LUISS

Corso di laurea in Strategic Management

Cattedra Processi di Digitalizzazione nella funzione AFC

L'impatto ambientale della Generative AI: fattori che influenzano consumo energetico, emissioni di CO_2 e uso idrico nella fase di inferenza

Prof. Giuseppe Mazzotta

RELATORE

Prof. Francesco Legrottaglie

CORRELATORE

Martina Capoccioni

CANDIDATO

Matricola: 780321

Anno Accademico 2024/2025

INDICE

Introduzione

L'intelligenza artificiale possiede un corpo, costituito da infrastrutture fisiche, come data center, reti elettriche, sistemi di raffreddamento, indispensabili per il corretto funzionamento. Questo corpo diventa tangibile quando incontra i limiti della disponibilità energetica e delle risorse ambientali, rendendo evidente che l'AI non è un'entità immateriale ma un processo profondamente radicato nel mondo fisico. Basti pensare a quello che accade in Irlanda, sede di alcuni importanti data server di colossi come Microsoft, Google, Meta e Amazon, dove il consumo di energia elettrica di questi impianti nel 2023 ha sorpassato per la prima volta quello di tutte le abitazioni del Paese messe insieme. Di fronte a questa pressione, EirGrid, l'operatore della rete elettrica irlandese, ha deciso di sospendere fino al 2028 i nuovi allacciamenti ad alto consumo nell'area di Dublino, per evitare sovraccarichi e possibili blackout. Improvvisamente ciò che consideriamo immateriale (reti neurali, infrastrutture di calcolo, flussi di dati) ha rivelato il suo peso reale.

Da questa consapevolezza prende avvio il presente lavoro, che indaga l'impatto ambientale dei modelli di Generative AI nella fase di inferenza, con l'obiettivo di misurare i consumi energetici e idrici e le emissioni di CO_2 associate a differenti configurazioni tecnologiche. La diffusione su scala globale dei modelli di Generative Artificial Intelligence (Generative AI) sta accelerando in modo esponenziale l'uso di infrastrutture di calcolo ad alta intensità energetica.: in pochi anni sono passati da applicazioni sperimentali a servizi integrati in motori di ricerca, piattaforme di produttività e strumenti creativi utilizzati da centinaia di milioni di persone.

Se il training rappresenta un momento sporadico ma fortemente energivoro, l'inferenza è invece un processo continuo, ripetuto miliardi di volte al giorno. La potenza computazionale necessaria per generare risposte in tempo reale trasforma un tema apparentemente immateriale in una questione concreta, misurabile in kWh, litri d'acqua e tonnellate di CO_2.

Questo elaborato si propone di analizzare in profondità tali impatti, offrendo un contributo originale su più livelli. Attraverso la raccolta di dati empirici tramite API pubbliche, che ha perm) di estrarre variabili micro su un insieme eterogeneo di modelli di Generative AI e di condurre un'analisi comparativa basata su osservazioni reali, vengono esaminati i fattori architetturali, operativi e infrastrutturali che influenzano la latenza e, di conseguenza, il fabbisogno energetico, le emissioni e il consumo idrico. In questa analisi la latenza (il tempo impiegato dal modello per generare una risposta) è stata assunta come indicatore del consumo energetico, così da mettere in luce il legame tra scelte progettuali e conseguenze ambientali.

L’analisi si estende inoltre alla dimensione spaziale e temporale, evidenziando come la localizzazione dei data center e la variabilità oraria della carbon intensity possano amplificare o attenuare l’impatto ambientale delle inferenze. Infine, la ricerca integra un approccio multidimensionale che, oltre a collegare consumi energetici ed emissioni di CO_2, considera anche l’utilizzo idrico (WUE) legato al raffreddamento dei data center, un elemento spesso trascurato in letteratura ma determinante per una valutazione complessiva della sostenibilità.

L’obiettivo finale è offrire strumenti conoscitivi e quantitativi per valutare e governare l’impatto ambientale della Generative AI, mostrando come scelte progettuali e operative, dalla struttura del modello alla localizzazione del data center, possano ridurre significativamente le emissioni senza sacrificare le capacità dei modelli. In tal modo, l’elaborato contribuisce a rendere più trasparente e sostenibile una delle tecnologie più decisive del nostro tempo, offrendo strumenti conoscitivi e analitici per compiere scelte più consapevoli e sostenibili nella progettazione, nell’adozione e nella gestione della Generative AI, orientando l’innovazione verso un minore impatto ambientale.

Capitolo 1 – Contesto tecnologico, energetico e normativo dell'Intelligenza Artificiale Generativa

1.1 Introduzione

"Artificial Intelligence is likely to be the best or worst thing ever to happen to humanity."
Stephen Hawking, *BBC Reith Lecture* (2016)

Questa affermazione sintetizza la portata epocale della sfida posta dall'intelligenza artificiale (AI): una tecnologia capace di generare benefici senza precedenti, ma anche di amplificare rischi ambientali, sociali e di governance. In meno di un decennio, i modelli di generative AI, in grado di creare testi, immagini, codice e contenuti multimediali, si sono diffusi su scala globale a un ritmo senza precedenti. Questa crescita, tuttavia, comporta un forte impatto in termini di **c**onsumi energetici e risorse ambientali, che rappresenta un nodo critico per la sostenibilità del settore digitale e per il raggiungimento degli obiettivi climatici internazionali.

Il presente capitolo fornisce il quadro descrittivo e analitico necessario a comprendere l'oggetto di studio della tesi: l'impatto ambientale dell'AI generativa.
L'analisi adotta una prospettiva interdisciplinare, combinando fonti tecniche, economiche e giuridiche, e si articola in tre parti principali.

1. Fondamenti tecnologici: vengono introdotte le definizioni di intelligenza artificiale e di generative AI, ricostruita la loro evoluzione storica e analizzati gli sviluppi più recenti, come gli AI agents e i modelli multimodali. Questa sezione illustra anche la nascita degli Small Language Models (SML), concepiti per ridurre i consumi energetici e consentire applicazioni più leggere.
2. Infrastrutture digitali e determinanti tecniche dei consumi: viene analizzato il ruolo dei data center come motore fisico dell'AI, dalle origini fino agli hyperscale cloud attuali. Si approfondiscono le scelte di localizzazione geografica, il mix energetico, le tecniche di raffreddamento e il confronto cloud vs on-premise, includendo le strategie di approvvigionamento energetico dei principali provider (Google, Microsoft, AWS) e i loro programmi di decarbonizzazione e 24/7 Carbon-Free Energy. All'interno di questa cornice vengono esaminati i fattori tecnici che determinano il consumo energetico dei modelli AI: architetture (Dense, Mixture-of-Experts, SML), hardware (GPU e TPU), caratteristiche funzionali (Chain-of-Thought, multimodalità), tecniche operative (batching, caching, prompt engineering).
3. Contesto energetico e normativo: la terza parte analizza i dati internazionali sulla domanda elettrica dei data center e le regole di contabilizzazione delle emissioni (Scope 1, 2 e 3) secondo GHG Protocol e ISO 14044. Segue un'analisi delle principali normative europee e internazionali

– AI Act, CSRD, ESRS E1, Energy Efficiency Directive, Regolamento europeo sui data center – evidenziandone punti di forza e limiti rispetto alla piena misurazione dell'impatto ambientale.

Questa architettura permette di collegare in modo continuo la dimensione tecnologica con quella energetica e regolatoria, mostrando come la sostenibilità dell'AI non dipenda solo dall'efficienza dei singoli modelli, ma anche dalle infrastrutture, dalle scelte di approvvigionamento energetico e dal quadro normativo.

1.2 Evoluzione dell'Intelligenza Artificiale

L'Intelligenza Artificiale (AI), nella letteratura scientifica, viene definita come l'insieme di sistemi informatici progettati per eseguire funzioni cognitive normalmente associate all'intelligenza umana, quali ragionamento, apprendimento, percezione e interazione[1]. Le principali istituzioni convergono su questa prospettiva: per l'OCSE (2019) l'AI è un sistema basato su macchine capace di influenzare l'ambiente attraverso l'interpretazione di dati; il NIST (2023) la descrive come la capacità di sistemi tecnologici di percepire, apprendere, ragionare e agire per conseguire obiettivi specifici; l'AI Act dell'Unione Europea (2024) la definisce come un insieme di software sviluppati con tecniche specifiche che possono generare output quali contenuti, previsioni, raccomandazioni o decisioni che influenzano l'ambiente fisico o virtuale.

Questa cornice comune, pur con sfumature diverse, evidenzia un elemento cardine: l'AI non è un'unica tecnologia, ma un percorso evolutivo che, nel tempo, ha integrato approcci, modelli e infrastrutture sempre più complessi

L'Intelligenza Artificiale (AI) non rappresenta un singolo paradigma tecnologico, ma un insieme stratificato di approcci che si sono sviluppati nel tempo, ognuno con capacità crescenti di apprendimento, autonomia e complessità computazionale. La letteratura tecnica e le linee guida internazionali consentono di distinguere quattro macrofamiglie principali, che si innestano l'una nell'altra e continuano a coesistere nell'attuale ecosistema digitale.

1.2.1 Rule-Based Algorithms

Sono i sistemi deterministici che hanno dominato le prime fasi dello sviluppo dell'AI, dagli anni '50 fino ai primi sistemi esperti degli anni '80. Operano secondo logiche if–then: dati uno stato iniziale e regole prefissate, producono un output senza capacità di apprendimento. Esempi tipici sono i filtri antispam, i motori di regole per la diagnostica industriale o i sistemi di controllo degli accessi[2].
Pur offrendo trasparenza e predicibilità, questi modelli richiedono che ogni scenario sia esplicitamente

[1] Russell, S. and Norvig, P. (2021) *Artificial Intelligence: A Modern Approach*. 4th edn. Boston: Pearson.
[2] Russell, S. & Norvig, P. (2021) *Artificial Intelligence: A Modern Approach*. 4th edn. Harlow: Pearson.

programmato, con limiti evidenti di adattabilità. In particolare, limiti hardware e scarsità di dati provocarono due lunghi periodi di rallentamento, noti come AI winter, che ridussero investimenti e interesse scientifico.

1.2.2 Machine Learning

Negli anni ’90 la combinazione di capacità di calcolo crescente e disponibilità di dataset digitali portò alla diffusione del Machine Learning (ML). Introdotto su larga scala a partire dagli anni 2000 segna la svolta verso la capacità di apprendere dai dati. Nei modelli ML l’algoritmo costruisce funzioni predittive a partire da esempi, con metodi supervisionati, non supervisionati o rinforzati (Goodfellow, Bengio & Courville, 2016[3]). Questa famiglia ha abilitato applicazioni come il riconoscimento facciale e recommendation systems. Tuttavia, la necessità di definire in anticipo le caratteristiche salienti dei dati e di alimentare il sistema con dataset ampi resta un vincolo, e la complessità computazionale cresce rapidamente con la dimensione dei dati.

1.2.3 Deep Learning

Il decennio successivo vide un salto qualitativo con il Deep Learning, reso possibile da GPU potenti e da enormi volumi di dati web. Il Deep Learning costituisce un sottoinsieme del ML basato su reti neurali profonde (deep neural networks) e rappresenta una delle maggiori rivoluzioni tecnologiche degli ultimi decenni. Attraverso più strati di neuroni artificiali, i modelli sono in grado di individuare pattern altamente complessi anche in dati non strutturati (immagini, suoni, testi) e di elaborare regole interne senza supervisione esplicita (LeCun, Bengio & Hinton, 2015[4]).
L’accuratezza e la capacità di generalizzazione ne hanno favorito l’adozione in campi come la guida autonoma e l’elaborazione del linguaggio naturale, ma al prezzo di un impatto energetico significativo, data l’enorme potenza di calcolo richiesta per l’addestramento e l’inferenza.

1.2.4 Generative AI e Large Language Models

A partire dal 2018 l’Intelligenza Artificiale ha conosciuto un’accelerazione senza precedenti con l’emergere dei Foundation Models, definiti da Bommasani et al. (2021)[5] come modelli addestrati su vasti insiemi di dati e riutilizzabili per una molteplicità di compiti downstream. La svolta è resa possibile dall’architettura

[3] Goodfellow, I., Bengio, Y. & Courville, A. (2016) Deep Learning. Cambridge, MA: MIT Press.

[4] LeCun, Y., Bengio, Y. & Hinton, G. (2015) ‘Deep learning’, *Nature*, 521(7553), pp. 436–444.

[5] Bommasani, R. et al. (2021) ‘On the opportunities and risks of foundation models’, *arXiv preprint* arXiv:2108.07258.

transformer (Vaswani et al., 2017), che ha introdotto il meccanismo di self-attention, consentendo di elaborare sequenze lunghe di dati e di catturare dipendenze contestuali in modo più efficiente rispetto alle reti neurali precedenti.

Questi modelli non si limitano a riconoscere schemi, ma costruiscono rappresentazioni generali del linguaggio e di altre modalità, riutilizzabili per compiti differenti, dalla traduzione automatica alla generazione di codice. L'evoluzione ha portato alla nascita della Generative AI, ossia sistemi capaci di produrre testi, immagini, audio e video originali e coerenti, con livelli di creatività e fluidità spesso paragonati a quelli umani (Brown et al., 2020; Ramesh et al., 2021).

All'interno della Generative AI si colloca la famiglia oggi più rappresentativa: i **Large Language Models (LLM).** Si tratta di reti neurali profonde basate su transformer e caratterizzate da una scala di parametri senza precedenti, che dai miliardi di GPT-2 sono passati a centinaia di miliardi, con stime non ufficiali che per i modelli più recenti superano il trilione. L'addestramento segue in genere tre fasi: un pre-training auto-supervisionato su corpora testuali e multilingue di dimensioni petabyte; un fine-tuning su compiti specifici; e, sempre più spesso, un'ottimizzazione attraverso reinforcement learning from human feedback (RLHF) per migliorare l'allineamento alle aspettative umane.

La crescita di scala è documentata da evidenze quantitative note come scaling laws: la capacità predittiva di un LLM cresce in modo log-lineare con l'aumento simultaneo di parametri, dati e potenza computazionale (Kaplan et al., 2020). Questa dinamica ha reso possibile, in soli cinque anni, il passaggio da modelli con qualche miliardo di parametri a sistemi che ne contano centinaia di miliardi e che, secondo proiezioni, raggiungeranno rapidamente ordini di grandezza superiori.

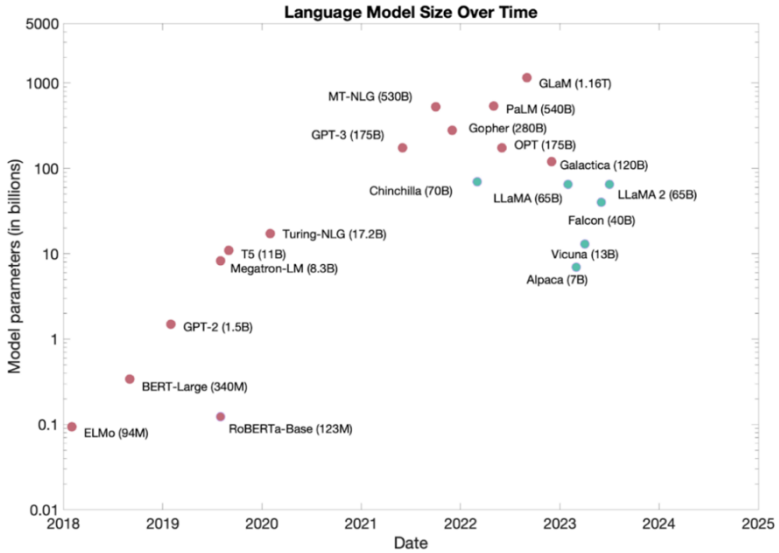


Figure 1. A selection of recently released models and their size in terms of number of parameters. Note that the y-axis is logarithmic and that model size has been growing exponentially. Of late, some model size has been substituted for longer training periods, as suggested by the markers in green.

Parallelamente, i modelli stanno acquisendo capacità multimodali, cioè la possibilità di elaborare e generare simultaneamente testo, immagini, audio e video. Modelli come Flamingo (2022), GPT-4 (OpenAI, 2023) e Gemini 1.5 (Google, 2024) sono esempi di architetture in grado di collegare dati visivi, sonori e linguistici in un unico spazio rappresentazionale. Questa integrazione consente una comprensione semantica più vicina a quella umana, capace di cogliere relazioni tra parole, immagini e suoni, e la creazione di output complessi, come descrizioni testuali di immagini o generazione di contenuti video a partire da istruzioni verbali.

Dal punto di vista infrastrutturale, la multimodalità richiede reti neurali più ampie e un uso coordinato di diversi tipi di processori (GPU, TPU, acceleratori specializzati), con conseguenti aumenti di domanda energetica e di fabbisogno di memoria. Questa potenza generativa e la capacità di adattarsi a contesti diversi spiegano perché i Foundation Models siano diventati la spina dorsale della Generative AI contemporanea. Tuttavia, le stesse caratteristiche che ne determinano il successo sollevano sfide ambientali significative. Il pre-training di un singolo modello di frontiera può richiedere centinaia di gigawattora di elettricità e produrre decine di migliaia di tonnellate di CO_2 equivalente (Patterson et al., 2021; IEA, 2024). E la fase di inferenza quotidiana, cioè le risposte generate in tempo reale per miliardi di query utente, tende nel tempo a superare i consumi del training, ampliando l'impatto energetico complessivo.

Questi aspetti, insieme ai fattori tecnici che li determinano (dimensione del modello, architettura Dense vs Mixture-of-Experts, complessità dei prompt, strategie di batching e caching), saranno analizzati nel dettaglio nella Sezione 1.5, dedicata alle determinanti tecniche del consumo energetico.

1.2.5 Verso AI Agents e multimodalità

L'evoluzione più recente vede l'introduzione di AI agent, cioè di sistemi che, partendo da un obiettivo generale fornito dall'utente, sono in grado di pianificare autonomamente le azioni necessarie, eseguirle in ambienti digitali o fisici e adattare il proprio comportamento in base ai risultati ottenuti. Rispetto ai Large Language Models tradizionali, che rispondono a singole query, gli AI agents gestiscono cicli di interazione prolungati, memorizzando lo stato del contesto; chiamano strumenti esterni (ad esempio API, database, ambienti di calcolo) per realizzare compiti complessi e prendono decisioni sequenziali, in modo simile a un processo di ragionamento umano (Wang et al., 2023).

Questa traiettoria storica implica un salto progressivo nei requisiti di calcolo e nei consumi elettrici, con effetti proporzionali sull'impatto climatico. Dai sistemi rule-based, a ridotto fabbisogno energetico, l'AI è passata al machine learning e al deep learning, che richiedono addestramenti molto più dispendiosi, fino agli odierni Large Language Models e alle architetture multimodali, che costituiscono il banco di prova più critico per i data center e per le politiche di decarbonizzazione dell'economia digitale.

1.3 Data center: infrastruttura, consumi e strategie energetiche

L'intelligenza artificiale è, nella sua essenza, una tecnologia digitale. Ma il suo funzionamento non è immateriale: ogni modello, ogni query e ogni byte elaborato si appoggiano a un'infrastruttura fisica globale energivora costituita da data center. Questi impianti, composti da complessi di edifici, server, apparati di rete, sistemi di raffreddamento e dispositivi di sicurezza, sono progettati per fornire capacità di calcolo continuativa e ad alte prestazioni. Nell'ultimo decennio il loro ruolo è diventato strategico per l'economia digitale, al punto che senza data center non esisterebbero servizi cloud, piattaforme di streaming, né l'odierna intelligenza artificiale generativa[6].

Dal punto di vista climatico, i data center rappresentano una componente crescente della domanda elettrica mondiale. Le stime più recenti dell'International Energy Agency (IEA, 2024) indicano un consumo di circa 460 TWh nel 2022 — pari a circa il 4% della domanda globale — con una possibile crescita fino a 1.000 TWh entro il 2030, spinta in larga misura dall'espansione delle applicazioni di intelligenza artificiale. L'intensità delle emissioni dipende non solo dalla quantità di energia assorbita, ma dal mix di generazione elettrica e dalla carbon intensity oraria delle reti di approvvigionamento (ElectricityMaps, 2024).

Questa sezione analizza i principali fattori che collegano i data center all'impatto ambientale dell'AI: la loro evoluzione storica e le caratteristiche architetturali, il ruolo specifico nell'addestramento e nell'inferenza dei modelli, l'influenza della localizzazione geografica e del mix energetico, le tecniche di raffreddamento, il confronto cloud vs on-premise e le strategie di approvvigionamento e decarbonizzazione adottate dai principali provider. L'obiettivo è fornire un quadro tecnico e documentato del legame fra infrastrutture di calcolo e sostenibilità climatica, che costituirà la base per le analisi quantitative e normative dei capitoli successivi.

1.3.1 Origine e sviluppo: dai server room ai cloud hyperscale

I data center sono l'infrastruttura fisica su cui poggiano i servizi digitali e, sempre più, l'intelligenza artificiale. La loro evoluzione, dagli esordi legati ai mainframe fino agli attuali complessi hyperscale, riflette oltre sessant'anni di progressi nell'informatica, nelle telecomunicazioni e nell'ingegneria energetica.

1.3.1.1 Evoluzione dei data center

[6] Uptime Institute (2023) *Global Data Center Survey*.

Negli anni ’60 e ’70 i data center coincidevano con sale server dedicate che ospitavano uno o pochi mainframe. Queste strutture, connessi tramite reti locali, richiedevano climatizzazione costante ma avevano dimensioni e consumi relativamente limitati. L’espansione dell’architettura client–server negli anni ’80–’90 portò alla diffusione di centri di calcolo aziendali (on-premise), ognuno con decine o centinaia di server e con Power Usage Effectiveness (PUE) spesso superiore a 2,0, segno di scarsa efficienza energetica.

L’esplosione del traffico Internet e la necessità di risorse elastiche portarono, nei primi anni 2000, al modello cloud computing. Secondo la definizione del National Institute of Standards and Technology (NIST), il cloud è “un modello che consente l’accesso su richiesta, comodo e ubiquo, a un insieme condiviso di risorse configurabili di elaborazione che possono essere rapidamente fornite e rilasciate con minimo sforzo di gestione” (Mell & Grance, 2011).

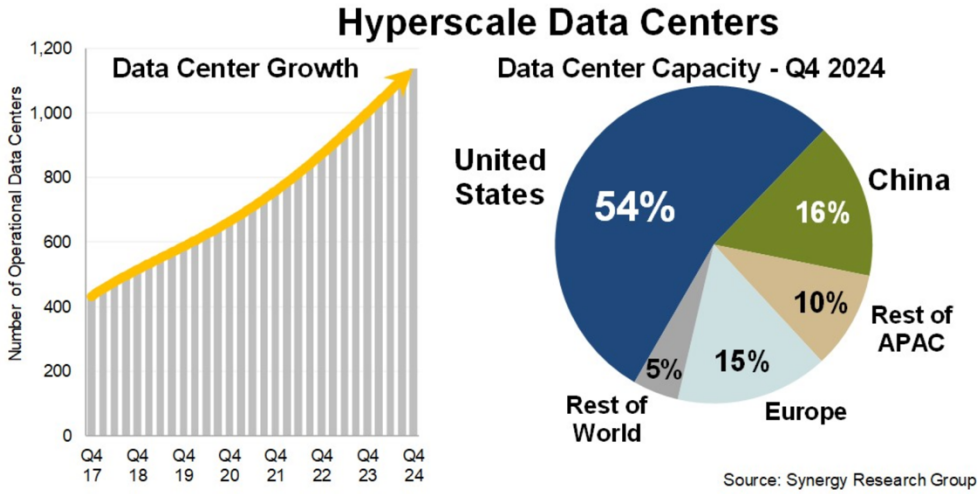
Nel 2006 Amazon inaugurò Amazon Web Services (AWS), introducendo il concetto di region: insiemi di data center geograficamente vicini, replicati e interconnessi, progettati per erogare potenza di calcolo e archiviazione on demand. Google Cloud e Microsoft Azure adottarono rapidamente lo stesso approccio. Questo passaggio ha segnato un punto di svolta sia nella scalabilità informatica sia nell’intensità energetica, creando le basi per l’AI contemporanea.[7].

A partire dalla metà degli anni 2010, i grandi provider hanno sviluppato il modello hyperscale, caratterizzato da:

- superfici >100.000 m²,
- potenza elettrica installata di 50–150 MW per campus,
- centinaia di migliaia di server organizzati in migliaia di rack.

Secondo Synergy Research Group (2024), nel 2023 esistevano circa 850 hyperscale data center nel mondo, più del triplo rispetto al 2015, con una crescita media annua del 13%. Le aree a maggiore densità si trovano negli Stati Uniti (Northern Virginia, Dallas), in Europa (Irlanda, Paesi Bassi) e in Asia (Singapore, Tokyo).

