, usa una calcolatrice. Per cercare un ristorante, usa Google Maps. Per tradurre una parola, usa un dizionario.

Ma attenzione: oggi “cercare su Google” non è sempre l'opzione più basso consumo. Dal 2025, Google Search ha integrato risposte AI di default.

Google è più green per “ricerche dirette o di navigazione” (meteo, orari, siti specifici), mentre l'AI può esserlo per “sintesi di argomenti complessi” che richiederebbero ore di navigazione manuale (tra pubblicità, video, ecc.).

L'AI è uno strumento potente, ma **non deve diventare il martello con cui vediamo ogni problema come un chiodo.**

Responsabilità aziendale

Se per gli utenti individuali parliamo di piccoli gesti quotidiani, per le aziende il discorso cambia scala, e con essa la responsabilità. Un'azienda di consulenza con 100 dipendenti, che usa indiscriminatamente GPT-4o per tutte le query, può arrivare a produrre 243 kg di CO₂

all'anno. Implementando una strategia intelligente invece – come ad esempio il routing automatico verso modelli leggeri per i task semplici, l'ottimizzazione dei prompt e il fine-tuning per casi d'uso ripetitivi – gli stessi risultati si ottengono inquinando solo 73 kg di CO₂. Un risparmio del 70% su licenze ed emissioni.

Se investitori e clienti guardano sempre più alle metriche ESG, avere una politica chiara sull'uso sostenibile dell'AI è anche strategico. Ma oltre che un **asset reputazionale**, serve una cultura d'impresa che consideri l'impatto ambientale come una responsabilità intrinseca.

Una scelta collettiva per un futuro efficiente e sostenibile

Google riporta che le emissioni per query di Gemini sono **diminuite di 44 volte** in un anno. Microsoft e AWS hanno impegni per il 100% di energia rinnovabile. Ma **la trasparenza resta scarsissima**. Dati auto-dichiarati o addirittura nessuna comunicazione (OpenAI e Anthropic non divulgano praticamente nulla sui consumi energetici dei loro modelli). E senza dati pubblici non c'è accountability. Non possiamo fidarci solo della buona volontà delle aziende. Attenzione poi al greenwashing!

L'Unione Europea sta provando a muoversi in questa direzione con la **Energy Efficiency Directive** del 2023 e il prossimo **Data Centre Energy Efficiency Package** del 2026.

Tuttavia, è fondamentale che la pressione arrivi anche dal basso: da cittadini informati che chiedono conto, da professionisti che inseriscono criteri di sostenibilità nei propri flussi, da investitori che premiano l'efficienza.

La politica e le risposte strutturali

L'intelligenza artificiale sta sbloccando scoperte scientifiche, migliorando diagnosi mediche, democratizzando competenze creative. Ma come ogni rivoluzione tecnologica, porta con sé dei problemi.

D'altra parte, sempre di più l'AI potrà ridurre i consumi. Applicata all'ottimizzazione delle reti elettriche, alla ricerca su batterie più efficienti o alla logistica industriale, potrebbe **risparmiare emissioni superiori a quelle che genera.** Ma questo scenario virtuoso non è automatico: richiede scelte politiche precise per orientare lo sviluppo AI verso questi utilizzi.

Ciascuno di noi può fare scelte più sostenibili: scegliere il modello giusto, scrivere prompt efficienti, chiedersi se l'AI serve davvero. Sono gesti semplici che, moltiplicati per milioni di utenti, fanno differenza. **Ma non possiamo pensare che la responsabilità ricada solo sulle spalle individuali.** Il problema è anche strutturale, e richiede risposte strutturali. Serve una regolamentazione seria che chieda ai provider trasparenza sui consumi, vincoli l'espansione dei data center a piani energetici verificati che rispettino le capacità dei territori e le comunità locali, e subordini lo sviluppo all'integrazione con fonti rinnovabili.

Non è un freno all'innovazione, anzi, ma la condizione per un'innovazione sostenibile: altrimenti rischiamo di trovarci con un'infrastruttura AI potentissima costruita su fondamenta fragili, reti elettriche al collasso, risorse idriche esaurite, emissioni fuori controllo. **Se l'infrastruttura non regge, non regge nemmeno l'innovazione che le AI promettono.**

Non dobbiamo scegliere tra azione individuale e advocacy politica, ma portarle avanti insieme. Le nostre scelte quotidiane dimostrano che un altro modo è possibile e creano la domanda che spinge il mercato. La pressione politica costruisce le regole del gioco che rendono la sostenibilità uno standard.

Ogni volta che usiamo l'AI, stiamo letteralmente accendendo una lampadina in un data center. Decidere che tipo di lampadina installare, e con quanta intensità accenderla, è una **scelta** che oggi è nelle nostre mani: **singularmente quando clicchiamo, collettivamente quando votiamo, protestiamo, scegliamo da chi comprare.**

Governiamo questa transizione ora, per assicurarci che l'intelligenza artificiale sia un alleato della terra e non un suo costo insostenibile.

– di Paola Furlan

Per approfondire

- [Energy and AI – IEA](#)
- [AI and the energy sector – Parlamento Europeo](#)