[7] Amazon Web Services (2023) *AWS and Sustainability*.



L'evoluzione verso il cloud e l'hyperscale ha reso possibili le attuali applicazioni di intelligenza artificiale generativa, che richiedono cluster di decine di migliaia di GPU e continuità di alimentazione elettrica su larga scala. Questo salto di scala implica ordini di grandezza maggiori nei consumi energetici e pone i data center al centro delle politiche di decarbonizzazione e gestione della domanda elettrica, aspetti che saranno analizzati nelle sottosezioni successive.

1.3.2 Ruolo per l'AI: GPU e TPU specializzate, potenza e scalabilità

La rapida ascesa dell'intelligenza artificiale generativa è strettamente legata all'evoluzione dei data center hyperscale, che offrono l'infrastruttura di calcolo necessaria per l'addestramento e l'inferenza dei Large Language Models (LLM) e dei modelli multimodali. Questa sezione analizza le componenti hardware critiche, i trend di potenza e le implicazioni energetiche.

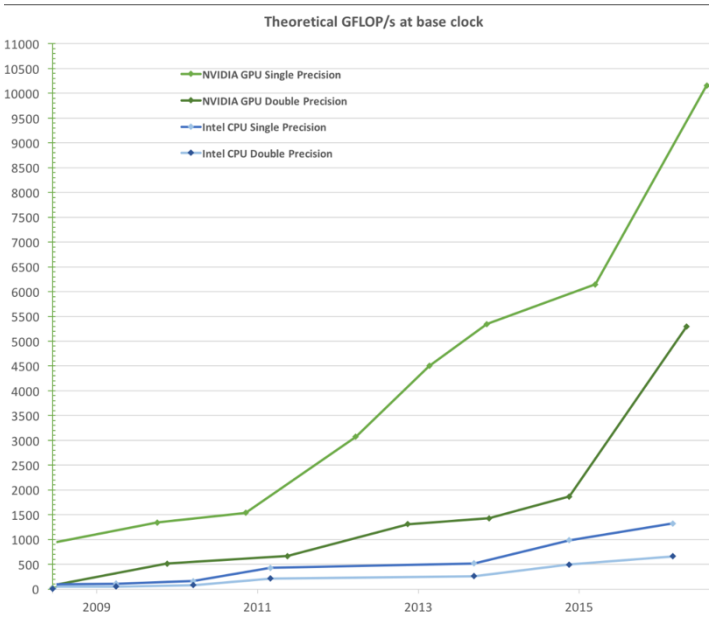
1.3.2.1 Acceleratori specializzati: GPU e TPU

I modelli di AI contemporanei richiedono calcolo altamente parallelo, che le tradizionali CPU non sono in grado di fornire in modo efficiente. Per questo motivo i data center destinati all'AI si basano su

- Graphics Processing Units (GPU), in particolare le serie Nvidia A100 e H100, progettate per operazioni di matrix multiplication e tensor computation;
- Tensor Processing Units (TPU), sviluppate da Google, ottimizzate per carichi di deep learning e in particolare per l'architettura transformer.

Le GPU di fascia più avanzata, come le Nvidia H100, erogano fino a 60 teraflop in precisione FP16 e sono progettate per eseguire miliardi di operazioni di moltiplicazione di matrici in parallelo, mentre le TPU

sviluppate da Google offrono un'ottimizzazione mirata per l'architettura transformer, cuore degli LLM (Nvidia, 2023; Google, 2023).



Il salto tecnologico non riguarda solo le singole unità, ma la loro integrazione in cluster di grandi dimensioni. Decine di migliaia di GPU o TPU, connesse tramite reti ad altissima velocità (InfiniBand o Ethernet 400 Gb/s), consentono di raggiungere potenze complessive dell'ordine degli exaflop, equivalenti a quelle dei più grandi supercomputer scientifici. Un singolo campus di questo tipo può assorbire centinaia di megawatt di elettricità, un fabbisogno istantaneo paragonabile a quello di un'intera città di medie dimensioni (Patterson et al., 2021; IEA, 2024).

Questa potenza è necessaria per l'addestramento dei modelli di frontiera, un processo che, secondo stime indipendenti, può richiedere centinaia di gigawattora per singolo training. Ma anche la successiva fase di inferenza, ossia l'esecuzione quotidiana di miliardi di query, comporta consumi cumulativi che nel tempo possono eguagliare o superare quelli del training iniziale. Le innovazioni in materia di densità computazionale e di raffreddamento avanzato (ad esempio il liquid cooling diretto) migliorano il rapporto tra calcolo ed energia, ma non compensano la crescita esponenziale della domanda, che continua a spingere verso un aumento assoluto del fabbisogno elettrico globale (IEA, 2024).

La convergenza tra acceleratori dedicati e architetture hyperscale colloca così i data center al centro delle sfide di sostenibilità. La scelta della localizzazione geografica, del mix di generazione elettrica e degli strumenti contrattuali di approvvigionamento rinnovabile (come i Power Purchase Agreements) diventa determinante per contenere le emissioni di Scope 2 e, indirettamente, di Scope 3. In questa prospettiva, le strategie dei grandi provider per alimentare tali infrastrutture con energia a basse emissioni, analizzate nelle sezioni successive, rappresentano un fattore chiave per conciliare l'espansione dell'AI con gli obiettivi di decarbonizzazione.

1.3.3 Localizzazione geografica, mix elettrico e carbon intensity

La localizzazione geografica dei data center incide in modo determinante sull'impronta climatica delle infrastrutture digitali. A parità di architettura e di efficienza hardware, le emissioni di CO_2 per chilowattora dipendono dal mix di generazione elettrica e dalla carbon intensity della rete locale.

Ogni Paese ha infatti un mix di generazione diverso, determinato da fattori storici, geografici, economici e politici, che incide direttamente sulle emissioni di CO_2 associate a ogni chilowattora consumato. Le differenze derivano innanzitutto dalla disponibilità di risorse energetiche primarie. Paesi ricchi di idroelettrico o geotermico, come Norvegia e Islanda, hanno potuto sviluppare una generazione elettrica quasi interamente rinnovabile, con intensità carbonica molto bassa (inferiore a 50 gCO_2/kWh secondo IEA, 2024). Altri Paesi, come Francia, hanno investito storicamente sul nucleare, ottenendo anch'essi una bassa intensità media. Al contrario, economie basate su carbone o gas naturale — ad esempio India, Sudafrica o alcune province cinesi — mantengono valori elevati, spesso superiori a 600–700 gCO_2/kWh.

Un secondo fattore è rappresentato dalle politiche energetiche e climatiche. Obiettivi nazionali di decarbonizzazione, incentivi alle rinnovabili, sistemi di carbon pricing e normative sulle emissioni influenzano la composizione del mix. L'Unione europea, con il pacchetto Fit for 55 e il mercato EU ETS, ha creato condizioni favorevoli per la penetrazione di solare ed eolico; gli Stati Uniti, invece, presentano forti differenze regionali, poiché le scelte sui piani energetici sono in gran parte affidate ai singoli Stati.

La configurazione della rete e la sua flessibilità giocano un ruolo aggiuntivo. Le reti elettriche con elevata interconnessione transfrontaliera (ad esempio nell'Europa centro-settentrionale) permettono di compensare la variabilità delle rinnovabili importando energia da Paesi a minore intensità carbonica. Dove le interconnessioni sono scarse, come in molte regioni asiatiche, la quota di carbone resta invece predominante per garantire continuità di servizio.

La crescente complessità energetica e ambientale dei data center ha portato, a partire dagli anni Duemila, alla necessità di metriche standardizzate per confrontare le prestazioni e guidare l'efficienza.
Nel 2007, il consorzio industriale The Green Grid introdusse il Power Usage Effectiveness (PUE), primo indicatore globale per misurare l'efficienza energetica complessiva. Il PUE è definito come il rapporto tra l'energia totale assorbita dal data center e quella utilizzata dall'IT; un PUE pari a 1,0 rappresenta la massima efficienza teorica.
Nel 2011, lo stesso consorzio propose il Water Usage Effectiveness (WUE), in risposta alle crescenti preoccupazioni per l'uso idrico dei sistemi di raffreddamento. Esprime i litri d'acqua consumati per chilowattora di energia IT erogata.

Nel 2010, fu introdotto il Carbon Usage Effectiveness (CUE), che collega i consumi elettrici del data center alle emissioni effettive di CO_2 sulla base del mix energetico.
Questi indicatori, oggi richiamati in norme di rendicontazione come la Corporate Sustainability Reporting Directive (CSRD), consentono di valutare in modo comparabile le prestazioni di impianti collocati in contesti geografici e climatici diversi.

Per l'intelligenza artificiale generativa, che richiede centinaia di gigawattora per il training e un flusso costante di energia per l'inferenza quotidiana, la localizzazione non è quindi una decisione neutra.

1.3.3.1 Standard di misurazione e rendicontazione delle emissioni (GHG Protocol e ISO 14044).

Sul piano della rendicontazione climatica, il riferimento principale è il GHG Protocol – Scope 2 Guidance (WRI/WBCSD, 2015), che definisce come calcolare le emissioni derivanti dall'energia elettrica acquistata. Lo standard distingue due metodologie complementari:

- Approccio location-based: calcola le emissioni moltiplicando il consumo di elettricità per il fattore di emissione medio della rete locale, espresso in gCO_2/kWh. Riflette le condizioni fisiche del sistema elettrico in cui il data center opera e fornisce una misura diretta dell'impatto territoriale.
- Approccio market-based: considera invece l'elettricità effettivamente acquistata sul mercato, tenendo conto di contratti specifici come i Power Purchase Agreements (PPA) o i certificati di origine rinnovabile (es. Garanzie d'Origine in UE, Renewable Energy Certificates negli USA). In questo modo un'azienda che finanzia nuova capacità rinnovabile o compra energia verde dedicata può dichiarare un'impronta inferiore rispetto alla media della rete.

Entrambi i metodi devono essere applicati e riportati, ma le implicazioni sono diverse. Il dato location-based è conservativo e fisico, descrive ciò che avviene realmente nel territorio e consente confronti fra Paesi. Il market-based premia invece le scelte di approvvigionamento sostenibile, segnalando l'impegno verso l'energia rinnovabile.

Nel contesto europeo, la Corporate Sustainability Reporting Directive (CSRD) — che dal 2024 amplia gli obblighi di disclosure non finanziaria — richiede che le imprese forniscano entrambi i valori, sottolineando l'importanza del location-based come indicatore dell'impatto effettivo. Per i cloud provider questa distinzione è cruciale: molte big tech comunicano prevalentemente dati market-based, ma i regolatori spingono per una doppia rendicontazione che mostri sia l'intensità della rete locale sia l'effetto delle strategie di acquisto di energia pulita.

Questa doppia prospettiva è particolarmente rilevante per i data center che ospitano modelli di intelligenza artificiale generativa, in quanto consente di valutare non solo l'impegno contrattuale verso la decarbonizzazione ma anche il contributo reale alla riduzione delle emissioni nei territori in cui l'energia viene effettivamente consumata.

D'altra parte, sul piano della valutazione del ciclo di vita, il riferimento principale è la ISO 14044:2006 (Environmental management- Life cycle assessment -Requirements and guidelines), sviluppata dall'International Organization for Standardization (ISO). Questa norma, nata come evoluzione della ISO 14040, stabilisce i principi e i requisiti per condurre un'analisi LCA (Life Cycle Assessment) completa, dalla produzione delle materie prime fino al fine vita del prodotto o dell'infrastruttura.

L'LCA si articola in quattro fasi metodologiche:

- Definizione di obiettivi e campo di applicazione, che identifica il sistema da analizzare e i confini del ciclo di vita (ad esempio, componenti hardware di un data center).
- Analisi dell'inventario, in cui si raccolgono dati quantitativi su consumi energetici, materiali e trasporti.
- Valutazione degli impatti, che traduce i flussi di materia ed energia in indicatori ambientali (ad esempio emissioni di gas serra, acidificazione, uso idrico).
- Interpretazione, nella quale i risultati vengono verificati, discussi e collegati a decisioni operative o strategiche.

La ISO 14044 richiede trasparenza nelle assunzioni, tracciabilità dei dati e criteri di qualità rigorosi. Nel contesto dei data center e dell'intelligenza artificiale generativa, la norma consente di stimare non solo le emissioni dirette e indirette di Scope 1 e Scope 2, ma anche le emissioni Scope 3, come quelle legate alla produzione di semiconduttori, alla costruzione degli edifici o allo smaltimento delle apparecchiature elettroniche.

L'adozione della ISO 14044 è oggi incoraggiata da diverse normative internazionali e regionali, tra cui la Corporate Sustainability Reporting Directive (CSRD), che riconosce la LCA come strumento chiave per una rendicontazione completa. Per i provider di cloud e per le infrastrutture che alimentano modelli di intelligenza artificiale, applicare questa norma significa poter documentare in modo scientifico l'impatto ambientale dell'intero ciclo di vita, andando oltre il solo consumo energetico e fornendo una base verificabile per obiettivi di riduzione delle emissioni.

1.3.4 Consumo energetico dei data center e tecniche di raffreddamento

I data center sono le infrastrutture fisiche che rendono possibile l'elaborazione, l'archiviazione e la distribuzione delle informazioni digitali, incluse le operazioni di addestramento e inferenza dei modelli di intelligenza artificiale generativa. Il loro impatto energetico è oggi una delle variabili decisive per la sostenibilità del settore ICT. Secondo l'International Energy Agency (IEA, 2024), l'insieme dei data center rappresenta circa il 4% della domanda elettrica mondiale e, nello scenario tendenziale, potrebbe raggiungere il 7–8% entro il 2030, con un contributo crescente proveniente dalle applicazioni di AI.

1.3.4.1 Fabbisogno energetico dei data center

Il fabbisogno energetico di un data center si compone di due macro-aree:

- Carico IT (Information Technology load)

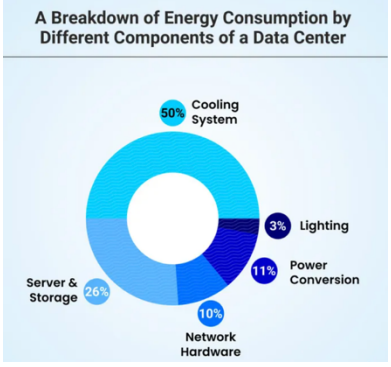
È l'energia destinata direttamente all'elaborazione e all'archiviazione dei dati. Comprende server, apparati di rete, sistemi di storage e, nei cluster per intelligenza artificiale, GPU e TPU ad alta densità. L'esecuzione di modelli di machine learning e deep learning comporta un'intensa attività di calcolo parallelo, con consumi che possono superare 30–40 kW per singolo rack e, nei progetti di frontiera per l'addestramento di LLM, arrivare a oltre 50 kW (Uptime Institute, 2023; 174 Power Global, 2025). La natura always-on di questi carichi fa sì che il consumo sia pressoché continuo.

- Infrastrutture di supporto (Facility load)

Accanto al carico IT, una quota significativa di energia è necessaria per garantire condizioni operative stabili e sicure. Le voci principali sono:

1. Sistemi di alimentazione e continuità (Power supply & UPS), che mantengono la fornitura anche in caso di interruzioni di rete.
2. Distribuzione elettrica interna, con trasformatori e sistemi di conversione che comportano inevitabili perdite.
3. Illuminazione e sicurezza.
4. Raffreddamento, cioè l'insieme di impianti dedicati alla rimozione del calore generato dalle apparecchiature.

Proprio il raffreddamento rappresenta in media tra il 30% e il 40% dell'energia totale assorbita da un data center, quota che può salire oltre il 50% nei siti con densità elevatissime o in climi caldi (IEA, 2024; Uptime Institute, 2023). La ragione risiede nella fisica elementare: ogni watt di potenza elettrica consumata dai server si trasforma in calore che deve essere dissipato per mantenere temperature operative comprese fra 18 °C e 27 °C.



1.3.4.2 Indicatori di efficienza

Per misurare l'efficienza complessiva è stato introdotto il Power Usage Effectiveness (PUE), definito come rapporto tra l'energia totale assorbita dalla struttura e l'energia effettivamente utilizzata per l'IT.
Un PUE ideale pari a 1,0 indicherebbe che ogni kWh consumato alimenta solo il carico IT.
I data center più avanzati registrano valori intorno a 1,1–1,2, mentre impianti meno ottimizzati possono superare 1,5 (The Green Grid, 2007; Uptime Institute, 2023).
Oltre al PUE, il Water Usage Effectiveness (WUE) quantifica il consumo idrico per chilowattora di energia IT, mentre il Carbon Usage Effectiveness (CUE) collega il consumo elettrico alle emissioni effettive di CO_2 in funzione del mix energetico.

1.3.4.3 Tecniche di raffreddamento e impatto ambientale

La rimozione del calore è una funzione vitale per ogni data center. qualsiasi watt assorbito dalle apparecchiature IT viene quasi interamente trasformato in calore; se non adeguatamente dissipato, può provocare malfunzionamenti, ridurre la vita utile dei componenti e persino arrestare i sistemi. Nel tempo si è sviluppata una vera e propria ingegneria della climatizzazione dei data center, con soluzioni che si sono evolute di pari passo con l'aumento delle densità di calcolo.

- **Raffreddamento ad aria tradizionale.**

È la tecnica storica, basata su condizionatori industriali e ventilazione forzata che convoglia aria fredda verso le apparecchiature. La sua robustezza e semplicità l'hanno resa lo standard per decenni, ma oggi mostra limiti evidenti. Quando la densità di potenza supera 20–25 kW per rack, occorrono enormi volumi di aria fredda e il PUE tende a crescere rapidamente, con conseguente aumento dei consumi e dei costi. Inoltre, la necessità di mantenere un ricircolo continuo comporta consumi elettrici indiretti significativi per ventilatori e compressori.

- **Containment a corridoio caldo/freddo (hot/cold aisle containment).**

Per migliorare l'efficienza dell'aria tradizionale, si è diffusa la tecnica dell'*hot/cold aisle containment*. Consiste nel separare fisicamente i corridoi che immettono aria fredda da quelli che estraggono l'aria calda, impedendo mescolamenti e consentendo di alzare di qualche grado la temperatura di mandata senza rischi. Questo accorgimento, ormai considerato una best practice, consente risparmi energetici nell'ordine del 10–20% e una maggiore uniformità termica.

- **Free cooling.**

Nei Paesi a clima temperato o freddo – ad esempio i Paesi nordici – è spesso possibile utilizzare direttamente l'aria esterna per una parte importante dell'anno. Questa soluzione, chiamata free cooling, riduce

drasticamente il ricorso a chiller e compressori meccanici, abbattendo sia il consumo elettrico sia quello idrico. In siti come la Svezia o la Finlandia, consente PUE prossimi a 1,1 senza sacrificare affidabilità.

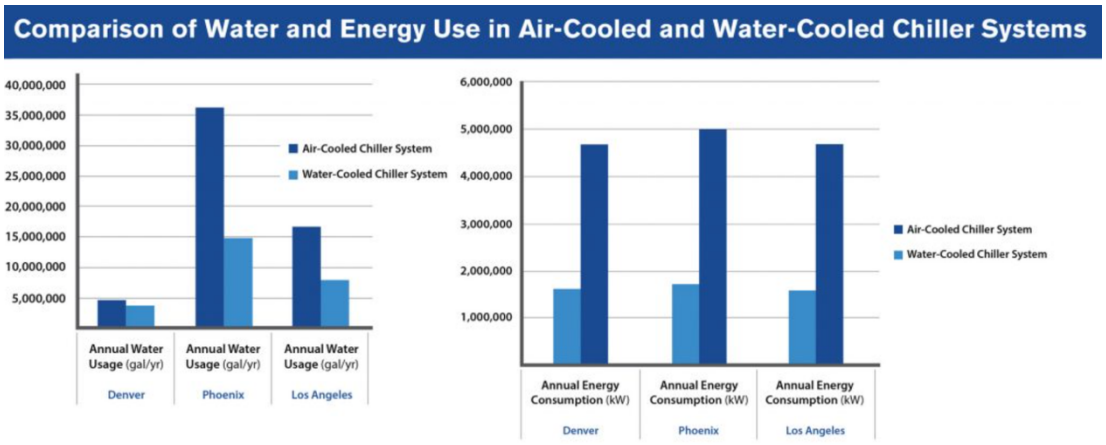
- **Raffreddamento a liquido diretto (direct-to-chip liquid cooling).**

L'esplosione della potenza di calcolo richiesta dall'intelligenza artificiale generativa ha spinto verso soluzioni più radicali. Il raffreddamento a liquido porta fluidi refrigeranti direttamente sulle superfici dei processori e delle GPU, estraendo il calore in maniera molto più efficace rispetto all'aria. È oggi una delle tecniche più promettenti per rack da 50 kW e oltre, perché mantiene basse le temperature operative anche in cluster ad altissima densità, con PUE che possono restare intorno a 1,1.

- **Immersion cooling.**

Rappresenta lo stadio più avanzato. I server sono completamente immersi in fluidi dielettrici che assorbono e trasferiscono il calore con estrema efficienza, eliminando quasi del tutto la necessità di ventilatori e condizionamento dell'aria. In test condotti su cluster AI di ultima generazione, questa tecnologia ha consentito di ridurre fino al 90% l'energia destinata alla ventilazione, oltre a prolungare la vita dei componenti grazie a una temperatura più stabile.

Ognuna di queste tecniche comporta impatti ambientali differenti. I sistemi evaporativi, ad esempio, possono richiedere fino a 4 litri d'acqua per ogni kWh di energia IT, mentre le soluzioni a circuito chiuso o l'uso combinato di free cooling e raffreddamento a liquido riducono sensibilmente il Water Usage Effectiveness (WUE). Anche la Carbon Usage Effectiveness (CUE) ne risente: un PUE più basso significa minori kWh totali da compensare, a parità di mix elettrico.



In prospettiva, l'adozione combinata di raffreddamento a liquido, free cooling e gestione intelligente dei carichi rappresenta una delle leve più efficaci per conciliare la crescita della domanda di calcolo con gli obiettivi di decarbonizzazione. Non si tratta solo di una scelta tecnologica, ma di una strategia climatica: ottimizzare il raffreddamento significa ridurre in modo diretto il consumo elettrico e, quindi, le emissioni associate, soprattutto in Paesi dove il mix energetico è ancora fortemente dipendente da fonti fossili. Comprendere a fondo la struttura dei consumi e le opzioni tecnologiche è quindi essenziale per valutare

l’impatto ambientale delle applicazioni di intelligenza artificiale generativa e per definire politiche efficaci di decarbonizzazione.

1.3.5 Cloud hyperscale vs on-premise

La crescita vertiginosa della domanda di calcolo, alimentata dall’intelligenza artificiale generativa e dai Large Language Models, ha trasformato le scelte infrastrutturali in una variabile chiave della sostenibilità digitale. Le organizzazioni che sviluppano o utilizzano modelli di AI si trovano oggi davanti a un bivio: affidarsi ai grandi provider di cloud hyperscale, come Amazon Web Services, Microsoft Azure e Google Cloud, oppure continuare a operare con data center on-premise, di proprietà e gestione diretta.

I Cloud hyperscale comprendono data center di grandissime dimensioni, progettati da operatori come Amazon Web Services (AWS), Microsoft Azure e Google Cloud, che forniscono capacità di calcolo e archiviazione “on demand” a clienti in tutto il mondo. Sono caratterizzati da una densità di server molto elevata, da economie di scala e da una gestione centralizzata dell’energia. Secondo Synergy Research Group (2024), nel 2023 erano attivi oltre 900 data center hyperscale nel mondo, con una crescita prevista di oltre il 50% entro il 2030.

I Data Center On-premise invece, sono di proprietà dell’azienda utilizzatrice, situati nei suoi stabilimenti o in spazi dedicati. Offrono un controllo diretto sulla sicurezza fisica e logica, ma richiedono ingenti investimenti iniziali e spesso presentano PUE più elevati, specie negli impianti meno recenti.

I grandi provider cloud dichiarano PUE medi tra 1,1 e 1,2, grazie a progettazione avanzata, raffreddamento ottimizzato e utilizzo di tecnologie come free cooling o immersion cooling (Uptime Institute, 2023). Mentre nei data center on-premise, soprattutto se datati, il PUE può variare tra 1,6 e 2,0, con conseguente maggiore consumo di energia a parità di carico IT.

Le ragioni sono strutturali: i provider cloud operano su scala globale, investono in tecnologie di raffreddamento all’avanguardia e localizzano i propri impianti in aree con condizioni climatiche e mix elettrico favorevoli.

Un ulteriore vantaggio del cloud hyperscale riguarda la capacità di approvvigionamento da fonti rinnovabili. Grandi operatori come Google, Microsoft e Amazon hanno sottoscritto Power Purchase Agreements (PPA) su larga scala e adottato programmi come il 24/7 Carbon-Free Energy, che mira a coprire in tempo reale il 100% della domanda con energia priva di carbonio. Questi strumenti consentono di ridurre le emissioni calcolate secondo il metodo market-based del GHG Protocol. Al contrario, un data center on-premise dipende in misura maggiore dal mix elettrico locale. Se l’azienda non ha la possibilità di stipulare contratti diretti di fornitura rinnovabile, il suo impatto climatico resterà legato ai valori location-based, spesso più alti.

La scelta tra cloud e on-premise si riflette direttamente sulla rendicontazione delle emissioni:

- Scope 2 (emissioni indirette da energia acquistata): In un ambiente cloud, le emissioni possono essere calcolate sia *location-based* (intensità media della rete dove è situato il data center) sia *market-based* (considerando PPA e certificati verdi del provider). In un data center on-premise, il valore location-based rappresenta quasi sempre la metrica dominante.
- Scope 3 (emissioni lungo la catena del valore): Nel modello cloud, una parte consistente di queste emissioni – produzione di hardware, costruzione e smaltimento – è contabilizzata dal provider e non dall'utente finale. Un data center on-premise trasferisce invece all'azienda proprietaria la responsabilità di includere tali voci nella propria rendicontazione.

La decisione tra cloud hyperscale e on-premise non è dunque neutrale. Per chi sviluppa o utilizza modelli di intelligenza artificiale generativa – caratterizzati da elevata densità di calcolo e necessità di disponibilità costante – queste considerazioni diventano determinanti nella valutazione di sostenibilità complessiva.

1.3.6 Alimentazione dei data center e strategie energetiche dei provider

La sostenibilità dei data center dipende in misura determinante dal tipo di energia che li alimenta. I principali provider globali – Google, Microsoft e Amazon Web Services (AWS) – hanno quindi sviluppato strategie mirate per garantire una fornitura sempre più pulita e continua, andando oltre il semplice acquisto di elettricità rinnovabile. Questi impianti, che devono garantire continuità 24 ore su 24 e 365 giorni l'anno, operano con margini di tolleranza minimi verso interruzioni o instabilità della rete. Di conseguenza, la disponibilità di energia affidabile, economicamente sostenibile e a basso impatto ambientale è diventata una condizione imprescindibile per la competitività dei provider. L'alimentazione dei data center è divenuta la variabile che più di ogni altra determina l'impronta ambientale reale delle infrastrutture digitali. Come abbiamo già visto, tecnologie hardware sempre più efficienti possono ridurre il consumo per operazione, ma, se l'energia che le alimenta proviene da fonti fossili, l'impatto resta elevato. In questa sezione analizziamo come i provider principali si stanno muovendo, quali strategie adottano e quali sono i numeri che rendono questi interventi non solo desiderabili, ma indispensabili.

1.3.6.1 Le modalità di approvvigionamento energetico

Le principali modalità che un provider può utilizzare per alimentare i propri data center comprendono:

- **Energia prelevata dalla rete pubblica**: la forma più diffusa,. Sebbene garantisca disponibilità costante, il suo impatto ambientale dipende fortemente dal mix di generazione locale. In molti paesi

europei la percentuale di rinnovabile nel mix è tra il 40-80 %, mentre in altre regioni il mix è fortemente dominato da carbone o gas.

- **Accordi di acquisto diretto (Power Purchase Agreements, PPA)**: contratti pluriennali con produttori di energia rinnovabile (eolica, solare, idroelettrica). Questi accordi permettono al provider di ottenere un flusso diretto di energia pulita, spesso con prezzo stabilizzato per molti anni.
- **Produzione on-site o vicina**: installazioni rinnovabili sul sito del data center o molto vicine, che riducono le perdite di trasmissione e migliorano l'affidabilità locale.
- **Sistemi di accumulo e storage**: batterie, accumulatori termici o sistemi di stoccaggio dell'energia rinnovabile che permettono di usare l'energia pulita anche quando la produzione locale diminuisce (es. durante ore di bassa insolazione o scarsa generazione eolica).
- **Gestione operativa della domanda**: spostamento o scheduling dei carichi ("batch jobs") in orari di energia più pulita, uso di tecnologie di load balancing che minimizzano i picchi, throttling delle funzioni non urgenti.

1.3.6.2 Dagli obiettivi di decarbonizzazione ai programmi 24/7 carbon-free

Negli ultimi dieci anni, i grandi operatori cloud hanno evoluto il concetto di "energia verde" da compensazione annuale a neutralità effettiva.

- Compensazione annuale: in passato era sufficiente acquistare in un anno un volume di elettricità rinnovabile pari ai consumi, anche se produzione e consumo non coincidevano temporalmente. L'energia verde immessa in rete in un giorno non necessariamente copriva i consumi notturni del data center. Questa pratica, tuttavia, non assicura che l'energia consumata in ogni singolo momento sia davvero pulita. Da qui nasce il nuovo obiettivo di matching orario (24/7 Carbon-Free Energy)
- 24/7 Carbon-Free Energy (CFE): oggi l'obiettivo è che ogni ora e ogni sito siano alimentati solo da energia realmente priva di carbonio. Significa allineare temporalmente e geograficamente domanda e offerta di elettricità pulita.

Google ha assunto un ruolo pionieristico, puntando a coprire con rinnovabili ogni ora di operatività entro il 2030, attraverso una combinazione di PPA, impianti locali e stoccaggio avanzato (Google, 2020). Microsoft sperimenta contratti orari, come quello con Powerex nello stato di Washington, che usa bacini idroelettrici come "batterie naturali" per redistribuire energia pulita (Microsoft, 2024a). Amazon, a sua volta, ha annunciato di aver raggiunto già nel 2023 il 100 % di energia rinnovabile "abbinata" ai propri consumi e nel 2024 è il primo acquirente corporate di rinnovabili al mondo, con oltre 500 progetti attivi (Amazon, 2024; 2025).

Le strategie descritte hanno effetti concreti. Google dichiara che circa il 70–75 % del suo fabbisogno elettrico globale è già coperto da fonti rinnovabili, mentre Microsoft e AWS hanno firmato contratti per decine di gigawatt di nuova capacità verde (Microsoft, 2024b; Amazon, 2025). Nonostante questi progressi, i report internazionali mostrano che il PUE medio globale (indice di efficienza complessiva) è sostanzialmente stabile attorno a 1,55–1,6, con miglioramenti concentrati nei data center hyperscale di ultima generazione (Uptime Institute, 2024).

Le sfide restano significative: l'intermittenza delle rinnovabili impone sistemi di accumulo costosi; la localizzazione in aree con buone risorse naturali non è sempre compatibile con le esigenze di latenza; i permessi e le regole di connessione alla rete rallentano molti progetti. E soprattutto, la crescita della domanda trainata dall'AI rischia di assorbire gran parte dei nuovi investimenti in rinnovabili, neutralizzando i benefici se non accompagnata da una rigorosa pianificazione della domanda.

Questo paragrafo ha analizzato come L'alimentazione energetica dei data center è dunque il vero crocevia tra intelligenza artificiale e clima. Non basta costruire server farm efficienti: occorre garantire che ogni kWh sia prodotto senza emissioni e nel momento stesso in cui il calcolo avviene. Le esperienze di Google, Microsoft e Amazon mostrano che questo obiettivo è tecnicamente possibile e industrialmente scalabile, ma richiede integrazione di fonti, accumulo e gestione intelligente del carico**,** oltre a forti investimenti regolatori e infrastrutturali.

1.4 Determinanti tecniche del consumo energetico nei modelli AI

La comprensione dell'impatto ambientale dell'AI non può limitarsi alla dimensione infrastrutturale o alle politiche di approvvigionamento energetico. Il consumo effettivo dipende anche da scelte di progettazione e di utilizzo dei modelli. Parametri come architettura e numero di parametri, lunghezza e complessità dei prompt, nonché funzionalità avanzate incidono in modo determinante sul fabbisogno di calcolo e quindi sui kWh necessari per ogni inferenza. Anche fattori operativi spesso invisibili all'utente finale — come batching, caching o parallelismo — giocano un ruolo cruciale nel determinare il consumo complessivo. Benché l'addestramento presenti un consumo energetico unitario elevato, la fase di inferenza, per la sua ripetitività quotidiana su scala globale, rappresenta nel tempo la componente dominante dell'impronta ambientale complessiva.
In questo paragrafo analizziamo i principali driver tecnici del consumo in inferenza, mostrando come le decisioni di ingegneria e di deployment possano ampliare o ridurre in modo significativo l'impatto ambientale di un modello, tenendo comunque conto delle scelte compiute in fase di training.

1.4.1 Dimensione e architettura del modello: Dense, Mixture-of-Experts e Small Language Models

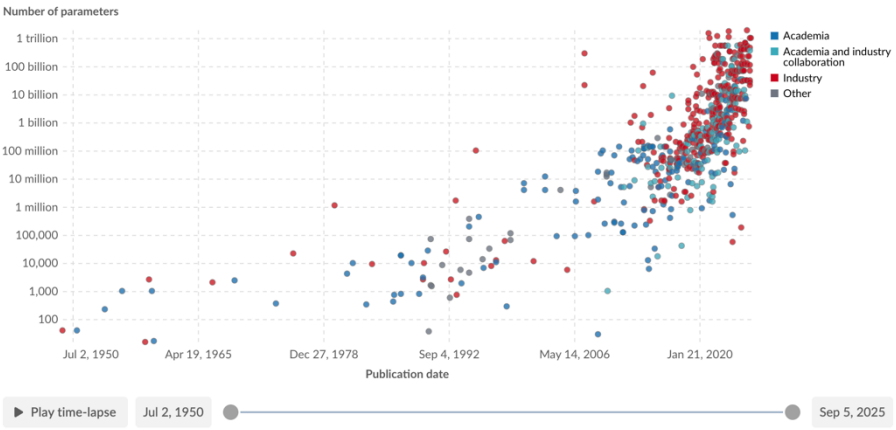
Tra i fattori che più incidono sul fabbisogno energetico dell'intelligenza artificiale c'è la struttura interna del modello, cioè il modo in cui i parametri vengono organizzati e attivati. Capire queste architetture non è solo una questione tecnica: significa comprendere quante operazioni matematiche il sistema deve compiere per rispondere a una singola richiesta e quindi quanta energia elettrica sarà consumata.

I primi grandi modelli linguistici si basano su un'architettura detta **dense**. In questo schema, tutti i parametri – talvolta centinaia di miliardi – partecipano a ogni inferenza. Ogni token di input attraversa l'intera rete di trasformatori (transformer layers), e ciascun livello esegue operazioni di auto-attenzione e feed-forward per generare l'output. La potenza espressiva cresce quasi proporzionalmente al numero di parametri, ma con essa cresce anche il numero di operazioni in floating point (FLOPs) e, di conseguenza, il consumo di elettricità.

Negli ultimi anni si è diffuso un approccio differente, noto come **Mixture-of-Experts** (MoE). L'idea – resa popolare da Google con la famiglia Switch Transformer e poi adottata anche da altri provider – è di attivare solo una piccola parte dei parametri per ogni query. Un gating network sceglie in tempo reale quali "esperti" specializzati sono necessari, mentre gli altri restano inattivi. In pratica, il modello può contenere anche centinaia di miliardi di parametri ma, per ogni richiesta, ne elabora solo una frazione. È un po' come avere

una grande enciclopedia e consultare solo le pagine utili invece di sfogliare l'intero volume: la conoscenza complessiva rimane enorme, ma il calcolo effettivo per singola risposta è più mirato.

Un'evoluzione parallela è rappresentata dai **Small Language Models (SML)**, progettati per fornire prestazioni mirate con un numero di parametri molto più contenuto. Pur basandosi anch'essi su architetture transformer, impiegano tecniche come pruning e distillazione del modello, che eliminano connessioni ridondanti e comprimono lo spazio di calcolo. Queste caratteristiche aprono la strada a un uso capillare in smartphone, robot domestici e applicazioni aziendali dedicate, dove il compromesso tra capacità linguistica e risorse disponibili deve essere ottimizzato.



Queste tre tipologie non rappresentano semplici varianti progettuali, ma strategie ingegneristiche che influenzano profondamente la "fisica" dell'elaborazione: stabiliscono quante moltiplicazioni di matrici vengono eseguite, in quale sequenza e con quali possibilità di parallelizzazione tra GPU. La quantità di operazioni elementari, e quindi il consumo energetico, non dipende dunque solo dal numero di parametri dichiarati, ma dal numero di parametri effettivamente attivati e dal modo in cui il calcolo viene orchestrato.

Comprendere la distinzione fra architetture Dense, Mixture-of-Experts e Small Language Models è quindi cruciale per qualsiasi valutazione seria dell'impronta ambientale dei modelli linguistici di nuova generazione. Senza questo quadro, il solo dato "X miliardi di parametri" rischia di essere fuorviante, perché non rivela quanta energia è effettivamente necessaria per ogni risposta generata.

1.4.2 Prompt engineering

Se l'architettura di un modello ne determina la capacità potenziale, il modo in cui viene formulata la richiesta (promtp) stabilisce quanta di questa capacità viene effettivamente mobilitata. La disciplina che studia e ottimizza queste interazioni è nota come prompt engineering, ed è l'insieme delle metodologie e

delle pratiche utilizzate per progettare input (prompt) in grado di guidare in modo ottimale il comportamento di modelli di intelligenza artificiale generativa, in particolare i Large Language Models (LLM). Questa disciplina è emersa con forza a partire dal 2020, parallelamente alla diffusione di modelli sempre più complessi, ed è oggi considerata una competenza chiave per sviluppatori, ricercatori e imprese che intendano integrare sistemi generativi nei propri processi.

I Large Language Models non comprendono il linguaggio come un essere umano. Generano testo prevedendo, token dopo token, la parola più probabile. Il prompt rappresenta l'interfaccia tra il pensiero umano e questo meccanismo predittivo: un testo che diventa una sequenza di istruzioni numeriche. La sua forma – lunghezza, ordine, chiarezza – decide quante operazioni matematiche il modello deve eseguire e, di conseguenza, quanta energia elettrica viene impiegata.
Un prompt ben progettato non solo migliora la qualità semantica della risposta, ma riduce la quantità di calcolo necessaria. Viceversa, un prompt ridondante o ambiguo può indurre il modello a produrre testi più lunghi o a compiere passaggi logici superflui, moltiplicando i cicli computazionali. In scenari con milioni di query quotidiane, la differenza fra una richiesta ottimizzata e una prolissa diventa significativa sia in termini economici sia ambientali.

1.4.2.1 Tecniche principali di Prompt engineering

Nel corso degli ultimi anni si è consolidata una vera e propria disciplina, articolata in diverse tecniche. Ognuna offre specifici vantaggi e comporta un differente profilo di consumo energetico.

- Zero-shot prompting

È la forma più essenziale: il modello riceve un'istruzione senza alcun esempio esplicativo. Esempio: "Riassumi questo articolo in tre frasi."
Si ha massima semplicità e brevità, quindi basso consumo energetico, ma la mancanza di contesto può generare risposte generiche o incoerenti, talvolta costringendo a ripetere la richiesta.

- Few-shot prompting

Il prompt fornisce alcuni esempi di input e output per guidare il comportamento del modello. In questo modo si ha maggiore coerenza e accuratezza. Dal piano dell'impatto energetico, questo è superiore allo zero-shot, poiché include più token, ma spesso compensato da minori correzioni successive.

- Chain-of-Thought prompting

Questa tecnica invita il modello a esplicitare i passaggi logici prima della risposta finale. Esempio: "Spiega la tua logica passo per passo prima di fornire il risultato del problema matematico."
È la tecnica migliore per il ragionamento complesso, ma l'esplicitazione del ragionamento moltiplica il numero di token da generare e quindi anche l'imaptto ambientale.

- Self-consistency prompting

Il modello genera più possibili soluzioni indipendenti a una stessa domanda e successivamente, seleziona la risposta più coerente e frequente.
Questa strategia può essere combinata con la catena di ragionamento per ottenere risposte ancora più affidabili, ma richiede più calcoli e quindi un dispendio energetico maggiore.

- ReAct prompting

Con il ReAct prompting (Reason + Act) il modello alterna fasi di ragionamento e fasi di azione, ad esempio interrogando una base dati esterna o eseguendo passaggi intermedi prima di dare la risposta finale. Questa tecnica è utile quando il compito richiede l'accesso a informazioni aggiornate o a più fonti. Il vantaggio è una maggiore accuratezza; lo svantaggio è che l'interazione con risorse esterne aumenta la complessità e può comportare ulteriori consumi energetici.

Il reasoning, inteso come capacità del modello di sviluppare catene logiche di passaggi, non è innato ma viene attivato dal prompt. Tecniche come il Chain-of-Thought o approcci come ReAct inducono deliberatamente il modello a ragionare passo-passo, aumentando il numero di token generati e le operazioni matematiche eseguite. Questa maggiore profondità analitica migliora l'accuratezza, ma comporta anche un incremento misurabile di consumo elettrico e di emissioni di CO_2.

1.4.2.2 Evoluzione e prospettive

Il prompt engineering non è una tecnica statica, ma un campo in continua trasformazione, spinto dalla rapida crescita dei modelli e dalle nuove esigenze applicative. Tre direzioni emergenti meritano particolare attenzione per le implicazioni sia funzionali sia energetiche.

Prompt multimodali

La prima tendenza riguarda la multimodalità: i modelli di ultima generazione sono in grado di ricevere e produrre non solo testo, ma anche immagini, audio e video. Ciò consente di affrontare compiti complessi come la descrizione di contenuti visivi, la traduzione automatica di segnali parlati o l'analisi simultanea di dati eterogenei.
Dal punto di vista energetico, tuttavia, l'integrazione di più canali informativi comporta un aumento della potenza di calcolo. Ogni modalità richiede specifici moduli di elaborazione e la loro combinazione impone calcoli aggiuntivi per correlare le informazioni. Di conseguenza, anche richieste apparentemente semplici – ad esempio "riassumi il contenuto di questa immagine e suggerisci un titolo" – attivano catene di inferenza più lunghe e consumano più elettricità rispetto a un prompt solo testuale.

Prompt generativi

Un secondo filone è rappresentato dai prompt generativi, in cui l'AI stessa formula e valuta i propri prompt per ottimizzare la qualità delle risposte. Questo approccio, già sperimentato nei laboratori di ricerca e in alcuni strumenti commerciali, apre la strada a sistemi capaci di auto-migliorarsi iterativamente, selezionando la strategia di interrogazione più efficace in base al compito.

L’autonomia, però, ha un costo: la produzione e la valutazione automatica di molteplici varianti di prompt richiede numerose inferenze interne, con un conseguente incremento del fabbisogno energetico. La sfida consiste quindi nel bilanciare il guadagno di precisione con la necessità di contenere l’energia spesa per il processo di auto-ottimizzazione.

Semantic prompting

La terza linea di sviluppo riguarda il semantic prompting, che integra il modello con basi di conoscenza esterne e dati aggiornati in tempo reale. In questo scenario il prompt non è più solo testo, ma una struttura semantica che attiva ricerche e richiami di dati attraverso API e database. Ciò amplia enormemente il potenziale informativo e consente di fornire risposte sempre contestuali, ma comporta anche operazioni di accesso e aggregazione dati aggiuntive, con impatti energetici da valutare attentamente.

In sintesi, il prompt engineering dimostra che l’efficienza ambientale dell’AI non dipende solo dalle infrastrutture, ma anche dalle scelte linguistiche e progettuali. La capacità di scrivere prompt chiari e mirati diventa quindi una competenza essenziale per ridurre l’impronta carbonica dell’intelligenza artificiale.

1.4.3 Efficienza di esecuzione: batching, caching e parallelismo

L’impatto ambientale dei modelli di intelligenza artificiale non dipende solo dalla dimensione del modello o dalla complessità delle query, ma anche da come le richieste vengono elaborate a livello operativo. Nel passaggio dal singolo prompt ai milioni di inferenze giornaliere, le strategie di esecuzione diventano determinanti per la domanda complessiva di energia. Tra queste, le più rilevanti sono batching, caching e parallelismo, tre tecniche di ottimizzazione che mirano a ridurre i tempi di calcolo e, se correttamente implementate, il consumo elettrico per unità di output.

1.4.3.1 Batching: eseguire più richieste in un unico ciclo

Il batching consiste nell’aggregare più richieste e processarle in un unico ciclo di inferenza, anziché una alla volta. In esoerimenti su server GPU, si è osservato che l’elaborazione simultanea di gruppi di input (batch) consente di sfruttare al massimo la capacità delle GPU, riducendo il numero complessivo di operazioni di avvio e sincronizzazione (Narayanan et al., 2021).
Ad esempio, l’esecuzione di dieci prompt singoli richiederebbe dieci avvii separati del modello, mentre un batch unico di dieci prompt richiede una sola inizializzazione, con un risparmio energetico significativo.

1.4.3.2 Caching: riutilizzare i risultati già calcolati

Il caching è una strategia che evita calcoli ripetuti memorizzando risultati di inferenze precedenti e riutilizzandoli quando la stessa o una simile richiesta si ripresenta. Un esempio tipico è l’elaborazione di

documenti molto lunghi: una volta calcolata l'embedding di una porzione di testo, questo vettore può essere riutilizzato per successive domande, evitando di ricalcolare ogni volta la rappresentazione interna. Secondo stime sperimentali su sistemi di ricerca semantica, il caching può ridurre il fabbisogno computazionale fino al 30–40% in scenari con query ripetitive (Henderson et al., 2023).
L'efficacia di questa tecnica dipende dalla ridondanza del traffico: maggiore è la frequenza di richieste identiche o molto simili, maggiore è il risparmio. È meno efficace in contesti di conversazione creativa, dove ogni prompt è univoco. Ma allo stesso tempo, bisogno considerare che il caching richiede memoria di archiviazione supplementare e un eccesso di dati salvati può annullare parte del vantaggio energetico, soprattutto se lo storage è dislocato su sistemi poco efficienti.

1.4.3.3 Parallelismo: distribuire il carico su più GPU e nodi

Il parallelismo consiste nel suddividere il calcolo tra più GPU o server per accelerare l'inferenza. Le principali modalità sono:

- Data parallelism, in cui ogni GPU elabora porzioni diverse di input.
- Model parallelism, in cui il modello stesso viene suddiviso e distribuito fra più nodi.
- Pipeline parallelism, che segmenta l'elaborazione in fasi consecutive.

Queste strategie permettono di gestire modelli di dimensioni superiori alla memoria di una singola GPU e di aumentare il throughput complessivo.
Dal punto di vista energetico il parallelismo può avere un duplice effetto: da un lato riduce il tempo di calcolo, fino al 40-60% (Google, 2023) e quindi le perdite di efficienza dovute all'inattività dei componenti; dall'altro, se mal configurato può causare overhead di comunicazione e sincronizzazione, aumentando il consumo totale (Patterson et al., 2021).

Dopo aver esaminato in profondità tutte le componenti tecnologiche che determinano l'impronta energetica e climatica dei modelli di intelligenza artificiale (dall'architettura e dalla dimensione dei modelli alla formulazione dei prompt, fino alle modalità di esecuzione e di gestione oeprativa) emerge un punto cruciale: la sfida ambientale non può essere lasciata alle sole scelte progettuali dei provider.

La portata di questi consumi, ormai misurabile e documentata, ha infatti raggiunto una dimensione sistemica, tale da spostare il dibattito ai tavoli decisionali internazionali e alle sedi legislative europee. È in questo contesto che prende forma la necessità di norme specifiche per l'intelligenza artificiale, con l'obiettivo di regolare, misurare e contenere l'impatto ambientale dei modelli generativi.

Il capitolo successivo approfondisce proprio questo passaggio: dall'analisi tecnica alle prime risposte normative, illustrando come l'Unione europea e gli organismi internazionali stiano iniziando a tradurre in obblighi giuridici e standard verificabili l'esigenza di una AI compatibile con gli obiettivi climatici globali.

1.5 Contesto normativo su rendicontazione ambientale, data center e intelligenza artificiale

La rapida espansione dell'intelligenza artificiale generativa richiede un quadro regolatorio capace di garantire trasparenza e responsabilità ambientale. Il contesto normativo europeo e internazionale, negli ultimi anni, ha introdotto obblighi di rendicontazione non finanziaria e standard di trasparenza ambientale sempre più articolati. Queste norme, nate in parte per finalità generali di sostenibilità, stanno diventando la base per affrontare l'impatto climatico ed energetico dei modelli di intelligenza artificiale.

Nella seguente sezione vengono esaminate le principali norme e iniziative che incidono, direttamente o indirettamente, sull'impatto ambientale dei modelli di AI, dalla rendicontazione ESG delle imprese, ai requisiti per i data center, fino alle prime regole specifiche per i modelli di intelligenza artificiale.

1.5.1 Direttiva NFRD e Direttiva CSRD: il pilastro europeo della rendicontazione non finanziaria

Il primo passo è rappresentato dalla Direttiva 2014/95/UE, conosciuta come Non-Financial Reporting Directive (NFRD). Essa impone alle grandi imprese di interesse pubblico, con oltre 500 dipendenti, l'obbligo di rendicontare informazioni non finanziarie riguardanti ambiente, questioni sociali, diritti umani e governance.
L'obiettivo è fornire agli stakeholder una visione completa dei rischi e degli impatti ambientali e sociali delle attività d'impresa (European Parliament, 2021[8]). Pur costituendo una svolta, la NFRD presentava due limiti principali:

- ampie discrezionalità nella scelta delle metriche, che riducevano la comparabilità tra aziende
- copertura limitata alle sole imprese di grandi dimensioni.

Per superare queste criticità, l'UE ha approvato nel 2022 la Corporate Sustainability Reporting Directive (CSRD), che entra in vigore in maniera progressiva dal 2024 in poi. La CSRD estende l'obbligo di reporting a circa 50.000 imprese europee, includendo società di medie dimensioni e società extra-UE con attività significative in Europa. La direttiva richiede dati dettagliati e standardizzati sulle emissioni (Scope 1, 2 e, in molti casi, 3), sui consumi energetici, sull'uso di risorse idriche e sull'impatto climatico complessivo

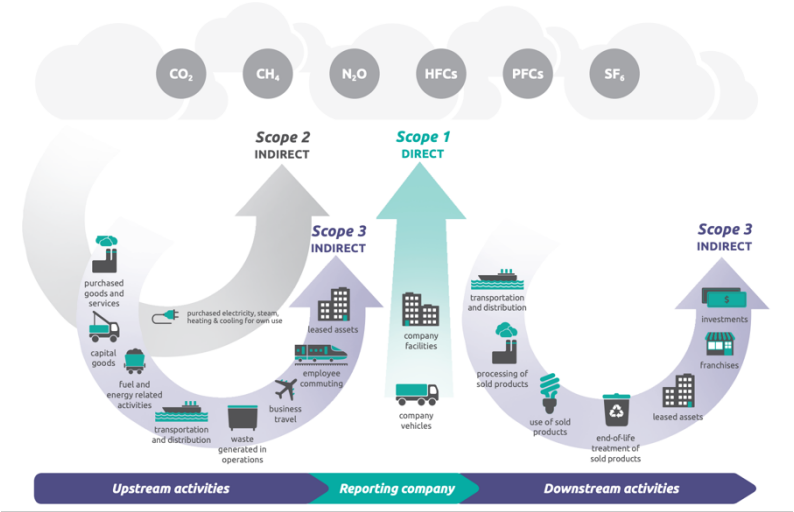
[8] European Parliament (2021) *Non-Financial Reporting Directive: Implementation and challenges*

(Digital Realty, 2024[9]). Le imprese devono redigere report conformi agli European Sustainability Reporting Standards (ESRS) e garantire verifica indipendente delle informazioni.

1.5.1.1 Nota metodologica sulle emissioni Scope 1, 2 e 3

La classificazione Scope 1–2–3 è stata definita dal Greenhouse Gas Protocol (GHG Protocol)[10], sviluppato alla fine degli anni '90 dal World Resources Institute (WRI) e dal World Business Council for Sustainable Development (WBCSD) con l'obiettivo di creare un framework comune e verificabile per calcolare e comunicare le emissioni di gas serra delle organizzazioni.

Figure [1.1] Overview of GHG Protocol scopes and emissions across the value chain



- Scope 1: emissioni dirette provenienti da fonti possedute o controllate dall'organizzazione (ad es. combustione di carburanti, perdite di gas refrigeranti).
- Scope 2: emissioni indirette derivanti dalla produzione dell'elettricità, calore o vapore acquistati e consumati.
- Scope 3: tutte le altre emissioni indirette lungo la catena del valore, a monte e a valle (ad es. produzione e smaltimento dell'hardware, logistica, uso dei prodotti da parte degli utenti finali).

Questa tassonomia, adottata nelle European Sustainability Reporting Standards (ESRS) e nella Corporate Sustainability Reporting Directive (CSRD), costituisce il riferimento tecnico per la rendicontazione climatica delle imprese.

[9] Digital Realty (2024) *EU sustainability regulations: CSRD and beyond*.

[10] Le categorie Scope 1, Scope 2 e Scope 3 nascono dal Greenhouse Gas Protocol (GHG Protocol), lo standard internazionale più usato per misurare e rendicontare le emissioni di gas serra.
Il GHG Protocol è stato sviluppato alla fine degli anni '90 da World Resources Institute (WRI) (USA) e World Business Council for Sustainable Development (WBCSD) (Svizzera), in collaborazione con governi, imprese e ONG.
La prima versione è stata pubblicata nel 2001 e aggiornata più volte, fino ad arrivare all'ultima edizione "Corporate Standard" del 2015.

1.5.1.2 European Sustainability Reporting Standards (ESRS)

Gli ESRS sono stati elaborati dall'EFRAG (European Financial Reporting Advisory Group) su incarico della Commissione Europea, nell'ambito del Corporate Sustainability Reporting Directive (CSRD). L'obiettivo è definire standard comuni, comparabili, trasparenti per la rendicontazione ambientale, sociale e di governance (ESG) di molte imprese europee, ben più numerose rispetto al vecchio obbligo NFRD. Ogni azienda soggetta alla CSRD dovrà redigere un bilancio di sostenibilità secondo questi standard, garantendo trasparenza, comparabilità e verificabilità delle informazioni.

Gli ESRS sono divisi in standard trasversali e standard tematici:

1. **Standard trasversali** (per tutte le aziende e tutti i settori)

- **ESRS 1 – Requisiti generali**: definisce principi di double materiality (materialità finanziaria e di impatto), trasparenza, tracciabilità dei dati e coerenza tra informativa finanziaria e non finanziaria.
- **ESRS 2 – General disclosures**: richiede di descrivere governance, ruoli e processi di controllo dei rischi ambientali, compresi meccanismi di due diligence e politiche di remunerazione legate alla sostenibilità.

2. **Standard ambientali specifici**

- **ESRS E1 – Climate change**: è il modulo più rilevante per l'energia e il clima. Impone di misurare e dichiarare:
    - emissioni di gas serra *Scope 1, 2 e 3*;
    - consumi energetici totali e composizione del mix elettrico;
    - obiettivi quantitativi di riduzione e progressi annuali.
- **ESRS E2 – Pollution**: obblighi relativi a inquinamento dell'aria, suolo e acqua.
- **ESRS E3 – Water and marine resources**: include consumo idrico e impatti su ecosistemi acquatici, molto rilevante per il raffreddamento dei data center.
- **ESRS E4 – Biodiversity and ecosystems**: tutela della biodiversità.
- **ESRS E5 – Resource use and circular economy**: copre l'uso di materie prime, rifiuti e riciclo dell'hardware.

Tra gli aspetti più significativi, lo standard ESRS E1 – Climate Change richiede di misurare e comunicare in modo dettagliato l'energia consumata e le emissioni di gas serra lungo tutto il ciclo di vita dei modelli, includendo non solo la fase di addestramento ma anche l'inferenza quotidiana, che rappresenta la quota in più rapida crescita dei consumi (EFRAG, 2023)[11]. La rendicontazione deve estendersi inoltre alle emissioni indirette (Scope 3), comprendendo la produzione, il trasporto e il fine vita dell'hardware utilizzato. Le aziende devono fissare obiettivi misurabili di riduzione e aggiornarli nel tempo, il che incoraggia scelte come l'adozione di data center alimentati da rinnovabili, l'uso di hardware più efficiente e la pianificazione dei carichi in fasce orarie a bassa carbon intensity[12].

[11] EFRAG (2023) *European Sustainability Reporting Standards: Delegated Act and Annexes*. Brussels: European Commission.
[12] Digital Realty (2024) *EU sustainability regulations: CSRD and beyond*.

Tutte le informazioni devono essere verificate da soggetti terzi, garantendo trasparenza e possibilità di confronto tra operatori. Restano però alcune criticità: gli standard sono recenti e mancano ancora linee guida univoche per distinguere con precisione le emissioni legate all'addestramento da quelle derivanti dall'inferenza; la stima delle emissioni Scope 3 è complessa; l'assicurazione indipendente richiede competenze specialistiche e può aumentare i costi iniziali[13].
Questa evoluzione crea comunque le basi giuridiche per includere, anche indirettamente, l'impatto energetico e idrico legato all'uso dell'AI, specialmente nei settori ad alta intensità digitale.

1.5.2 Regolamento UE sui data center e Energy Efficiency Directive (EED)

Un tassello fondamentale della strategia europea per la sostenibilità digitale è rappresentato dal Regolamento (UE) 2024/1364 sui data center, che introduce obblighi vincolanti di rendicontazione ambientale per tutti i centri dati con un consumo IT pari o superiore a 500 kW. A partire da settembre 2024, i gestori devono compilare e trasmettere, su base annuale, una serie di indicatori chiave di performance (KPIs) relativi all'efficienza e all'impatto ambientale delle proprie strutture. Tra le informazioni richieste figurano:

- l'**efficienza energetica complessiva**, misurata ad esempio attraverso il Power Usage Effectiveness (**PUE**);
- le **caratteristiche dei sistemi di raffreddamento** e i relativi consumi;
- il **volume totale di energia assorbita** e il **mix di fonti** impiegato;
- i **volumi di traffico dati trattati** e la capacità operativa.

I dati devono essere inviati a un registro nazionale e, in prospettiva, confluiranno in una piattaforma europea centralizzata per consentire il monitoraggio e il confronto tra Paesi (Tera Energy, 2024[14]). L'obiettivo è creare un quadro trasparente e comparabile delle performance ambientali dei data center, che sono l'infrastruttura critica per l'addestramento e l'inferenza dei modelli di intelligenza artificiale generativa.
Complementare a questo regolamento è la Energy Efficiency Directive (EED), una direttiva quadro già vigente in varie revisioni, che stabilisce obiettivi vincolanti di efficienza energetica per l'intera Unione Europea. La versione più recente rafforza le disposizioni riguardanti il settore ICT, imponendo:

- monitoraggio costante della performance energetica;
- piani di miglioramento continuo, con traguardi misurabili di riduzione dei consumi;
- criteri per la riqualificazione energetica delle infrastrutture esistenti.

[13] KPMG (2024) *European Sustainability Reporting Standards (ESRS): key insights*.
[14] Tera Energy (2024) *Nuove normative UE per il reporting sulla sostenibilità dei data center*.

La EED mira così a integrare gli obiettivi climatici europei nella gestione operativa dei data center, favorendo l'adozione di fonti rinnovabili e di soluzioni tecnologiche ad alta efficienza (GRESB, 2024[15]). Queste norme hanno ricadute dirette sulla filiera della generative AI. I modelli di grandi dimensioni dipendono infatti da data center ad alta capacità, sia per l'addestramento sia per l'inferenza quotidiana. L'obbligo di rendicontazione dei consumi energetici e delle emissioni impone ai provider di intelligenza artificiale di selezionare con maggiore attenzione le infrastrutture, favorendo quelle che offrono metriche certificate di efficienza e trasparenza. Inoltre, la combinazione tra reporting annuale e obiettivi di miglioramento continuo spinge l'intero settore verso pratiche di ottimizzazione del carico di lavoro, pianificazione degli orari in base alla carbon intensity e adozione di tecniche di raffreddamento a minore impatto idrico.

1.5.3 Regolamento europeo sull'AI (AI Act)

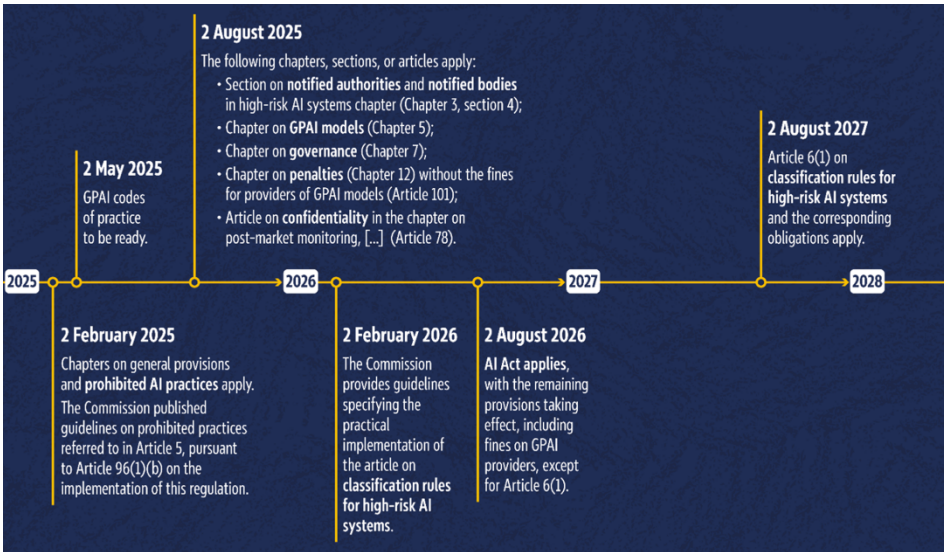
Il Regolamento (UE) 2024/1689, noto come Artificial Intelligence Act (AI Act), pubblicato nella Gazzetta Ufficiale dell'Unione Europea il 12 luglio 2024 ed entrato in vigore il 1° agosto 2024, stabilisce per la prima volta nell'UE un quadro normativo organico per l'intelligenza artificiale. L'obiettivo è garantire che lo sviluppo e l'uso dei sistemi di AI, inclusi i modelli generativi di uso generale (General-Purpose AI Models, GPAI), avvengano in modo sicuro, trasparente e rispettoso dei diritti fondamentali, evitando applicazioni indiscriminate o lesive. Sebbene la ratio originaria del regolamento sia prevalentemente legata alla sicurezza, alla tutela dei dati personali e alla prevenzione dei rischi sociali, l'AI Act contiene disposizioni innovative di rilievo ambientale, che toccano direttamente anche i modelli generativi di uso generale.

Il regolamento è prescrittivo: fissa obblighi puntuali e direttamente applicabili in tutti gli Stati membri, pur lasciando a ciascuno il compito di designare le autorità nazionali competenti e di coordinare i controlli. Gli obblighi si applicano con scadenze progressive[16], definite a livello europeo:

- 2 novembre 2024 – gli Stati membri devono indicare e notificare alla Commissione europea le autorità nazionali competenti.
- 2 febbraio 2025 – entrano in vigore i divieti per i sistemi di AI a rischio inaccettabile, come il social scoring o la manipolazione cognitiva, e le prime disposizioni su alfabetizzazione e trasparenza.
- 2 maggio 2025 – la Commissione deve pubblicare i codici di condotta (Codes of Practice) per i modelli di uso generale.
- 2 agosto 2025 – decorrono gli obblighi specifici per i GPAI, inclusa la predisposizione di documentazione tecnica dettagliata, la stima o misurazione del consumo energetico e l'adozione di sistemi di logging e audit indipendenti.

[15] GRESB (2024) *Sustainable data centers: ESG compliance and futureproofing for success*.
[16] European Commission (2024) *EU Artificial Intelligence Act: Implementation timeline*.

- 2 agosto 2026 – applicazione generalizzata dell'AI Act: tutte le imprese e le pubbliche amministrazioni che forniscono o utilizzano sistemi di intelligenza artificiale nell'Unione Europea devono essere conformi.
- 2 agosto 2027 – termine ultimo per l'adeguamento dei modelli GPAI già presenti sul mercato prima del 2 agosto 2025.



Sul piano della sostenibilità, il regolamento introduce tre novità giuridiche rilevanti:

1. **Riconoscimento del valore della protezione ambientale**
   Il legislatore europeo nelle General e Final Previsions (Chapter 1 e 13, Art 1- 4 e 102-113) afferma che l'AI deve essere sviluppata e utilizzata in coerenza con i valori dell'Unione, comprendendo esplicitamente la protezione dell'ambiente. Questo principio, pur non essendo una norma immediatamente precettiva, fornisce una base interpretativa che legittima l'inclusione dell'impatto ambientale nei successivi atti di esecuzione e negli standard tecnici (European Union, 2024[17]).
2. **Obblighi per i modelli di uso generale (GPAI)**
   I fornitori di grandi modelli generativi sono tenuti a predisporre una **documentazione tecnica dettagliata**, comprensiva, laddove tecnicamente possibile, di **stime o misurazioni del consumo energetico e dell'uso delle risorse** durante **addestramento, distribuzione e inferenza**. Il regolamento prevede inoltre l'adozione di sistemi di logging che permettano la tracciabilità e l'audit indipendente delle performance.
3. **Sviluppo di standard e codici di condotta per l'efficienza**
   La Commissione europea è incaricata di collaborare con gli organismi di normazione per elaborare **standard europei** e **codici di condotta** che includano indicatori di performance ambientale, obiettivi di riduzione dei consumi e pratiche di progettazione sostenibile, promuovendo la convergenza verso metriche uniformi.

[17] European Union (2024) *Regulation (EU) 2024/1689 laying down harmonised rules on Artificial Intelligence (AI Act)*. Official Journal of the European Union, 12 July.

Questo regolamento può essere considerato un primo pilastro normativo a livello europeo nella governance ambientale dell'intelligenza artificiale. Al tempo stesso, però, non garantisce da solo la neutralità climatica del settore. Per colmare il divario sarà cruciale l'adozione di atti delegati e **standard europei** dettagliati che specifichino metriche e limiti, la **sinergia con altre normative** (CSRD, ESRS, regolamento sui data center, EED) per una rendicontazione piena e un **coordinamento internazionale**, che eviti il rischio di "delocalizzazione" dei consumi verso Paesi con normative meno esigenti.
Nonostante le innovazioni, l'AI Act non è ancora una disciplina ambientale completa. Le principali criticità sono:

- Assenza di obblighi quantitativi precisi: il regolamento menziona il consumo energetico, ma non definisce metodologie univoche né fissa soglie o limiti.
- Copertura parziale dell'inferenza: molte disposizioni si concentrano sulla fase di addestramento, mentre l'uso quotidiano, che rappresenta la quota in più rapida crescita dei consumi, non è sempre oggetto di obblighi espliciti.
- Scope 3 e catena di fornitura: non vi è un obbligo generalizzato di rendicontare le emissioni indirette, come la produzione e il riciclo dell'hardware[18].
- Enforcement da definire: i meccanismi di controllo e le sanzioni specifiche per le violazioni ambientali non sono ancora dettagliati.

Queste lacune confermano che, pur segnando una svolta concettuale, l'AI Act resta più vicino alla regolazione della sicurezza e della protezione dei diritti che a una disciplina compiutamente ambientale.

1.5.4 Iniziative extra-UE e raccomandazioni internazionali

Sul piano globale si stanno sviluppando cornici normative e linee guida complementari che rafforzano l'orientamento verso una governance ambientale dell'AI:

- Stati Uniti – L'*Artificial Intelligence Environmental Impacts Act of 2024* prevede la costituzione di un consorzio pubblico-privato incaricato di sviluppare sistemi di reporting volontario per i consumi energetici e le emissioni associate ai modelli di intelligenza artificiale. L'iniziativa, pur non vincolante, punta a costruire metriche comuni e a fornire dati trasparenti a decisori pubblici e investitori, ponendo le basi per possibili obblighi federali futuri (United States Congress, 2024[19]).

[18] Alder, N., Ebert, K., Herbrich, R. and Hacker, P. (2024) *AI, Climate, and Regulation: From Data Centers to the AI Act*. arXiv preprint arXiv:2410.06681.
[19] United States Congress (2024) *Artificial Intelligence Environmental Impacts Act of 2024*. 118th Congress, Washington D.C.

- Organizzazione per la Cooperazione e lo Sviluppo Economico (OECD). Le sue raccomandazioni sollecitano l'adozione di metriche standardizzate e l'introduzione di obblighi di disclosure per quantificare l'impatto ambientale dei sistemi di AI, dalle fasi di addestramento fino all'utilizzo quotidiano (OECD, 2024[20]).
- Programma Ambiente delle Nazioni Unite (UNEP), invita a una trasparenza globale e a forme di cooperazione multilaterale, affinché governi e imprese condividano dati e metodologie per ridurre l'impronta climatica delle tecnologie digitali e dei data center (UNEP, 2024[21]).

Queste iniziative, seppure con approcci differenti, convergono su un principio comune: l'efficienza tecnologica non è sufficiente se non accompagnata da misurazione rigorosa, reporting e responsabilità pubblica. In prospettiva, il coordinamento tra il quadro vincolante europeo e le raccomandazioni internazionali potrà favorire la nascita di standard globali di rendicontazione ambientale per l'intelligenza artificiale.

[20] OECD (2024) *Measuring the environmental impacts of artificial intelligence: compute and applications*. Paris: OECD Publishing.
[21] UNEP (2024) *AI has an environmental problem: Here's what the world can do about it*. United Nations Environment Programme

Capitolo 2 – Analisi della Letteratura e Gap di Ricerca

2.1 Introduzione

L'interesse scientifico verso l'impatto ambientale dell'intelligenza artificiale ha iniziato a consolidarsi soprattutto intorno al 2019-2020, quando l'AI generativa, e in particolare i Large Language Models (LLM), ha suscitato un dibattito non solo per le potenzialità applicative ma anche per i costi ambientali connessi. Se fino ad allora l'attenzione era quasi esclusivamente sull'accuratezza dei modelli, in quegli anni alcuni lavori pionieristici hanno mostrato che l'addestramento di grandi reti neurali poteva comportare costi energetici ed emissivi molto elevati. La crescita esponenziale della domanda di potenza computazionale, legata sia al training che all'inferenza, ha reso necessario un approccio critico che vada oltre l'accuratezza dei modelli, includendo come variabile imprescindibile il consumo energetico.

2.2 Revisione della Letteratura

Per chiarezza espositiva, i contributi della letteratura scientifica legata a questo tema si articolano in tre direttici principali: i lavori pionieristici che hanno introdotto il problema, le misurazioni sperimentali che hanno tentato di quantificare l'impatto ambientale e gli studi più recenti che hanno cercato di proporre standard e framework di riferimento, in ottica di rendicontazione industriale.

2.2.1 I primi studi sull'impatto energetico e ambientale dell'IA (2019–2020)

La ricerca sull'impatto ambientale dell'intelligenza artificiale ha preso avvio con alcuni lavori pionieristici che hanno quantificato per la prima volta i costi energetici del deep learning. In questo contesto si collocano i contributi fondativi di Strubell et al. (2019) e Schwartz et al. (2020), che hanno avuto il merito di aprire il dibattito collegando l'avanzamento dell'IA ai suoi costi ambientali.

Strubell, Ganesh e McCallum (2019) hanno stimato il consumo energetico e le emissioni derivanti dall'addestramento di modelli di NLP, algoritmi di Intelligenza Artificiale che consentono alle macchine di comprendere, interpretare e generare il linguaggio umano. Essi utilizzano il Machine Learning e il Deep Learning per analizzare testi e imparare il significato delle parole nel contesto. Tra i modelli più diffusi ci sono i modelli basati su reti neurali come i Transformer, che hanno rivoluzionato il campo con prestazioni elevate e si sono evoluti nei modelli linguistici di grandi dimensioni (LLM). Lo studio ha evidenziato come il costo ambientale potesse raggiungere ordini di grandezza comparabili al ciclo di vita di più automobili (Strubell, Emma; Ganesh, Ananya; McCallum, Andrew, *Energy and Policy Considerations for Deep Learning in NLP*, 2019). Il risultato principale è stato quello di quantificare un ordine di grandezza, concludendo che l'addestramento di un singolo modello può generare fino a 284 tonnellate di CO_2. Da questo momento si capisce che i progressi dell'IA avevano un costo materiale, finora ignorato, che non poteva più essere trascurato.

Se Strubell et al. (2019) avevano mostrato per la prima volta che l'addestramento di modelli NLP poteva avere un'impronta paragonabile al ciclo di vita di più automobili, Schwartz et al. (2020) hanno fatto un passo oltre, proponendo il paradigma del Green AI e sostenendo che la ricerca non dovesse più essere guidata soltanto dall'accuratezza, ma anche dall'efficienza. La loro posizione è stata rivoluzionaria perché non proponeva una misurazione puntuale, bensì un cambio di paradigma: considerare i costi computazionali come metrica di valutazione al pari delle performance (Schwartz, Roy; Dodge, Jesse; Smith, Noah; Etzioni, Oren, *Green AI*, 2020). Hanno dimostrato che in molti casi è possibile ottenere performance simili con modelli più piccoli, meno energivori e meno costosi. Nello studio mostrano che se la comunità scientifica avesse continuato a valutare i modelli esclusivamente sulla base di metriche di performance (BLEU, F1, accuracy), si sarebbe rischiato di premiare modelli sempre più grandi e costosi dal punto di vista energetico, anche quando i miglioramenti ottenuti erano marginali. In questo senso, la competizione accademica e industriale a spingere continuamente in avanti i record rischiava di trasformarsi in una corsa insostenibile dal punto di vista ambientale.

Questa riflessione sulla necessità di considerare l'efficienza come metrica centrale trova punto di interesse negli studi di Patterson et al. (2021) che hanno iniziato a spostare l'attenzione non solo sul modello, ma anche sul contesto in cui esso opera.

Patterson et al. (2021) hanno mostrato che l'impatto ambientale dei modelli non dipende soltanto dalla loro dimensione o architettura, ma anche dal contesto geografico e infrastrutturale in cui vengono eseguiti. Gli autori evidenziano come il mix energetico della regione, l'efficienza dei data center e la generazione hardware possano modificare di ordini di grandezza le emissioni associate a uno stesso carico di lavoro (Patterson et al., Carbon Footprint of AI, 2021). Il risultato principale è la dimostrazione che la sostenibilità dell'IA non può essere valutata in astratto, ma deve tener conto delle condizioni operative, preparando così il terreno per le analisi LCA e per i successivi interventi di carbon-aware scheduling.

2.2.2 Dalla fase di training all'inferenza: misurazioni sperimentali e analisi empiriche

Dopo i primi studi che avevano quantificato l'impatto ambientale dell'addestramento dei modelli e introdotto il paradigma del Green AI, l'attenzione della comunità scientifica si è progressivamente spostata oltre il training, verso una prospettiva più ampia che considera l'intero ciclo di vita dei modelli. In particolare, la fase di inferenza ha iniziato a emergere come componente centrale, poiché ripetuta milioni di volte al giorno su scala globale, con un impatto cumulativo spesso superiore a quello del training stesso.

Un contributo emblematico in questa direzione è lo studio di Luccioni, Viguier e Ligozat (2023) dedicato al modello BLOOM, un LLM open-source di 176 miliardi di parametri sviluppato come alternativa trasparente ai grandi modelli proprietari. Gli autori hanno applicato un approccio di Life Cycle Assessment (LCA), cioè

una metodologia che tiene conto di tutte le fasi: dalla produzione dell'hardware, all'addestramento, fino all'utilizzo quotidiano (inferenza). Questa prospettiva ha permesso di spostare lo sguardo da una valutazione episodica dei costi (tipicamente centrata sul training) a una visione integrata e continua dell'impatto.
I risultati hanno mostrato un dato sorprendente: in scenari realistici di largo utilizzo, la fase di inferenza può superare quella di addestramento in termini di emissioni totali di CO_2. Il motivo è semplice ma potente: mentre il training è un evento concentrato e relativamente raro, l'inferenza si ripete milioni di volte al giorno su scala globale, accumulando consumi che, nel tempo, pesano di più. BLOOM, in questo senso, è diventato un "caso di scuola" perché ha permesso di osservare empiricamente come la dimensione dell'impatto non dipenda solo dal costo iniziale di addestramento, ma dal modo in cui il modello viene distribuito e utilizzato.
Questo studio ha segnato un punto di svolta nella letteratura: ha ribaltato l'assunto implicito che i costi ambientali fossero legati soprattutto all'addestramento, evidenziando invece che la vera sfida si colloca nella fase d'uso quotidiano. L'apporto alla conoscenza non è stato solo quello di quantificare nuove grandezze, ma di spostare l'agenda di ricerca verso l'inferenza come oggetto centrale di analisi.

Lo studio su BLOOM ha segnato un punto di svolta, mostrando che l'inferenza poteva avere un peso maggiore del training e che il mix energetico era un fattore decisivo. Queste intuizioni sono state poi generalizzate da Wu et al. (2024), che su Nature hanno analizzato modelli proprietari di larga scala come GPT-3 e PaLM, giungendo a risultati analoghi ma su scala comparativa più ampia. L'integrazione dei due contributi rafforza l'evidenza che l'impatto ambientale non dipende soltanto dalla grandezza del modello, ma dal contesto infrastrutturale e geografico in cui viene addestrato e servito (Wu et al. (2024) – *The Carbon Footprint of Large Language Models).*

Mentre Luccioni (2023) e Wu (2024) hanno spostato il focus sull'impatto complessivo dei modelli, in parallelo si è sviluppata una seconda linea di ricerca, basata su misurazioni sperimentali a livello di query. È in questo contesto che si colloca il lavoro di Samsi et al. (2023), con il progetto From Words to Watts. Gli autori hanno condotto una serie di benchmark sperimentali per analizzare la relazione tra la lunghezza dei prompt forniti ai modelli linguistici e l'energia effettivamente consumata durante l'elaborazione.

Il risultato più significativo è che il consumo cresce in maniera quasi lineare con il numero di token in input: ogni token aggiuntivo diventa un'unità di misura del costo energetico. In altri termini, il prompt non è soltanto un fattore di accuratezza o di contenuto, ma diventa anche un driver ambientale. Questo ha portato a un primo cambio di prospettiva: per valutare la sostenibilità dell'IA, non basta più guardare al modello, bisogna guardare anche al tipo di query che gli viene sottoposta.

Dal punto di vista metodologico, lo studio di Samsi et al. è rilevante perché utilizza un approccio sperimentale ingegneristico, basato su misure dirette di potenza e consumo su GPU (in particolare NVIDIA V100 e A100), anziché limitarsi a stime teoriche. Ciò ha permesso di produrre dati concreti, ripetibili e comparabili, che hanno rafforzato l'affidabilità delle conclusioni.

Dopo i primi benchmark sperimentali, la letteratura ha cominciato a esplorare il consumo energetico dell'inferenza in modo più fine, considerando non soltanto l'input, ma anche ciò che il modello produce. In questa direzione si colloca il lavoro di Husom et al. (2024), che con il framework MELODI (Measuring the Energy cost of Large language mOdel DIalogs) ha proposto un approccio innovativo di profilazione in tempo reale.

Gli autori hanno sviluppato un sistema capace di monitorare in maniera diretta e continua l'assorbimento energetico delle GPU e delle CPU durante l'esecuzione delle query. Questo ha permesso di misurare con grande precisione il costo energetico di ogni singola richiesta all'LLM, non più solo come media aggregata. La forza del loro contributo è proprio nell'aver portato l'analisi "al livello della query", restituendo un quadro micro che mostra quanta variabilità possa esserci anche tra richieste apparentemente simili.

I risultati hanno evidenziato che la lunghezza dell'output è un driver determinante: a parità di input, due risposte di lunghezza diversa possono avere consumi radicalmente differenti. In altre parole, il peso energetico non dipende solo da quanto chiediamo al modello, ma soprattutto da quanto il modello sceglie di generare. Questo ha ribaltato una certa percezione diffusa: se i token in input erano già stati riconosciuti come variabile critica (Samsi et al., 2023), Husom et al. hanno dimostrato che il vero costo risiede nei token in output, che risultano ancora più onerosi in termini di energia.

Un altro elemento interessante emerso è la variabilità elevata tra query simili. Lo stesso modello, con input analoghi, può consumare quantità diverse di energia in funzione delle strategie di generazione, dei tempi di calcolo e dei percorsi interni della rete. Questo suggerisce che la sostenibilità non dipende soltanto dal design architetturale o dall'hardware, ma anche da come il modello viene effettivamente usato e dai parametri di inferenza.

Tra gli studi più recenti si colloca anche il contributo di Sánchez-Mompó et al. (2025), che ampliano il dibattito analizzando le differenze tra modelli generativi e discriminativi. I modelli generativi, come i Large Language Models, sono progettati per produrre nuovi contenuti (testo, immagini, suoni) a partire da un prompt, estendendo quindi l'informazione oltre ciò che è fornito in input. I modelli discriminativi, al contrario, svolgono compiti di classificazione o riconoscimento. Questi modelli ricevono un input e devono assegnarlo a una categoria già nota, ad esempio distinguere tra spam e non-spam in un'email, o riconoscere un oggetto in un'immagine. L'analisi mostra che i modelli generativi risultano sistematicamente più energivori, poiché l'elaborazione autoregressiva e la lunghezza tipica delle sequenze in output comportano consumi molto superiori a parità di compito (*Sánchez-Mompó, Mavromatis, Li et al., Green MLOps to Green GenOps, 2025).* Pur basandosi su un campione limitato, questo studio conferma quanto emerso dalle misurazioni query-level: il costo ambientale cresce in modo proporzionale al numero di token generati. Il risultato principale è quello di chiarire che i modelli generativi, e in particolare gli LLM, rappresentano oggi il fulcro del problema ambientale dell'IA, fornendo così una giustificazione ulteriore alla scelta di questa ricerca di concentrarsi sulla latenza come variabile critica.

2.2.3 Benchmark industriali e approcci integrati: dal consumo di energia alla water footprint

Parallelamente agli studi accademici, il mondo industriale ha riconosciuto la necessità di introdurre standard comuni di misurazione per le prestazioni e i consumi dei modelli di intelligenza artificiale. Questo ruolo è stato assunto da MLCommons, un consorzio no-profit che dal 2018 sviluppa benchmark di riferimento per la comunità scientifica e industriale. La loro suite più rilevante in questo contesto è MLPerf Inference, giunta nel 2024 alla versione 5.1.
I risultati più rilevanti emersi negli ultimi report riguardano le differenze tra generazioni di GPU. In particolare, il passaggio da NVIDIA A100 a H100 ha mostrato salti significativi in termini di prestazioni per watt: a parità di carico, gli H100 consumano meno energia per token generato e permettono tempi di latenza inferiori. Questo tipo di evidenza, pur non accompagnata da un'analisi econometrica, ha avuto un impatto enorme perché ha fornito alla comunità scientifica dati solidi per sostenere che la scelta dell'hardware è una delle leve più immediate per ridurre i costi ambientali dell'inferenza.
Dal punto di vista metodologico, MLPerf ha introdotto anche la categoria "Power", che richiede ai partecipanti di riportare non solo throughput e latenza, ma anche consumo elettrico misurato. Anche se la partecipazione a questa categoria è ancora limitata, ha segnato un passo decisivo verso la trasparenza sui costi ambientali.
Il limite di MLPerf è che rimane un benchmark pensato per l'industria: non offre spiegazioni sui meccanismi che producono i consumi, né considera variabili ambientali come carbon intensity o WUE. Tuttavia, la sua importanza sta nell'aver reso l'efficienza un criterio competitivo, inserendola di fatto nell'agenda tecnologica globale.

Dopo i primi contributi focalizzati sui consumi in addestramento o sulla misurazione di energia prodotta per singola query, la letteratura ha iniziato a spostarsi verso un approccio più ampio, indagando su quale fosse l'impatto complessivo dell'intero ciclo di vita dei modelli. È in questo contesto che si colloca il contributo di Faiz et al. (2023), con il progetto LLMCarbon, uno dei primi a stimare in maniera integrata l'impronta carbonica dei Large Language Models lungo tutte le fasi, dall'addestramento all'inferenza fino alle emissioni embodied. L'obiettivo dichiarato era quello di rispondere a una domanda innovativa: non solo quanto consuma un'inferenza, ma quanto pesa complessivamente un modello linguistico lungo l'intero suo ciclo di vita.

Questa impostazione ha permesso di mostrare un risultato controintuitivo fino a quel momento, ma fondamentale per gli studi successivi: in scenari reali di larga adozione, l'inferenza può avere un impatto ambientale superiore a quello del training. Se in passato la letteratura aveva concentrato quasi tutta l'attenzione sull'addestramento, *LLMCarbon* ha dimostrato che l'uso quotidiano di milioni di query al giorno genera un carico emissivo continuo che, accumulandosi, finisce per superare quello, pur ingente, della fase di training iniziale.

Accanto agli studi accademici, anche i grandi provider hanno iniziato a rendere disponibili dati sui propri data center. Un aspetto di crescente rilevanza riguarda l'impatto idrico dei data center, calcolato come i litri d'acqua utilizzati per il raffreddamento per ogni kWh IT consumato, tradizionalmente trascurato rispetto alla carbon footprint. Nei propri Environmental Sustainability Reports 2024–2025, Microsoft ha introdotto dati sistematici su PUE (Power Usage Effectiveness) e soprattutto su WUE (Water Usage Effectiveness), sottolineando che la quantità d'acqua impiegata per il raffreddamento rappresenta una variabile critica tanto quanto l'energia elettrica consumata. Questo ha permesso di portare l'attenzione non solo sulla carbon footprint ma anche sulla water footprint, evidenziando che la scarsità idrica può rappresentare un vincolo critico quanto le emissioni di CO_2. Sebbene si tratti di dati aziendali, quindi presentati in forma aggregata e non sempre verificabili, la loro pubblicazione mostra come l'attenzione al WUE stia entrando a pieno titolo nell'agenda industriale, in parallelo agli avanzamenti della ricerca accademica. In questo senso, l'attenzione al WUE non è soltanto una scelta metodologica innovativa sul piano scientifico, ma riflette una consapevolezza ormai consolidata anche nelle pratiche aziendali, rafforzando la legittimità di considerare l'acqua come componente essenziale della sostenibilità dei modelli di intelligenza artificiale.

2.3 Gap di ricerca e Obiettivo dello Studio

Dalla letteratura emerge come l'attenzione si sia concentrata prevalentemente sulla fase di addestramento dei modelli, lasciando l'inferenza in secondo piano. Eppure, quest'ultima rappresenta oggi la componente dominante del ciclo di vita degli LLM in termini di consumo energetico ed emissioni. Le analisi disponibili, inoltre, risultano frammentate e spesso basate su singoli casi studio o su stime aggregate, senza fornire strumenti sistematici e replicabili in grado di restituire un quadro comparabile tra modelli e provider.

Un ulteriore limite riguarda la scarsità di benchmark empirici fondati su dati reali: la maggior parte degli studi utilizza stime teoriche o simulazioni, trascurando l'analisi di query effettive. Anche laddove vengono considerati dati operativi, l'attenzione si limita a grandezze aggregate come il consumo complessivo di energia, senza includere variabili micro quali la latenza, il numero di token in input e in output o le strategie di generazione, che costituiscono invece i veri determinanti della sostenibilità.

Altro aspetto critico è l'assenza di una prospettiva geografica e temporale: la letteratura considera quasi esclusivamente medie annuali di carbon intensity, senza tener conto di come il luogo di esecuzione e l'orario di utilizzo possano modificare radicalmente l'impatto ambientale. Questo approccio ignora il fatto che la disponibilità di energia rinnovabile varia in modo significativo sia tra Paesi sia all'interno della stessa giornata, con conseguenze dirette sulla carbon footprint delle inferenze. A ciò si aggiunge la quasi totale trascuratezza della dimensione idrica: il consumo d'acqua necessario al raffreddamento dei data center,

misurato tramite Water Usage Effectiveness (WUE), rimane marginale negli studi, pur rappresentando un vincolo sempre più rilevante in termini di sostenibilità.

Questa ricerca si propone di colmare tali lacune, offrendo un contributo originale su più livelli. In primo luogo, raccoglie dati empirici tramite API pubbliche, per estrarre variabili micro, su un insieme eterogeneo di modelli di Generative AI, consentendo un'analisi comparativa fondata su osservazioni reali. In secondo luogo, tratta la latenza non come semplice input tecnico, ma come variabile di output e indicatore chiave dell'efficienza, stimata in funzione di fattori architetturali (Dense vs MoE), funzionali (reasoning, multimodalità, chain-of-thought), hardware (GPU e configurazioni) e operativi (token in input e output).

L'analisi si estende inoltre alla dimensione spaziale e temporale, evidenziando come la localizzazione dei data center e la variabilità oraria della carbon intensity possano amplificare o attenuare l'impatto ambientale delle inferenze. Infine, la ricerca integra un approccio multidimensionale che, oltre a collegare consumi energetici ed emissioni di CO_2, considera anche l'utilizzo idrico (WUE) legato al raffreddamento dei data center, un elemento spesso trascurato in letteratura ma determinante per una valutazione complessiva della sostenibilità.

In sintesi, manca ancora un quadro analitico che colleghi le caratteristiche tecniche dei modelli alle condizioni infrastrutturali e geografiche in cui vengono eseguiti, nonostante i principali provider cloud offrano già strumenti per orientare queste scelte. Questo vuoto appare particolarmente rilevante in un contesto in cui le imprese, spinte da normative come la CSRD e la EED, sono chiamate a integrare nei propri bilanci di sostenibilità informazioni concrete sulle emissioni digitali.

- *Research Question*

L'obiettivo di questa ricerca è fornire un quadro empirico che consenta decisioni più consapevoli, rispondendo alla seguente domanda di ricerca: "*Quali fattori influenzano l'impatto ambientale – in termini di consumo energetico, consumo idrico ed emissioni di CO_2 – dei modelli di Generative AI nella fase di inferenza?"* In particolare, si indaga come le variabili architetturali, operative e infrastrutturali incidano sulla latenza e, di conseguenza, sull'impatto ambientale dei modelli di Intelligenza Artificiale Generativa.

Per rispondere in modo rigoroso alla domanda di ricerca, il lavoro si è tradotto in un'analisi empirica volta a quantificare l'effetto delle variabili architetturali, operative e infrastrutturali sui consumi energetici, idrici e sulle emissioni di CO_2 dei modelli di Generative AI nella fase di inferenza.
Il capitolo che segue descrive la metodologia adottata, illustrando il percorso di raccolta e analisi dei dati che ne ha reso possibile l'analisi dettagliata.

Capitolo 3 – Metodologia e Disegno Sperimentale

3.1 Introduzione metodologica

L'obiettivo del lavoro è misurare, su un insieme di modelli generativi di Intelligenza Artificiale, accessibili tramite API pubbliche, una serie di indicatori primari di esecuzione: latenza end-to-end, numero di token in input e in output, velocità di generazione. Tali misure sono state aggregate per combinazione modello × tipologia di prompt (breve, medio, lungo), con lo scopo di costruire un dataset solido e comparabile, base per successive analisi di efficienza energetica e impatto ambientale.

Per raccogliere dati comparabili e robusti, è stato adottato un disegno sperimentale strutturato. Ogni modello è stato interrogato su tre diverse tipologie di prompt (approfondimento nel capitolo successivo), con più run replicati in più giornate in differenti fascie orarie, per attenuare la variabilità legata al carico delle piattaforme cloud. Prima dei run validi è stata sempre eseguita una chiamata di warm-up, utile a stabilizzare connessioni e caching lato provider: i risultati di questa fase sono stati scartati e non confluiscono nell'analisi.

Per assicurare che i dati raccolti fossero affidabili e comparabili, si è posta particolare attenzione a una serie di accorgimenti metodologici. Un elemento centrale è stato l'utilizzo della **randomizzazione parziale dell'ordine dei run**: in altre parole, non si è seguito un ordine fisso nella sequenza di esecuzione delle prove per ciascun modello e prompt, ma si è introdotto un criterio di rotazione. Questo accorgimento ha permesso di ridurre il rischio che fattori esterni (come picchi di traffico, congestioni momentanee dei server o variazioni nell'allocazione delle risorse cloud) influenzassero sistematicamente un determinato modello o tipologia di prompt. In tal modo, la variabilità è stata distribuita in maniera più equa e le statistiche finali risultano meno esposte a bias temporali.
A completare questa scelta, tra un run e l'altro è stata introdotta una pausa di circa un secondo. Tale intervallo, apparentemente minimo, ha avuto un duplice scopo: da un lato mitigare gli effetti dei meccanismi di rate limit imposti dai provider, dall'altro ridurre l'impatto di eventuali caching lato server, che avrebbe potuto accelerare artificialmente le risposte. Infine, tutti i run che presentavano anomalie (come timeout, output incompleto o errori di connessione) sono stati esclusi e ripetuti. Questa decisione ha garantito che le statistiche finali riflettessero soltanto comportamenti stabili e rappresentativi dei modelli analizzati.

Il campione preso in analisi comprende circa 2400 osservazioni: 10 run per prompt (breve, medio, lungo), eseguiti in 5 giornate differenti per 16 modelli analizzati.

Accanto agli accorgimenti metodologici, è stato necessario definire un insieme di parametri operativi comuni, pensati per assicurare omogeneità e confrontabilità tra modelli e scenari di test. Tali parametri riguardano sia le piattaforme di accesso ai modelli sia le modalità di esecuzione delle prove:

- **API considerate**: OpenAI (GPT-5, GPT-5-mini, GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, GPT-4o, o3, GPT-3.5-turbo), Anthropic (Claude-3.7 Sonnet), DeepSeek (DeepSeek-R1), Together AI (hosting per LLaMA-3.2-90B Vision Instruct, Mistral-7B, Mixtral-8×7B), Mistral API (endpoint nativo).
- **Tipologia di prompt (appronfondimento nel capitolo successivo)**: breve (≈300 token in output), medio (≈1000), lungo (≈1500).
- **Ripetizioni**: 5 run per ogni combinazione modello × prompt, in modo da ottenere statistiche descrittive robuste.
- **Warm-up**: 1 run preliminare escluso dall'analisi, utilizzato solo per stabilizzare la connessione.
- **Output**: raccolta dei dati grezzi e successiva elaborazione in dataset strutturato, con calcolo di medie e deviazioni standard per ciascun modello e scenario.

Il dataset così costruito consente non solo di confrontare le prestazioni dei modelli, ma anche di collegare i risultati sperimentali a fattori ambientali e infrastrutturali.

I dati così raccolti rappresentano il primo passo per applicare i moltiplicatori ambientali – Power Usage Effectiveness (PUE), Water Usage Effectiveness (WUE) e Carbon Intensity Factor (CIF). L'integrazione di questi parametri consente di tradurre misure tecniche, come latenza o numero di token elaborati, in grandezze più significative dal punto di vista della sostenibilità: consumo di energia, utilizzo di acqua per il raffreddamento e quantità di emissioni di CO_2 associate all'inferenza di ciascun modello.

3.2Campione e Fonte Dati

3.2.1 Definizione dei Prompt

Per la misurazione sperimentale delle prestazioni e dei consumi energetici dei modelli di linguaggio, è stato necessario definire un set di prompt standardizzati, in grado di rappresentare tre diversi livelli di complessità e lunghezza: breve, medio e lungo. Questa scelta risponde a due esigenze metodologiche: da un lato, garantire la comparabilità tra modelli differenti, dall'altro, simulare casi d'uso realistici, simili a quelli che emergono nei contesti accademici e professionali.

1. **Prompt breve**. Il primo prompt consiste in una semplice richiesta di traduzione dall'italiano all'inglese: "La LUISS Guido Carli è una delle università italiane più riconosciute a livello internazionale per gli studi in economia, diritto e scienze politiche." In questo caso la lunghezza prevista è di circa 100 token in input e 300 token in output. Il compito è stato scelto per valutare la rapidità del modello e la sua capacità di produrre una risposta concisa e corretta. La traduzione è infatti un'attività linguistica elementare, ma al tempo stesso permette di misurare latenza e throughput in uno scenario di interazione quotidiana con un modello di GenAI.

2. **Prompt medio**. Il secondo prompt è stato progettato per rappresentare un compito di maggiore complessità, tipico delle attività accademiche e manageriali. Al modello è stato chiesto: “Spiega in modo approfondito le differenze tra algoritmi rule-based, machine learning e deep learning. Fornisci esempi concreti di applicazioni, evidenziando vantaggi e limiti di ciascun approccio. Concludi con un confronto che illustri come queste tecniche vengano applicate oggi in contesti accademici e aziendali.” La lunghezza prevista è di circa 1.000 token in input e 1.000 token in output. Tale compito permette di valutare la capacità del modello di produrre un testo articolato, ben strutturato e coerente, affrontando temi che combinano aspetti teorici e applicativi. Inoltre, la distinzione tra algoritmi rule-based, ML e DL rappresenta un classico esempio di comparazione concettuale.

3. **Prompt lungo**, per simulare l’utilizzo in contesti di ricerca avanzata, è stato definito un terzo prompt basato su un testo accademico esteso. È stato scelto come documento di riferimento l’articolo di Utterback e Abernathy (1975), *A dynamic model of process and product innovation*, considerato un classico della letteratura sul management dell’innovazione. Il compito assegnato ai modelli consisteva nel produrre un riassunto critico di circa 1.500 token in output, a partire da un input di circa 10.000 token. In particolare, si richiedeva di evidenziare: (i) la tesi principale, (ii) le fonti utilizzate, (iii) i punti di forza, (iv) i limiti metodologici e (v) possibili sviluppi futuri della ricerca. La scelta di questo paper risponde a un duplice criterio: da un lato, si tratta di un contributo breve ma denso di contenuti, con oltre 9.000 citazioni in letteratura, che costituisce un terreno di prova ideale per la sintesi critica; dall’altro, la sua natura teorica e concettuale permette di testare la capacità dei modelli di identificare argomentazioni complesse, limiti metodologici e prospettive future.

3.2.2 Indicatori Primari

Il punto di partenza dell'elaborazione dei dati raccolti attraverso i run sperimentali è costituito dalle metriche fornite direttamente dalle API durante l'esecuzione dei test: il numero di token in input, i token prodotti in output, la latenza (espressa in millisecondi) e la velocità di generazione (token per secondo). Questi parametri descrivono le prestazioni operative del modello e rappresentano la base su cui vengono calcolati i consumi energetici e le emissioni.

A ciascun modello viene poi associato un valore di potenza tipico dell'hardware su cui si presume venga eseguita l'inferenza. Tali valori derivano dalla letteratura tecnica e dalle specifiche dei produttori: ad esempio, per le GPU NVIDIA A100 si assume un assorbimento di circa 400 watt, mentre per le H100 il valore di riferimento è pari a 700 watt. Questo passaggio consente di stimare il consumo energetico dell'elaborazione.

L'energia consumata a livello IT viene calcolata secondo la formula:

$$Energia_{IT}(Wh) = Potenza_{GPU}(W) \times \frac{Latenza(ms)}{1000 \times 3600}$$

La latenza viene convertita dapprima in secondi e successivamente in ore, così da ottenere un consumo espresso in wattora. Si tratta del fabbisogno energetico strettamente legato all'attività della GPU durante l'inferenza.

Per risalire all'energia effettivamente utilizzata a livello di data center, si applica il coefficiente PUE (Power Usage Effectiveness), che tiene conto dei consumi aggiuntivi per raffreddamento, alimentazione ausiliaria e infrastruttura di supporto. La formula diventa quindi:

$$Energia_{Totale}(Wh) = Energia_{IT}(Wh) \times PUE$$

A partire da tale valore, si stimano le emissioni climalteranti mediante i fattori di emissione regionali. L'equazione utilizzata è:

$$Emissioni(gCO_2e) = \left(\frac{Energia_{Totale}(Wh)}{1000}\right) \times EF$$

dove EF (Emission Factor) rappresenta la quantità media di anidride carbonica equivalente emessa per la produzione di 1 kWh di elettricità in una determinata area geografica.

Parallelamente, il consumo idrico associato al funzionamento dei data center viene stimato sulla base dell'indice WUE (Water Usage Effectiveness), espresso in litri per kWh. La formula adottata è:

$$Consumo\ H_2O(mL) = Energia_{Totale}(kWh) \times WUE \times 1000$$

in cui il fattore moltiplicativo 1000 converte i litri in millilitri, al fine di rendere più immediato il confronto tra esperimenti di diversa durata.

Infine, per consentire l'analisi comparativa tra modelli e scenari di output di diversa lunghezza, si procede alla normalizzazione dei valori di consumo energetico, emissioni e uso idrico per 100 token generati. In questo modo, gli indicatori risultano indipendenti dalla dimensione assoluta della risposta prodotta dal modello, permettendo una valutazione equa delle efficienze relative.

3.2.3 Modelli Selezionati

Lo studio ha preso in considerazione un insieme eterogeneo di modelli di linguaggio di grandi dimensioni, comprendente soluzioni proprietarie (OpenAI, Anthropic, DeepSeek) e open-source (Meta, Mistral), tutte accessibili tramite API pubbliche o piattaforme di hosting gestito (Together AI).

Prima di presentare i singoli modelli, è utile descrivere brevemente i **principali provider** che hanno reso possibile l'analisi:

- **OpenAI**: provider leader di mercato, autore della serie GPT. Rappresenta il riferimento commerciale per accuratezza e diffusione, con modelli che spaziano da soluzioni compatte (mini, nano) fino a sistemi multimodali di ultima generazione (GPT-4o e GPT-5).
- **Anthropic**: start-up specializzata in modelli sicuri ed eco-efficienti, focalizzata sulla famiglia Claude, che offre contesti di input molto estesi e ottimizzazioni di consumo.
- **DeepSeek**: realtà emergente cinese, nota per modelli di grande scala come R1. La sua inclusione consente di rappresentare scenari a elevata intensità energetica.
- **Meta**: con la famiglia LLaMA ha reso disponibili modelli open-source di grande rilevanza per la ricerca, offrendo la possibilità di confronto diretto con le soluzioni proprietarie.
- **Mistral AI**: start-up europea che si è distinta per l'attenzione all'efficienza computazionale e per l'introduzione di architetture innovative, dai modelli compatti (7B) alle versioni mixture-of-experts (8×7B).

- **Together AI**: piattaforma di hosting gestito che ha reso disponibili tramite API i modelli open-source, assicurando uniformità di logging e condizioni di test simili a quelle dei provider proprietari.

A partire da questo contesto, nelle sezioni seguenti vengono presentati i **modelli selezionati**, ciascuno accompagnato da una descrizione sintetica delle principali caratteristiche e delle motivazioni che ne hanno giustificato l'inclusione nello studio.

- GPT-5 e GPT-5-mini *(OpenAI, 2025)*

Rappresentano la frontiera più avanzata dell'offerta OpenAI e dello stato dell'arte nel settore. GPT-5 è stato incluso come riferimento più recente per accuratezza e diffusione, mentre la variante "mini" consente di esplorare i compromessi tra prestazioni e consumo, offrendo un punto di confronto essenziale per capire l'efficienza di modelli più leggeri.

- GPT-4.1, GPT-4.1-mini e GPT-4.1-nano *(OpenAI, 2024–2025)*

Questi modelli permettono di analizzare la scalabilità interna a una stessa famiglia. La versione standard rappresenta il modello di riferimento.

- GPT-4o *(OpenAI, 2024)*

È stato selezionato come modello multimodale di riferimento, in grado di gestire testo, immagini, audio e video. La sua inclusione si giustifica per la centralità che ha assunto nell'offerta OpenAI e per la vasta diffusione presso gli utenti, elementi che lo rendono un benchmark imprescindibile.

- GPT-3.5-turbo *(OpenAI, 2022)*

Scelto come baseline storica, permette di osservare il divario evolutivo rispetto alle generazioni più recenti. La sua inclusione rende evidente come si siano modificati negli anni i rapporti tra prestazioni, costi computazionali e impatti ambientali.

- o3 *(OpenAI, 2025)*

Questo modello di ragionamento avanzato è stato incluso perché rappresenta uno scenario ad alta intensità energetica, utile per valutare il comportamento dei modelli più potenti e complessi.

- Claude-3.7 Sonnet *(Anthropic, 2025)*

Claude-3.7 Sonnet si distingue per l'attenzione all'efficienza e alla sicurezza, oltre che per la capacità di gestire contesti di input molto ampi (200k token). Rappresenta quindi un modello complementare che permette di testare approcci diversi a livello di architettura e filosofia di sviluppo.

- DeepSeek-R1 *(DeepSeek, 2025)*

È stato scelto come esempio di scenario ad alta intensità computazionale. Pur essendo meno efficiente di altri modelli, è utile per evidenziare come i consumi possano variare in funzione dell'architettura (Mixture fo Experts), ma anche delle infrastrutture di deployment, in questo caso basate su GPU H800 in Cina.

- LLaMA-3.2-90B *Vision Instruct (Meta, 2024)*

Rappresenta il più grande modello open-source incluso nell'analisi. La sua scelta permette di osservare le prestazioni e i consumi di un modello open molto grande.

- Mistral-7B *(Mistral AI, 2023)*

È un modello compatto, noto per l'efficienza e la rapidità di risposta. È stato selezionato per evidenziare i vantaggi delle architetture leggere e ottimizzate, che pur con risorse ridotte riescono a garantire prestazioni competitive.

- Mixtral-8×7B *(Mistral AI, 2024)*

Questo modello introduce un'architettura mixture-of-experts (MoE), che attiva solo una parte dei suoi parametri a ogni inferenza. È stato incluso perché rappresenta un approccio innovativo in termini di scalabilità ed efficienza, consentendo di confrontare strategie architetturali differenti.

Modello	**Provider**	**Tipo**	**Anno**	**Modalità di generazione**	**HW inferenza**	**Potenza stimata (W)**
GPT-5	OpenAI	Proprietario	2025	Multimodale (testo, immagini, audio, codice)	NVIDIA H100/H200 (cluster 8+ GPU, Azure)	~700 W × GPU (≈5.6–7 kW tot.)
GPT-5-mini	OpenAI	Proprietario	2025	Testo + multimodale ridotto	NVIDIA A100/H200 (1–2 GPU)	~400–700 W × GPU
GPT-4.1	OpenAI	Proprietario	2024	Multimodale standard	NVIDIA H100 (4–8 GPU)	~700 W × GPU (≈2.8–5.6 kW)
GPT-4.1-mini	OpenAI	Proprietario	2024	Multimodale light	NVIDIA A100 (1–2 GPU)	~400 W × GPU (≈400–800 W)

Modello	Provider	Tipo	Anno	Modalità di generazione	HW inferenza	Potenza stimata (W)
GPT-4.1-nano	OpenAI	Proprietario	2024	Testo	NVIDIA L4 / A10G (1 GPU)	150–250 W
GPT-4o	OpenAI	Proprietario	2024	Multimodale avanzato (testo, immagini, audio, video)	NVIDIA H100 (8+ GPU)	~700 W × GPU (≈5.6 kW+)
o3	OpenAI	Proprietario	2025	Testo e ragionamento avanzato	NVIDIA H100/H200 (8+ GPU)	~700–1000 W × GPU
GPT-3.5-turbo	OpenAI	Proprietario	2022	Testo	NVIDIA A100 (1–4 GPU)	~400 W × GPU (≈400–1600 W)
Claude-3.7 Sonnet	Anthropic	Proprietario	2025	Testo, contesti estesi (200k input)	NVIDIA A100/H100 (1 GPU)	400–700 W
DeepSeek-R1	DeepSeek	Proprietario	2025	Testo e ragionamento	NVIDIA H800 (cluster 4–8 GPU, Cina)	~700–1000 W × GPU
LLaMA-3.2-90B Vision Inst	Meta	Open-source	2024	Multimodale vision + testo	NVIDIA H100/H200 (8 GPU, AWS/Together)	~700 W × GPU (≈5.6 kW)
Mistral-7B	Mistral	Open-source	2023	Testo	NVIDIA A100 (1 GPU) o L4 (1 GPU)	200–400 W
Mixtral-8×7B	Mistral	Open-source	2024	Testo con mixture-of-experts	NVIDIA A100 (2 GPU, 80 GB)	~400 W × GPU (≈800 W)

La selezione dei modelli ha seguito un criterio di rappresentatività: includere modelli di diverse taglie e generazioni, con architetture e obiettivi differenti, così da avere un confronto bilanciato tra accuratezza, efficienza ed eco-sostenibilità. Alcuni modelli (GPT-4o, GPT-5) rappresentano lo stato dell'arte; altri (GPT-3.5-turbo) fungono da baseline storica. Alcuni modelli, come DeepSeek-R1 e o3, sono stati inseriti perché permettono di esplorare il comportamento dei sistemi in scenari di inferenza particolarmente intensivi, offrendo un termine di paragone fondamentale per valutare le differenze di eco-efficienza

3.2.4 Caratteristiche Modelli di Generative AI e Provider

Nell'analisi dei modelli proprietari accessibili esclusivamente tramite API, uno dei principali ostacoli metodologici è l'assenza di informazioni dirette sulle risorse computazionali effettivamente impiegate dai provider. OpenAI, Anthropic, DeepSeek, ecc infatti, non forniscono dettagli pubblici né sul numero di GPU attivate per query né sulla loro efficienza operativa.

Per superare questa problematica, è stato adottato un approccio infrastruttura-aware, che associa ciascun modello ad una "classe hardware" tipica (A100, H100, TPUv4) in base alla complessità e alla scala del modello. La stessa strategia è stata utilizzata in report di settore, come *MLPerf Power Benchmark* (MLCommons, 2023)[22] e nelle schede tecniche dei principali produttori di GPU (NVIDIA, 2023–2024)[23].

3.2.5 Classificazione hardware dei modelli

I modelli sono stati classificati in tre macro-categorie di dimensione:

- Piccoli (≤10 miliardi di parametri, uso ottimizzato e contesti ristretti), tipicamente eseguiti su GPU NVIDIA A100 con un consumo medio di circa 300 W.
- Medi (10–20 miliardi di parametri, o equivalenti proprietari), associati a GPU A100 da 80 GB o configurazioni analoghe, con potenza stimata di circa 400 W.
- Grandi (≥70 miliardi di parametri o equivalenti multimodali proprietari), eseguiti su NVIDIA H100 o TPUv4, con potenze comprese tra 600–700 W.

Un aspetto metodologico cruciale di questo lavoro riguarda la **stima dell'hardware associato a ciascun modello di linguaggio di grandi dimensioni**. Infatti, non tutti i modelli forniscono informazioni trasparenti e pubbliche circa l'infrastruttura computazionale utilizzata in fase di inferenza, e questo rende necessaria una procedura di attribuzione indiretta, basata su evidenze empiriche e fonti secondarie.

Nel caso dei **modelli open-source** (ad esempio la famiglia *LLaMA* di Meta o *Mistral*), il numero di parametri e l'architettura sono resi pubblici, e la comunità scientifica dispone di benchmark dettagliati sulle performance e sui requisiti hardware. In questi casi, la stima può essere effettuata in maniera relativamente

[22] MLCommons, 2023. *MLPerf Inference delivers power efficiency and performance gains (v3.0 results)*. MLCommons.
MLCommons, n.d. *Power Working Group*. MLCommons.

[23] NVIDIA Corporation, 2023. *NVIDIA H100 Tensor Core GPU — Datasheet*. NVIDIA Corporation.
NVIDIA Corporation, 2024, *NVIDIA A100 Tensor Core GPU — Datasheet*. NVIDIA Corporation.
NVIDIA Corporation, 2020. *NVIDIA DGX A100 — Data Sheet*. NVIDIA Corporation.
Google Cloud (n.d.) *TPU v4 — Documentazione*. Google Cloud.

precisa, facendo riferimento a documentazione tecnica e a test riproducibili condotti dalla comunità di ricerca (Meta AI, 2023; Mistral AI, 2023).[24]

Diversa è la situazione dei **modelli proprietari,** come GPT-4o, Claude o DeepSeek, i quali vengono resi disponibili esclusivamente tramite API e senza dettagli ufficiali circa l'hardware utilizzato.

3.2.5.1 Modelli Open - Source

Uno degli elementi distintivi dei modelli open-source è la disponibilità pubblica dei dati fondamentali che li descrivono: numero di parametri, architettura di riferimento, dimensioni dei pesi e caratteristiche principali per l'inferenza.

Le informazioni su parametri e architettura sono rese disponibili principalmente attraverso:

- **articoli scientifici su arXiv**, che descrivono i modelli in maniera sistematica e citano i dati fondamentali.
- **model cards e repository ufficiali su Hugging Face**, dove vengono caricati i pesi, i tokenizer e le istruzioni per l'uso.

Per la famiglia **LLaMA-2**, Meta ha rilasciato nel luglio 2023 il paper *LLaMA 2: Open Foundation and Fine-Tuned Chat Models*[25], in cui sono esplicitati i tre modelli da 7, 13 e 70 miliardi di parametri. Nel documento vengono descritti i dataset di addestramento, le modifiche architetturali e vengono indicati i requisiti di memoria associati alle varie taglie. Parallelamente, i pesi sono stati resi disponibili su Hugging Face, rendendo possibile l'uso diretto dei modelli in contesti accademici e industriali.

Analogamente, Mistral AI ha pubblicato il proprio modello **Mistral-7B** nel settembre 2023, corredato da un paper tecnico[26] che ne descrive dimensione (7,3 miliardi di parametri) e innovazioni architetturali pensate per migliorare l'efficienza, come la sliding-window attention. Anche in questo caso i pesi sono stati caricati su Hugging Face con licenza Apache-2.0, favorendo l'adozione e la sperimentazione da parte della comunità.

Infine, nel dicembre 2023, sempre Mistral AI ha introdotto **Mixtral-8×7B:** si tratta di uno dei primi modelli open a implementare un'architettura sparse mixture-of-experts, con otto esperti da 7 miliardi ciascuno per layer, ma con soli due attivati per token.

---

[24] Meta AI, 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models.* Meta AI.
Mistral AI, 2023. *Mistral 7B*. Mistral AI.

[25] Touvron, H. *et al.*, 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv:2307.09288.

[26] Jiang, A.Q. *et al.*, 2023. *Mistral 7B*. arXiv:2310.06825.

Più recentemente, benchmark pubblici come quelli di MLCommons (*MLPerf Inference v4.0 Results*, 2024) hanno incluso modelli LLaMA in configurazioni standardizzate (8× H100), offrendo dati diretti e verificabili sui consumi.

Caratteristiche principali dei modelli

Pur senza entrare nei dettagli tecnici, è utile ricordare le caratteristiche fondamentali dei tre modelli considerati.

- **LLaMA-2 (Meta, 2023):** disponibile in tre taglie (7B, 13B, 70B), è un modello denso della famiglia transformer, progettato per essere versatile e facilmente adattabile. La versione da 7B è leggera e compatibile con una singola GPU A100, la 13B rappresenta un compromesso tra accuratezza e costi computazionali, mentre la 70B raggiunge performance molto elevate, ma richiede configurazioni multi-GPU.
- **Mistral-7B (Mistral AI, 2023):** modello da 7,3 miliardi di parametri, ottimizzato per l'efficienza di inferenza. Introdotto con licenza Apache-2.0, integra meccanismi come la sliding-window attention, che consente di gestire sequenze più lunghe con un costo computazionale contenuto. È stato progettato per funzionare anche su GPU più economiche, favorendo l'adozione diffusa.
- **Mixtral-8×7B (Mistral AI, 2024):** modello basato su un'architettura mixture-of-experts, che combina la capacità teorica di un modello molto grande (46,7 miliardi di parametri) con un costo computazionale per token equivalente a circa 12,9 miliardi di parametri.

3.2.5.2 Modelli Proprietari

Per questi modelli, la stima dell'hardware è stata effettuata basandosi su tre criteri principali:

1. **Classificazione dimensionale**: i modelli sono stati suddivisi in "piccoli", "medi" e "grandi" a seconda della scala parametrica stimata, della collocazione commerciale e del posizionamento competitivo. Ad esempio, GPT-4.1 nano è stato collocato nella categoria "piccoli" poiché progettato per applicazioni leggere e a basso consumo (OpenAI, 2024), mentre modelli come GPT-5 o DeepSeek-R1 sono stati attribuiti alla classe "grande" in virtù delle elevate esigenze computazionali.
2. **Associazione a GPU e infrastrutture tipiche**: si è fatto riferimento ai datasheet ufficiali NVIDIA (2023; 2024) e a studi sul consumo energetico dei data center [27]. In particolare, ai modelli più piccoli sono state associate GPU di fascia A100 (400 W di TDP), ai modelli medi GPU H100 (700 W TDP),

[27] Shehabi et al., 2016 – LBNL Lawrence Berkeley National Laboratory, "United States Data Center Energy Usage Report Patterson et al., 2021 – "Carbon Emissions and Large Neural Network Training".

mentre per i modelli più grandi e sperimentali si è fatto riferimento a configurazioni con H800 (utilizzate in Cina, secondo DeepSeek, 2025) o TPUv4 (Google, 2023). Questo approccio è coerente con i dati riportati da MLCommons nei benchmark MLPerf Power *(MLCommons, 2024)*.

3. **Convalida tramite evidenze sperimentali**: i risultati dei nostri test black-box (latenza, throughput e token generati per secondo) sono stati confrontati con report di sostenibilità pubblicati dai principali hyperscaler (Google Environmental Report 2023; Microsoft Sustainability Report 2023). Tale triangolazione consente di ridurre il margine di errore nella stima e di assicurare coerenza interna tra fonti accademiche, documenti industriali e osservazioni empiriche.

Questa scelta metodologica è necessaria perché l'impatto ambientale di un modello non dipende unicamente dalla sua efficienza algoritmica, ma anche dalle caratteristiche dell'hardware su cui viene eseguito. Ne consegue che una stima affidabile del consumo deve necessariamente includere l'hardware, anche laddove non vi siano informazioni ufficiali disponibili.

Modello	Classe dimensionale	Hardware stimato	Potenza stimata (W)	Fonte principale
GPT-5	Grande	NVIDIA H100, 8+ GPU	700 W × GPU	OpenAI (2024); NVIDIA, *H100 Datasheet* (2023);
GPT-5-mini	Medio	NVIDIA A100, 1–2 GPU	400 W × GPU	OpenAI (2024); NVIDIA, *A100 Datasheet* (2020); MLCommons (2024)
GPT-4.1	Grande	NVIDIA H100, 4–8 GPU	700 W × GPU	OpenAI (2024); MLCommons, *Inference v3.0* (2023)
GPT-4.1-mini	Medio	NVIDIA A100, 1–2 GPU	400 W × GPU	NVIDIA, *A100 Datasheet* (2020); OpenAI API docs (2024)
GPT-4.1-nano	Piccolo	NVIDIA L4 / A10G, 1 GPU	150–250 W	Google, *Cloud GPU L4 Datasheet* (2023); OpenAI (2024)
GPT-4o	Grande	NVIDIA H100, 8+ GPU	700 W × GPU	OpenAI, *Product blog* (2024); NVIDIA (2023)
O3	Grande	NVIDIA H100, 8+ GPU	700–1000 W	OpenAI (2025); Deng et al., *How Hungry is AI?* (2025)
GPT-3.5-turbo	Medio	NVIDIA A100, 1–4 GPU	400 W × GPU	OpenAI (2022); MLCommons, *Inference Benchmark* (2022)
Claude-3.7 Sonnet (20250219)	Piccolo	NVIDIA A100 / H100 singola	400–700 W	Anthropic, *Claude 3.7 Model Card* (2025); MLPerf (2024)

Modello	Classe dimensionale	Hardware stimato	Potenza stimata (W)	Fonte principale
Claude-Sonnet-4 (20250514)	Medio	NVIDIA H100, 1–2 GPU	700 W × GPU	Anthropic, *Sustainability Report* (2025); MLCommons (2024)
Claude-Opus-4 (20250514)	Grande	NVIDIA H100, 8 GPU	700 W × GPU	Anthropic, *Technical Whitepaper* (2025)
Claude-Opus-4.1 (20250805)	Grande	NVIDIA H100, 8–16 GPU	700 W × GPU	Anthropic, *Model Update* (2025); Uptime Institute (2024)
DeepSeek-Reasoner	Grande	NVIDIA H800, 2.048 GPU (training); inferenza su cluster ridotti H100/H800	700–1000 W	DeepSeek, *Technical Whitepaper* (2024); MIT Tech Review (2025)
LLaMA-2-7B (Meta, 2023)	Piccolo	NVIDIA A100, 1 GPU	~400 W	Touvron et al., LLaMA 2 (2023), arXiv:2307.09288; Meta AI ai.meta.com/llama; Hugging Face meta-llama
LLaMA-2-13B (Meta, 2023)	Medio	NVIDIA A100, 2 GPU	~800 W	Touvron et al., LLaMA 2 (2023), arXiv:2307.09288; Meta AI ai.meta.com/llama
LLaMA-2-70B (Meta, 2023)	Grande	NVIDIA H100, 8 GPU	~5600 W	Touvron et al., LLaMA 2 (2023); MLCommons, MLPerf Inference v4.0 (2024); Hugging Face meta-llama
Mistral-7B (Mistral AI, 2023)	Piccolo	NVIDIA A100, 1 GPU (opp. L4 24 GB)	400 W (opp. 200 W)	Jiang et al., Mistral 7B (2023), arXiv:2310.06825; Mistral AI mistral.ai; Hugging Face Mistral-7B
Mixtral-8×7B (Mistral AI, 2024)	Medio	NVIDIA A100, 2 GPU (80 GB)	~800 W	Jiang et al., Mixtral of Experts (2024), arXiv:2401.04088; Mistral AI mistral.ai; Hugging Face Mixtral-8x7B

L’adozione di questo schema consente di collegare i dati empirici raccolti tramite i run di inferenza (token, latenza, throughput) con le ipotesi di consumo hardware. In particolare, si potrà stimare il consumo energetico per run utilizzando la formula:

$$Energia\ (kWh) = \frac{Potenza\_GPU \times Numero\_GPU \times Tempo\_sec}{3600}$$

dove il tempo di esecuzione è quello registrato sperimentalmente, mentre potenza e numero di GPU sono derivati dalle stime riportate in tabella.

3.2.6 Selezione e analisi Hosting per singolo Provider

Provider / Modello	Hosting primario dichiarato	Fonte ufficiale / documentazione
OpenAI (GPT-3.5, GPT-4o, GPT-4.1, GPT-5, O3, Nano/Mini varianti)	**Microsoft Azure** – in particolare regioni **US-East (Virginia)** e **West Europe (Paesi Bassi)** per motivi di latenza	OpenAI Blog (2020), partnership con Microsoft; Microsoft Azure Sustainability Docs (2023)
Anthropic (Claude 3.5, 3.7, 4 Sonnet, Opus, Haiku)	**Amazon Web Services (AWS)** – principalmente **us-east-1 (N. Virginia)**; dal 2023 parte del carico anche su **Microsoft Azure**	AWS Blog (2023), Anthropic Model Cards (2023–2025)
DeepSeek (Reasoner, Chat)	**Cluster proprietari in Cina**, alimentati da GPU **NVIDIA H800**. Le regioni operative non sono documentate, ma inferenza stimata in **AP-South (India)** o **AP-East (Hong Kong)** per vicinanza all'utenza asiatica	MIT Tech Review (2024), DeepSeek Whitepaper (2024)
Meta (LLaMA-2 7B/13B/70B – open source)	**Eseguibile ovunque**. Meta distribuisce i pesi ma non fornisce inferenza gestita. In pratica: **white-box locale** o su provider cloud (Azure, GCP, AWS)	Meta AI Blog (2023), Hugging Face Model Cards
Mistral (Mistral-7B, Mixtral)	Simile a Meta: open source, **white-box**. Nessuna infrastruttura di hosting proprietaria dichiarata.	Mistral AI Release Notes (2023)
Microsoft (Phi-3 Mini, ecc.)	**Microsoft Azure** – soprattutto regioni **West Europe (Paesi Bassi)** e **US-East (Virginia)**	Microsoft Build Announcements (2024), Azure Docs

Un aspetto cruciale per la stima degli impatti ambientali dei modelli linguistici riguarda l'individuazione delle infrastrutture di hosting sulle quali tali modelli vengono eseguiti. Ne consegue che, in mancanza di dati dettagliati forniti dai provider, la scelta della regione di riferimento deve basarsi su criteri solidi e citabili.

In questo progetto, la scelta per tale stima è ricaduta nell'associare a ciascun modello, proprio l'**hosting primario dichiarato** dal provider stesso. L'analisi della documentazione pubblica conferma che i principali provider hanno reso noto almeno in termini generali il loro partner infrastrutturale e la regione di riferimento.

OpenAI ha dichiarato ufficialmente fin dal 2020 la propria partnership strategica con **Microsoft Azure**, specificando che i modelli della serie GPT vengono eseguiti su cluster GPU ospitati nei data center Azure di **US-East (Virginia)** e, per l'utenza europea, in **West Europe (Paesi Bassi)**[28]. Analogamente, **Anthropic** ha

[28] OpenAI Blog, 2020;

reso noto nel 2023 di aver stretto una collaborazione con **Amazon Web Services (AWS)**, localizzando gran parte delle proprie attività di inferenza nella regione **us-east-1 (N. Virginia)**, la stessa che rappresenta storicamente il cuore della rete globale di AWS. [29]Più recentemente, Anthropic ha iniziato a distribuire parte dei carichi anche su Azure, ma la documentazione tecnica e i model cards disponibili indicano come primario il legame con AWS (Anthropic, *Claude Model Card*, 2025).

Un caso differente è rappresentato da **DeepSeek**, azienda cinese che ha acquisito notorietà per l'utilizzo massiccio di GPU NVIDIA H800 durante il training dei propri modelli. In assenza di una dichiarazione ufficiale di hosting per l'inferenza, si ricorre a stime basate su fonti giornalistiche e whitepaper tecnici[30]: l'infrastruttura sembra essere collocata prevalentemente in **Cina continentale**, con possibili nodi di distribuzione verso **Hong Kong (AP-East**) o **Mumbai (AP-South)** per servire l'utenza internazionale *(MIT Tech Review, 2024; DeepSeek Whitepaper, 2024).*

Al contrario, modelli **open-source** come **Meta LLaMA-**2 o **Mistral** non prevedono un hosting centralizzato. In tali casi, i pesi vengono messi a disposizione della comunità e l'esecuzione può avvenire localmente o su cloud scelti dall'utente (Azure, AWS, Google Cloud).

Infine, i modelli sviluppati direttamente da Microsoft, come **Phi-3 Mini,** sono nativamente disponibile su Azure. Tra le prime regioni operative per questi servizi figurano East US (Virginia) e West Europe (Paesi Bassi); l'elenco si è ampliato nel tempo con ulteriori regioni europee e nord-americane, come dichiarato nei comunicati ufficiali dell'azienda (Microsoft Built)[31].

In sintesi, la ricostruzione degli hosting primari è un passaggio metodologico necessario per applicare correttamente i moltiplicatori ambientali.

Microsoft, *Azure Sustainability Report*, 2023

[29] Anthropic, *Claude Model Card*, 2025.

[30] MIT Technology Review, 2024. *Articolo su DeepSeek e l'uso di GPU H800. MIT Technology Review.*
DeepSeek AI, 2024. DeepSeek – *Technical Report - DeepSeek-V2, DeepSeek AI.*
DeepSeek AI, 2024. DeepSeek – *Technical Report - DeepSeek-R1, DeepSeek AI.*

[31] Microsoft, 2024a. *Build 2024 – Book of News*. Microsoft.
Microsoft, 2024b. *Introducing Phi-3*. Microsoft.
Microsoft, 2024c. *New models added to the Phi-3 family, available on Microsoft Azure*. Microsoft.
Microsoft, 2024d. *Azure AI Foundry – Model Catalog: Region availability*. Microsoft.
Microsoft Q&A, 2023. *Azure OpenAI Service – Region availability (official thread)*. Microsoft.

3.2.7 Localizzazione geografica dei data center e definizione dei metodi di stima

La stima dei consumi energetici e dell'impatto ambientale dei modelli di intelligenza artificiale non può prescindere dalla considerazione della localizzazione geografica dei data center in cui i modelli stessi vengono eseguiti. Infatti, l'intensità carbonica dell'energia elettrica (CIF – *Carbon Intensity Factor*) e l'efficienza complessiva delle infrastrutture (PUE – *Power Usage Effectiveness*, WUE – *Water Usage Effectiveness*) variano in maniera significativa a seconda della regione geografica, delle politiche energetiche locali e della disponibilità di fonti rinnovabili.

3.2.7.1 Metodo 1 – Hosting primario dichiarato

La prima strategia seguita in questo lavoro consiste nell'assumere, per ogni provider, la localizzazione primaria dei propri data center, cioè quella indicata nei benchmark ufficiali o nei documenti tecnici delle aziende. Ad esempio, i report di MLCommons e gli studi condotti da Google e Microsoft adottano come riferimento i data center statunitensi, spesso collocati in regioni come Iowa, Virginia o Oregon, in cui l'approvvigionamento energetico è largamente monitorato (MLCommons, *Inference v3.0 Results*, 2023).

Per **OpenAI**, la letteratura indica un'infrastruttura prevalentemente basata su cluster **Microsoft Azure** situati negli Stati Uniti. La regione **US-East (Virginia)** rappresenta l'hub primario dichiarato, a cui si affiancano anche altre regioni strategiche come Iowa e Oregon, per motivi di ridondanza e disponibilità energetica (Microsoft, *Sustainability Report*, 2023).

Analogamente, **Anthropic** si appoggia principalmente ad **AWS**, con disponibilità prioritaria in regioni USA come **US-East-1 (N. Virginia)**, indicata come default per gran parte delle API (Anthropic, *Model Card*, 2025).

Per **DeepSeek**, i documenti tecnici e le analisi di mercato confermano che i server sono localizzati in **Cina continentale**, principalmente su cluster di GPU H800/H100 distribuiti in grandi hub tecnologici come Pechino e Shenzhen (DeepSeek, *Technical Whitepaper*, 2024; MIT Technology Review, 2025).

In questo scenario, i valori ambientali utilizzati derivano da database consolidati come l'**IEA (Emission Factors Database, 2024)** e da paper tecnici del **Lawrence Berkeley National Laboratory** per PUE e WUE. Questo approccio risulta fondamentale perché consente di collocare i risultati della sperimentazione in continuità con gli standard internazionali e con i dati riportati nei benchmark accademici.

3.2.7.2 Metodo 2 – Provider e Regione effettiva per i test sperimentali

Parallelamente, si è scelto di adottare un secondo metodo di stima che riflette il contesto effettivo della sperimentazione condotta. Le richieste API sono state lanciate dall'Italia, ed è quindi plausibile che i provider abbiano instradato i carichi verso regioni europee.

Per **OpenAI**, ciò significa con ogni probabilità un routing verso cluster **Azure localizzati in Europa occidentale,** principalmente nei **Paesi Bassi** o in **Irlanda** (Microsoft Azure, 2024)[32]. In particolare, i nodi europei di Azure in West Europe (Paesi Bassi) e North Europe (Irlanda) sono i più utilizzati per servire il traffico API proveniente dall'Europa, come confermato anche da studi indipendenti di misurazione delle latenze.

Per **Anthropic**, che utilizza **AWS**, la regione più probabile è **eu-west-1 (Irlanda),** uno degli hub principali per l'utenza europea. AWS ha inoltre altre sedi europee (come Francoforte e Milano), ma l'Irlanda rimane la più usata per servizi AI grazie alla maggiore disponibilità di GPU di nuova generazione (Amazon Web Services, 2024)[33].

In questo caso, i valori di **PUE, WUE e CIF** sono stati stimati sulla base delle metriche europee: ad esempio, il **CIF medio europeo** pubblicato dall'IEA (2024) e i valori di efficienza idrica riportati dal **LBNL**[34] per data center situati in aree temperate. Questo metodo permette di catturare la specificità del contesto reale in cui sono stati raccolti i dati sperimentali, evitando il rischio di sottostimare o sovrastimare l'impatto rispetto al reale flusso delle richieste.

3.2.7.3 Il caso particolare di DeepSeek

Un discorso a parte va fatto per **DeepSeek**. Al momento, questo provider non dispone di edge server o infrastrutture distribuite in Europa. I suoi servizi vengono erogati principalmente dalla **Cina**, come riportato nei documenti tecnici ufficiali e nelle analisi di settore (DeepSeek, 2024; MIT Technology Review, 2025)[35]. Pertanto, per questo modello non è possibile applicare il Metodo 2: l'unico scenario realistico resta quello del Metodo 1, basato sul provider primario in Asia.

Secondo le fonti, i cluster operativi si trovano in particolare a **Pechin**o e **Shenzhen**, zone in cui si concentrano i poli industriali e scientifici cinesi dedicati all'AI. Questo significa che i valori ambientali stimati per DeepSeek devono essere basati su CIF, PUE e WUE cinesi, che presentano caratteristiche molto

[32] Microsoft Azure, *Regional Infrastructure Overview*, 2024
[33] Amazon Web Services, *Global Infrastructure*, 2024
[34] Shehabi et al., 2016 – LBNL Lawrence Berkeley National Laboratory , "United States Data Center Energy Usage Report
[35] DeepSeek, *Technical Whitepaper*, 2024

diverse rispetto a quelle europee e statunitensi (più alta intensità carbonica, maggior uso di carbone nel mix elettrico).

Questa scelta introduce una distinzione importante: mentre per OpenAI e Anthropic è possibile condurre un confronto fra scenario "globale" e scenario "locale europeo", per DeepSeek tale confronto non è praticabile, il che limita in parte la comparabilità dei risultati.

Secondo fonti accademiche, i principali cluster DeepSeek si trovano a **Pechino e Shenzhen**, all'interno di hub tecnologici supportati dal governo cinese e alimentati da infrastrutture ad alta densità di GPU H800 e H100. Durante il training del modello, DeepSeek ha dichiarato di aver impiegato oltre 2.048 GPU H800 e 2,7 milioni di ore GPU, con un fabbisogno energetico complessivo senza precedenti per il contesto asiatico (DeepSeek, 2024)[36].

Il punto critico riguarda la **Carbon Intensity Factor (CIF),** cioè l'intensità carbonica dell'energia elettrica. La rete elettrica cinese, nonostante i progressi compiuti sul fronte delle rinnovabili, rimane fortemente dipendente dal carbone, che nel 2023 copriva circa il **60% del mix elettrico nazionale** (International Energy Agency, 2024)[37] . Questo si traduce in un CIF stimato fra **550 e 650 gCO_2/kWh,** significativamente più elevato rispetto a quello medio europeo (300–350 gCO_2/kWh) e statunitense (380–450 gCO_2/kWh) (IEA, 2024).[38]

Sul piano dell'efficienza, i data center cinesi mostrano **PUE (Power Usage Effectiveness)** generalmente più alti rispetto agli hyperscaler occidentali. Studi congiunti di LBNL e IEA indicano per la Cina valori tipici compresi fra **1.4 e 1.6**, mentre i principali operatori globali (Google, Microsoft, Amazon) hanno ormai raggiunto valori prossimi a 1.1–1.2 (LBNL, 2023 [39]).

Per quanto riguarda il **WUE (Water Usage Effectiveness),** le stime sono ancora più incerte. Secondo un'analisi condotta da Greenpeace East Asia, i data center cinesi che utilizzano raffreddamento evaporativo possono raggiungere valori compresi fra **0.6 e 0.8 L/kWh**, cioè circa il doppio rispetto agli standard dei data center hyperscaler europei e statunitensi (Greenpeace East Asia, *Thirsty Data: Water Use in China's AI Infrastructure*, 2022, p. 9).[40]

Queste osservazioni portano a delle implicazioni a livello metodologico: l'assenza di dati granulari ufficiali implica che, per DeepSeek, non sia possibile utilizzare le metriche consolidate disponibili per Stati Uniti ed Europa. Nel presente lavoro, sono state quindi adottate **due strategie parallele**:

[36] DeepSeek, *Technical Whitepaper*, 2024, p. 4
[37] International Energy Agency – IEA, *China Energy Outlook*, 2024, p. 29
[38] IEA, *Emission Factors Database*, 2024).
[39] Lawrence Berkeley National Laboratory – LBNL, *Data Center Energy Report*, 2023, p. 17
[40] Greenpeace East Asia, *Thirsty Data: Water Use in China's AI Infrastructure*, 2022, p. 9

1. **Scenario benchmark internazionale:** si assumono i valori medi degli hyperscaler globali (PUE = 1.1, WUE = 0.39 L/kWh, CIF = 400 gCO_2/kWh), come riportato da MLCommons e LBNL, per garantire comparabilità con gli altri modelli analizzati.
2. **Scenario realistico:** si utilizzano le stime derivate dalla letteratura (CIF = 600 gCO_2/kWh, PUE = 1.5, WUE = 0.7 L/kWh), che riflettono meglio la specificità del contesto energetico e infrastrutturale cinese.

Questa duplice impostazione consente di distinguere fra una valutazione "globale" – utile per comparare i risultati con la letteratura internazionale – e una valutazione "locale" più vicina alla realtà operativa di DeepSeek. La differenza fra i due scenari risulta cruciale per discutere le implicazioni ambientali dei modelli AI sviluppati e operati in Cina, in particolare quando confrontati con quelli statunitensi ed europei.

3.2.8 Attribuzione dei fattori ambientali ai modelli open-source

A differenza dei modelli appena visti, per i quali esiste un'infrastruttura di hosting centralizzata dichiarata (Microsoft Azure, AWS, ecc), i modelli open-source vengono distribuiti sotto forma di pesi pubblici accompagnati da documentazione tecnica, senza che sia previsto un servizio ufficiale di inferenza gestita. In altri termini, l'utente che intende impiegare questi modelli è libero di scaricarne i pesi e di eseguirli localmente, su un proprio cluster ad alta performance, oppure su un provider cloud a scelta, come Azure, AWS o Google Cloud.

Questa caratteristica pone un problema metodologico rilevante: non esistendo un hosting primario unico, non vi è un riferimento diretto per l'attribuzione dei valori di PUE (Power Usage Effectiveness), WUE (Water Usage Effectiveness) e CIF (Carbon Intensity Factor).

L'approccio scelto è quello di assegnazione di un provider ipotetico, così da garantire la comparabilità con i modelli proprietari. In questo caso, i modelli open-source vengono attribuiti a una regione cloud coerente: nello scenario globale, l'attribuzione è a Microsoft Azure US-East (Virginia), con l'applicazione dei valori medi di PUE, WUE e CIF relativi agli Stati Uniti. Nello scenario europeo, l'attribuzione è invece ai data center collocati in Europa occidentale, come Azure West Europe (Paesi Bassi) o Google Cloud Europe-West (Paesi Bassi o Irlanda), utilizzando i corrispondenti valori ambientali europei.

Il rilascio della famiglia **LLaMA-2** è avvenuto da parte di Microsoft nel luglio. L'annuncio ufficiale, pubblicato sul **Meta AI Blog** (*Meta AI Blog, 2023*), ha reso esplicito che LLaMA-2 sarebbe stato immediatamente disponibile all'interno del ***Microsoft Azure AI Model Catalog***, come parte di una partnership strutturata tra le due aziende. Questa scelta ha sancito Azure come il provider primario di riferimento per l'adozione del modello in scenari di inferenza gestita, pur senza limitare la natura open del progetto: i pesi sono stati infatti resi liberamente scaricabili tramite Hugging Face, permettendo l'esecuzione

anche su altre piattaforme, tra cui AWS e Google Cloud. In termini metodologici, la presenza di un accordo formale con Microsoft motiva l'attribuzione di LLaMA-2 ad **Azure US-East (Virginia)** nello scenario globale e ad **Azure West Europe (Paesi Bassi/Irlanda)** nello scenario europeo.

Il percorso di **Mistral AI** è invece differente. L'azienda francese, fondata nel 2023, ha adottato una filosofia di distribuzione completamente open, rilasciando i pesi di **Mistral-7B** e successivamente di **Mixtral-8×7B** su Hugging Face e attraverso il proprio sito ufficiale (Mistral AI Blog, 2023; 2024). Questa scelta garantisce un'esecuzione potenzialmente indipendente da qualsiasi provider, sia in ambienti locali che su cloud di terze parti. Tuttavia, a partire dal 2024, Mistral AI ha annunciato collaborazioni con **Amazon Web Services**, integrando i modelli all'interno della piattaforma **SageMaker JumpStart,** e con **Databricks Model Serving,** ampliando così la disponibilità dei propri modelli in ecosistemi cloud già consolidati. Non risulta invece una partnership esclusiva con Azure simile a quella stretta tra Meta e Microsoft. Per questo motivo, sebbene i modelli Mistral rimangano tecnicamente eseguibili ovunque, la vicinanza operativa maggiore è oggi verso l'ecosistema **AWS**, che ha integrato le soluzioni Mistral nelle proprie offerte di machine learning gestito.

In sintesi, **Meta** si colloca chiaramente nell'orbita di **Azure**, grazie a un accordo formale di distribuzione annunciato pubblicamente, mentre **Mistral AI** si avvicina ad **AWS**, pur mantenendo una filosofia open e distribuendo i pesi liberamente. Questa differenza riflette strategie industriali divergenti: da un lato un'integrazione strutturata con un hyperscaler specifico (Meta-Microsoft), dall'altro un modello più distribuito ma che trova in AWS la sua collocazione più naturale.

Questa scelta metodologica ha una duplice giustificazione. Da un lato, consente di mantenere la coerenza tra modelli open-source e modelli proprietari, eliminando il rischio di produrre valori non confrontabili. Dall'altro, riflette la reale flessibilità dei modelli open-source, che possono essere effettivamente eseguiti su qualunque infrastruttura cloud disponibile, senza vincoli tecnologici o legali. In tal senso, l'attribuzione ai provider già considerati per gli altri modelli non è arbitraria, ma rappresenta una decisione di armonizzazione necessaria per permettere un confronto corretto.

3.2.9 Considerazioni metodologiche

L'utilizzo combinato di due metodi – hosting primario e hosting effettivo – risponde a due esigenze complementari. Da un lato, il Metodo di attribuzione dell'hosting primario garantisce la continuità con la letteratura e la confrontabilità dei dati con i principali benchmark accademici e industriali. Dall'altro, il Metodo di attribuzione dell'hosting effettivo riflette più fedelmente il contesto sperimentale specifico di questa ricerca, caratterizzato da richieste effettuate dall'Italia e quindi soggette a instradamento europeo.

Provider / Modello	Metodo 1 – Provider/Regione principale (letteratura/benchmark)	Fonte principale	Metodo 2 – Regione effettiva più probabile (routing europeo)	Fonte principale
OpenAI (GPT-x, O3)	Microsoft Azure US-East (Virginia)	MLCommons (2023); Deng et al., *How Hungry is AI?* (2025)	Microsoft Azure West Europe (Paesi Bassi/Irlanda)	Microsoft Azure, *Regional Infrastructure Overview* (2024)
Anthropic (Claude)	AWS US-East-1 (N. Virginia)	Anthropic, *Model Cards* (2025); AWS Global Infra	AWS eu-west-1 (Irlanda)	AWS, *Global Infrastructure* (2024)
DeepSeek (Reasoner, Chat)	Cina continentale / AP-South, cluster H100/H800 in data center locali	DeepSeek, *Technical Whitepaper* (2024); MIT Tech Review (2025)	Non disponibile (nessun edge europeo)	–
Meta (LLaMA-x)	Nessun hosting centralizzato. Attribuzione metodologica: Azure US-East (Virginia), coerente con la partnership ufficiale Azure AI Model Catalog	Touvron et al., LLaMA 2 (2023), arXiv:2307.09288; Meta AI Blog (2023); Microsoft Azure Blog (2023)	Attribuzione metodologica: Azure West Europe (Paesi Bassi)	Azure, Regional Infrastructure Overview (2024); Google Cloud, Europe Regions (2024)
Mistral	Nessun hosting centralizzato. Attribuzione metodologica: US-East (Virginia), in linea con i deployment cloud più diffusi.	Jiang et al., Mistral 7B (2023); Jiang et al., Mixtral of Experts (2024); Mistral AI Blog (2023–2024)	Attribuzione metodologica: AWS eu-west-1 (Irlanda), coerente con routing europeo	Azure, Regional Infrastructure Overview (2024); AWS, Global Infrastructure (2024)

3.2.10 Limiti di certezza nella localizzazione dei provider

Un aspetto cruciale da sottolineare è che non esiste alcuna certezza assoluta sulla localizzazione esatta dei data center a cui vengono instradate le richieste dei modelli di intelligenza artificiale. OpenAI, Anthropic e DeepSeek non rendono pubblica la mappa completa delle infrastrutture utilizzate per l'erogazione dei loro modelli. Ciò che è possibile ricostruire si basa su una combinazione di **dichiarazioni ufficiali dei provider**, documentazione tecnica relativa alle piattaforme cloud su cui essi si appoggiano.

3.3 Costruzione degli indici di sostenibilità: Impatto ambientale (energia, CO_2, acqua) dei modelli

Nella sezione che segue vengono presentate le analisi volte a stimare quali tra i 16 modelli di linguaggio di grandi dimensioni abbiano avuto il maggiore impatto ambientale in termini di energia consumata, CO_2 emessa e consumo idrico.

Poiché non sono disponibili dati ufficiali sui data center e sulle infrastrutture di inferenza, le stime si basano, come deascritto precedentemente, su ipotesi di localizzazione geografica e di hardware (cluster GPU) e sull'applicazione di fattori di conversione riconosciuti:

- PUE – Power Usage Effectiveness, che esprime il rapporto tra energia totale consumata da un data center e quella effettivamente utilizzata per il calcolo;
- WUE – Water Usage Effectiveness, che misura i litri d'acqua impiegati per il raffreddamento per ogni kWh;
- CIF – Carbon Intensity Factor, che quantifica i grammi di CO_2 emessi per kWh in funzione del mix energetico locale.

Questa sezione non risponde direttamente alla domanda di ricerca principale – *quali fattori influenzano l'impatto ambientale dei modelli di generative AI in fase di inferenza?* – che viene affrontata in modo completo nei capitoli successivi.
Questa sezione offre un estratto sintetico, utile a comprendere l'impatto medio delle query sperimentali effettuate sui 16 modelli analizzati, evidenziando gli ordini di grandezza delle risorse coinvolte e fornendo il quadro empirico su cui poggiano le discussioni successive.
Nella sezione Appendice è presente un approfondimento sull'impatto ambientale dei 16 modelli analizzati, con dettagli, metodologie, formule utilizzate e risultati integrali.

In queste analisi, sono stati analizzati, come anticipato, due scenari geografici:

- Scenario europeo: instradamento delle query verso data center situati principalmente nei Paesi Bassi e in Irlanda.,caratterizzati da mix elettrici con basse emissioni medie e da PUE prossimi a 1,2.
- Scenario globale: include anche i principali hub statunitensi (ad es. Virginia e Iowa), con un Carbon Intensity Factor leggermente superiore.

Per ogni scenario sono stati costruiti tre ranking paralleli:

- Energia (Wh/100 tok): Wh consumati per 100 token generati.
- Emissioni di CO_2 (gCO_2e/100 tok) : gCO_2e rilasciati per 100 token:
- Consumo idrico (mL/100 tok): mL d'acqua utilizzati per il raffreddamento per 100 token.

I valori sono derivati combinando i dati sperimentali (latenza, token, throughput) con le ipotesi di potenza hardware (da 1 GPU A100 ≈400 W per i modelli più piccoli fino a cluster 8+ GPU H100 ≈5,6 kW per i più grandi) e con i fattori PUE, WUE e CIF derivati dalla letteratura.

3.3.1 Ranking energetico, emissivo e idrico

Nello scenario europeo, i modelli che si sono dimostrati più efficienti sono quelli di piccola taglia e ottimizzati: GPT-4.1 nano (0,09 Wh/100 tok), Mistral-7B (0,14) e GPT-3.5-turbo (0,20).

Seguono varianti intermedie come GPT-4.1 mini e Claude Sonnet (0,36–0,47), quindi GPT-5 mini (0,51). La fascia alta comprende GPT-5 e GPT-4.1 (≈3 Wh/100 tok), GPT-4o (6,86), LLaMA-3.2-90B Vision Instruct (7,46), Claude-Opus-4 e Claude-Opus-4.1 (≈9–9,6) e, all'estremo, DeepSeek-Reasoner (11,08). Andamento analogo si riscontra per CO_2 e acqua, confermando che la scala del modello e la configurazione hardware sono i principali driver di impatto.

Nello Scenario globale, includendo cluster statunitensi, i consumi e le emissioni crescono leggermente per molti modelli, a causa di un CIF medio più elevato negli USA rispetto ai valori europei.

Un'analisi separata è stata condotta per DeepSeek, con cui si stima per la Cina un CIF ≈600 gCO_2/kWh, PUE ≈1.5 e WUE ≈0.7 L/kWh, con un impatto fino al 50 % superiore rispetto allo scenario globale standard, facendo di DeepSeek-Reasoner il modello più impattante dell'intero campione.

3.3.2 Indicatori compositi e sintetici

Per fornire una **visione integrata** delle tre dimensioni ambientali considerate – **consumo energetico (Wh/100 token)**, **emissioni di CO_2 (gCO_2e/100 token)** e **consumo idrico (mL/100 token)** – è stato sviluppato un **Indice di Sintesi "Impatto Ambientale"**.
L'idea alla base dell'indice è quindi quella di integrare le tre dimensioni in un'unica misura che ne riassuma l'impatto complessivo, permettendo di riconoscere rapidamente i modelli più efficienti da quelli con maggiore impronta ambientale.

Il processo di calcolo si sviluppa in tre passaggi logici.

1. Normalizzazione min–max: Anzitutto ogni indicatore elementare – energia, CO_2 e acqua – è stato trasformato su una scala comune 0–1, così che valori molto diversi potessero essere confrontati tra loro. Questa procedura, nota come min–max normalization, è ampiamente utilizzata in campo ambientale ed economico perché elimina l'influenza delle unità di misura e mette tutti gli indicatori sullo stesso piano.

2. Aggregazione delle tre dimensioni: I valori normalizzati sono stati quindi aggregati in media aritmetica, attribuendo uguale peso (1/3) a ciascuna componente. Questa scelta risponde all'esigenza di neutralità metodologica: non si è voluto predefinire una priorità tra energia, emissioni e consumo idrico, lasciando che il risultato finale rifletta in modo equilibrato le tre dimensioni.

3. Calcolo dell'indice complessivo

Grazie a questa costruzione, l'indice permette di collocare ogni modello su una scala unica di sostenibilità, dove i valori prossimi a 0 segnalano un impatto basso e bilanciato su tutte le dimensioni e valori vicini a 1 indicano un profilo ambientale critico e uniforme nelle tre metriche.

In questo modo il lettore può cogliere in un solo colpo d'occhio le differenze sostanziali tra modelli, senza dover analizzare separatamente energia, CO_2 e acqua.

I risulati ottenuti da questa analisi applicando l'indice al campione di 16 modelli sono i seguenti:

- Modelli più efficienti: GPT-4.1 nano, Mistral-7B e GPT-3.5-turbo, che combinano bassi consumi, basse emissioni e minima variabilità.
- Modelli di fascia intermedia: GPT-4.1 mini, Claude Sonnet e GPT-5 mini, con impatto moderato ma contenuto.
- Modelli meno sostenibili: GPT-4o, LLaMA-3.2-90B Vision Instruct, Claude-Opus-4/4.1 e, in particolare, DeepSeek-Reasoner, che registra l'impronta più alta.

Questa graduatoria riproduce in modo compatto le tendenze già emerse dai singoli ranking (energia, CO_2 e acqua), ma offre una lettura più immediata e un supporto comparativo più efficace.

3.4 Analisi MMLU – Accuratezza cognitiva dei modelli

Una volta identificati i modelli ambientalmente efficienti, diventa fondamentale capire quanto tali modelli siano anche performanti dal punto di vista della qualità del testo generato. Viene introdotta una nuova variabile: l'efficienza prestazionale, intesa come rapidità e accuratezza cognitiva. Per quest'ultima, è stato scelto il benchmark Massive Multitask Language Understanding (MMLU), standard de-facto per misurare la capacità dei modelli linguistici di rispondere correttamente su oltre cinquanta discipline accademiche. MMLU rappresenta quindi una proxy sintetica della qualità e della completezza delle risposte.

Per ciascuno dei 16 modelli studiati sono stati individuati i valori MMLU disponibili attraverso fonti ufficiali (es. GPT-4.1: 90,2 %; GPT-4.1 mini: 87,5 %; GPT-4.1 nano: 80,1 %); leaderboard indipendenti (es. O3: 85,6 % su MMLU-Pro; Claude-Opus-4.1: 87,8 %) e stime ragionate quando mancavano dati ufficiali (es. GPT-5 mini: 86–88 % stimati sulla base del rapporto medio mini/main in altri benchmark; Mistral-7B: ~62,6 % sulla base di comparazioni con LLaMA-2-13B e LLaMA-34B).

I punteggi sono poi normalizzati su scala 0–1 per renderli confrontabili con l'indice ambientale sintetico e permettere successive elaborazioni.

Dall'analisi svolta appare che le prestazioni migliori sono state registrate da GPT-4.1 (90,2 %), DeepSeek-Reasoner (90,8 % standard MMLU) e GPT-5 (≈87 % su MMLU-Pro), seguiti dalle varianti Claude-Opus e Claude-Sonnet-4 (86–89 %). Le Prestazioni peggiori appartengono a GPT-4.1 nano (80,1 %), Claude-3.7 Sonnet (82,7 %), Mixtral-8×7B (71,3 %), GPT-3.5-turbo (~70 %) e Mistral-7B (~62,6 %).

Questi dati mostrano che le migliori performance cognitive appartengono ai modelli di fascia alta, ma non sempre coincidono con la maggiore sostenibilità ambientale.

I punteggi MMLU non sono solo una classifica di accuratezza, ma vengono integrati con l'indice di impatto ambientale per creare un indicatore composito, utile a individuare i modelli che offrono il miglior compromesso tra qualità e sostenibilità. In questo modo permettono di valutare trade-off importanti: un modello può essere più sostenibile a livello ambientale ma meno preciso, o viceversa.

3.5 Analisi di Pareto applicata ai modelli selezionati

Nella sua formulazione classica, il criterio di Pareto individua le combinazioni di beni o di fattori produttivi in cui nessuna dimensione può essere migliorata senza peggiorarne un'altra. Applicata ai modelli linguistici, la stessa logica consente di misurare il trade-off tra prestazioni cognitive e impatto ecologico.

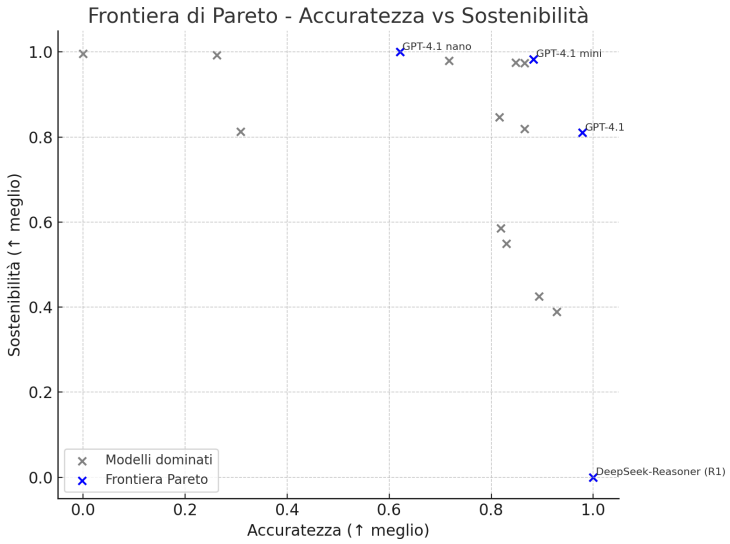
Ogni modello è rappresentato come un punto su un piano a due dimensioni, dove l'asse orizzontale indica la sostenibilità (impatto ambientale normalizzato) e quello verticale **l'**accuratezza (punteggio MMLU normalizzato). Un modello A domina un modello B quando possiede accuratezza maggiore o uguale e sostenibilità maggiore o uguale, con almeno un miglioramento stretto in una delle due dimensioni.

I modelli non dominati da alcun altro costituiscono la frontiera di Pareto: l'insieme delle soluzioni efficienti che esprimono i migliori compromessi possibili.

L'applicazione della regola di dominanza ai 16 modelli considerati ha portato a stabilire che solo quattro modelli appartengono alla frontiera di Pareto:

- DeepSeek-Reasoner R1:Massima accuratezza, sostenibilità minima.
- GPT-4.1:Buon equilibrio fra accuratezza e impatto.
- GPT-4.1 mini: Compromesso più spostato verso la sostenibilità, con accuratezza comunque elevata.
- GPT-4.1 nano: Massima sostenibilità, accuratezza più contenuta.

Tutti gli altri modelli risultano dominati, cioè superati in entrambe le metriche da almeno una di queste quattro alternative.



L'analisi ha mostrato che la frontiera è composta da un numero limitato di modelli, a conferma del fatto che non esiste una soluzione universalmente ottimale, bensì un insieme di compromessi che riflettono priorità differenti. Tale risultato offre alle imprese un quadro di riferimento chiaro all'interno del quale definire le proprie strategie: la scelta potrà orientarsi verso modelli che privilegiano la sostenibilità, quando la reputazione ambientale e gli obiettivi ESG assumono un ruolo competitivo, oppure verso modelli che massimizzano l'accuratezza, nei casi in cui la qualità cognitiva rappresenti un fattore critico di successo. In alternativa, la decisione potrà ricadere su compromessi intermedi, qualora l'obiettivo sia trovare un equilibrio tra prestazioni e responsabilità ambientale. In questa prospettiva, la frontiera di Pareto non costituisce soltanto uno strumento tecnico di analisi, ma si configura come un dispositivo decisionale, utile a orientare le strategie in un contesto in cui innovazione e responsabilità ambientale devono procedere insieme.

In questa parte della tesi abbiamo misurato e confrontato l'impatto ambientale e l'accuratezza dei modelli generativi sul mercato, ma non abbiamo ancora spiegato i fattori che lo determinano. Le prossime sezioni affronterà proprio questo aspetto, per comprendere in profondità le cause e i meccanismi alla base delle differenze osservate. L'Appendice, all'allegato 2, raccoglie i dati completi, le formule e gli approfondimenti metodologici a supporto delle analisi qui presentate.

4.1 Modello di Regressione per l'analisi della latenza

Al fine di analizzare i fattori che incidono sulla latenza dei Language Model, è stato adottato un modello di regressione lineare. La metodologia OLS consente di stimare l'effetto medio esercitato dalle variabili indipendenti (numero di token in input e in output) sulla variabile dipendente (latenza misurata in millisecondi).
Tale approccio consente di isolare i contributi dei token generati e dei token forniti in input, verificando quale dei due incida maggiormente sul tempo di risposta

Modello di Regressione Lineare:

$$latenza_{ms} = \alpha + \beta 1 \cdot input\_token + \beta 2 \cdot output\_token + \varepsilon_i$$

- α = intercetta, rappresenta la latenza attesa se input=0 e output=0.
- β_1 = variazione media della latenza se cresce di 1 unità l'input token, a output costante.
- β_2 = variazione media della latenza se cresce di 1 unità l'output token, a input costante.

I risultati evidenziano che la latenza cresce in maniera statisticamente significativa al crescere dei token di output, mentre l'effetto dei token di input risulta molto ridotto e non significativo al livello del 5%. In altri termini, il tempo di risposta dei modelli è guidato soprattutto dal processo di generazione dell'output, mentre la fase di elaborazione dell'input ha un peso marginale. In particolar modo, Ogni token di output aggiuntivo comporta un incremento di 19,12 millisecondi nella latenza.

Intercetta: 665,5 ms (p=0,193)

β_1 (input tokens)**:** 0,252 ms per token (p=0,090)

β_2 (output tokens)**:** 19,212 ms per token (p<0,001)

R^2 = 0,531**;** R^2 aggiustato = 0,527

Tuttavia, il valore di R^2 (0,53) indica che il modello spiega solo poco più della metà della variabilità osservata nella latenza. Questo suggerisce che vi siano altri fattori rilevanti non inclusi nel modello di base.

Per stimare i fattori che influenzano la latenza dei modelli generativi è stato costruito un modello di regressione OLS che combina variabili quantitative e qualitative.

Per verificare l'assenza di multicollinearità tra le variabili indipendenti, è stato calcolato il Variance Inflation Factor (VIF). Tutti i valori relativi alle variabili utilizzate nel modello, risultano inferiori alla soglia critica di 5, indicando che non sussistono problemi di collinearità tali da compromettere la stabilità delle stime.

Sono stati verificati i residui tramite test di normalità e di eteroschedasticità (Breusch–Pagan). In alternativa, modelli GLM con link log o regressioni quantiliche potrebbero rappresentare estensioni utili, ma si è scelto di mantenere l'OLS per coerenza e leggibilità dei risultati."

Le variabili quantitative sono state considerate in forma continua, mentre per le variabili qualitative si è proceduto alla trasformazione in variabili dummy, individuando una categoria di riferimento per ciascuna, in modo da consentire l'interpretazione dei coefficienti come differenziali rispetto alla baseline.

- **Input tokens**
  Numero di token contenuti nel prompt fornito al modello. Rappresenta la dimensione dell'input da elaborare nella fase iniziale di inferenza.
  Si tratta di una variabile quantitativa continua.
- **Output tokens**
  Numero di token generati in output dal modello.
  Si tratta di una variabile quantitativa continua.
- **Parametri effettivi (espressi in B)**
  Dimensione complessiva dell'architettura del modello espressa in miliardi di parametri. Indica la complessità del modello, spesso utilizzata come proxy della scala del modello.
  Si tratta di una variabile quantitativa continua.
- **GPU effettive**
  Numero di GPU allocate per l'inferenza. Si tratta di una variabile quantitativa continua.
- **Architettura**
  Variabile dummy che assume valore 1 se il modello utilizza un'architettura *Mixture of Experts* (MoE) e 0 se adotta una architettura *Dense*. La categoria di riferimento è Dense (0).
- **CoT (Chain-of-Thought)**
  Variabile dummy che vale 1 per modelli che generano catene di ragionamento esplicite e 0 se non le generano. La categoria presa come riferimento è *CoT = No*.
- **Multimodalità**
  Variabile dummy che vale 1 per richieste multimodali (testo + altre modalità) e 0 per richieste solo testuali. La categoria presa come riferimento è *Modelli solo testo*.
- **HardWare assegnato** (A100, H100, L4)
  Variabili dummy che distinguono tra diverse tipologie di GPU utilizzate per l'inferenza. In particolare, sono state codificate dummies per A100 e L4, con valore 1 se il modello utilizza quel tipo di GPU.
  La categoria di riferimento è *GPU H100*.

Al fine di analizzare in maniera più completa i fattori che incidono sulla latenza dei Language Model, è stato adottato un modello di regressione lineare multivariato stimato con la metodologia OLS. Questo approccio consente di stimare l'effetto medio esercitato non solo dal numero di token in input e in output, ma anche da variabili strutturali (parametri del modello, architettura) e infrastrutturali (GPU e hardware) sulla variabile dipendente latenza (espressa in millisecondi).

Modello di Regressione Lineare:

$$\text{latency}_{ms} = \alpha + \beta_1 \cdot \text{input\_tokens} + \beta_2 \cdot \text{output\_tokens} + \beta_3 \cdot \text{Parametri} + \beta_4 \cdot \text{GPU}$$
$$+\beta_5 \cdot \text{Architettura} + \beta_6 \cdot \text{CoT} + \beta_7 \cdot \text{Multimodalità} + \beta_8 \cdot \text{HW} + \varepsilon_i$$

- α **(intercetta)**: rappresenta il valore atteso della variabile dipendente (latenza) quando tutte le variabili indipendenti assumono valore zero.

- β_1 **(input tokens)**: misura la variazione attesa della latenza associata all'incremento unitario degli input tokens, a parità delle altre variabili indipendenti.

- β_2 **(output tokens)**: misura la variazione attesa della latenza associata all'incremento unitario negli output tokens, mantenendo costanti le altre variabili.

- β_3 **(Parametri)**: misura la variazione attesa della latenza all'incremento unitario del numero di parametri, mantenendo costanti le altre variabili.

- β_4 **(GPU)**: misura la variazione attesa della latenza all'incremento unitario del numero di GPU, mantenendo costanti le altre variabili**.**

- β_5 **–** β_9: rappresentano i coefficienti associati alle variabili qualitative e strutturali. Ogni coefficiente stima la differenza media nella latenza tra la categoria considerata e la categoria di riferimento, mantenendo costanti le altre variabili.

$$latency_{ms} = 1.608 \cdot 10^4 + 0.29 \cdot input\_tokens + 18.49 \cdot output\_tokens + 49.50$$
$$\cdot Parametri\_effettivi - 2603 \cdot GPU - 7742 \cdot Architettura\_MoE + 22870 \cdot CoT\_Si$$
$$+ 7627 \cdot Multimodalita\_Si - 17370 \cdot HW\_A100 - 24750 \cdot HW\_L4 + \varepsilon_i$$

I coefficienti riportati rappresentano stime medie condizionate e non valori deterministici. Vanno quindi interpretati come effetti marginali attesi, validi ceteris paribus, e soggetti a un margine di incertezza residua. Le stime dei coefficienti sono state valutate al livello di significatività convenzionale del 5% (IC 95%).

L’analisi di regressione mostra che ogni incremento unitario nel numero di token generati si associa a un aumento medio della latenza pari a circa 18,5 millisecondi ($p<0,001$). Questo indica che la fase di generazione è il fattore più oneroso in termini di tempo, mentre l’elaborazione dell’input ha un peso marginale. Infatti, gli input tokens hanno un coefficiente positivo ma molto ridotto: ogni token aggiuntivo è associato a un incremento medio di 0,29 millisecondi ($p<0,01$), indicando che il contributo dell’input è trascurabile rispetto all’output.

La complessità del modello, misurata attraverso i parametri effettivi, ha un effetto significativo: ogni incremento unitario (espresso in miliardi di parametri) si traduce in un aumento medio della latenza di circa 49 ms ($p<0,001$).

In senso opposto, il numero di GPU effettive assegnate al modello ha un effetto riduttivo: ciascuna GPU aggiuntiva riduce in media la latenza di 2603 millisecondi ($p<0,001$), mostrando il fatto che l’assegnazione di più risorse hardware consente di ridurre sensibilmente i tempi di inferenza.

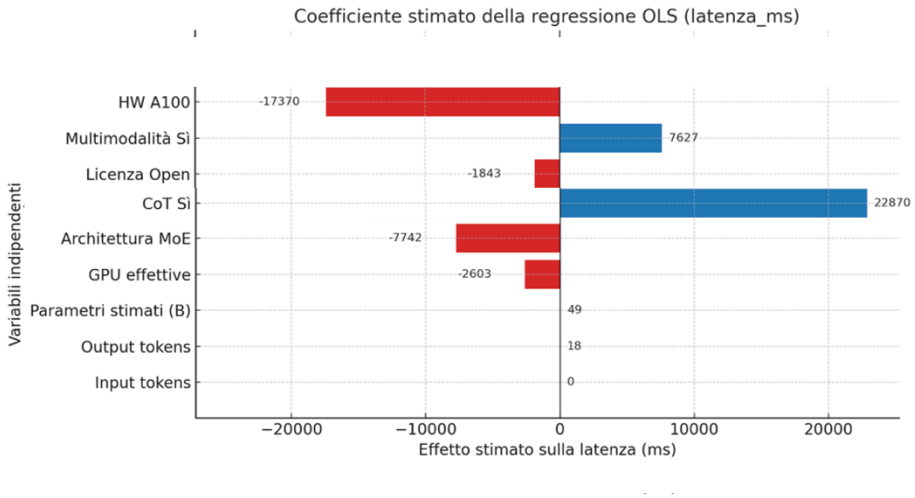
Tra le variabili qualitative emergono risultati interessanti. I modelli con architettura MoE registrano una latenza inferiore di circa 7742 millisecondi (quasi 8 secondi in meno) rispetto ai modelli Dense ($p<0,001$), a parità delle altre condizioni. Questo punto è cruciale perché evidenzia come la “dimensione stimata” del modello non sia sempre un buon predittore della latenza: ciò che conta è la porzione di parametri realmente attivata a runtime.

La Chain-of-Thought comporta un incremento consistente pari a 22.870 ms, a parità di condizioni. Il modello, anziché restituire direttamente una risposta, genera passaggi intermedi espliciti di ragionamento, allungando sensibilmente la sequenza di output e quindi i tempi complessivi di inferenza quindi, l’output è più lungo e richiede più token da generare, più operazioni interne, verosimilmente più accesso alla memoria, più calcolo sequenziale dato che la produzione di passaggi intermedi prolunga la sequenza di output.

Anche la Multimodalità è associata a un aumento sostanziale dei tempi di risposta, pari a 7627 millisecondi ($p<0,001$). L’integrazione di modalità diverse (testo, immagini, audio) richiede pipeline più complesse e l’attivazione di sottoreti aggiuntive rispetto a un modello puramente testuale. Questo è dovuto dalla presenza di overhead legati alla gestione e fusione di input eterogenei. Ciò conferma che la multimodalità, pur rappresentando un vantaggio in termini di versatilità applicativa, ha un costo prestazionale significativo.

Nel complesso, il modello raggiunge un R^2 di 0,81, spiegando circa l’81% della variabilità osservata.

Tuttavia, una quota di variabilità residua rimane non catturata, verosimilmente legata a fattori non osservabili (batching dinamico, caching, autoscaling), non accessibili tramite API pubbliche.

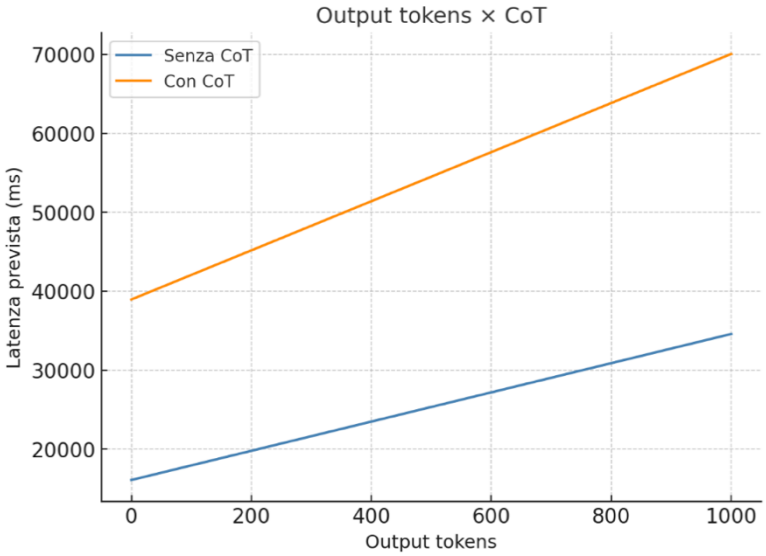


- (CoT × Multimodalità)

In seguito, è stato introdotto un termine di interazione (CoT × Multimodalità) per verificare se la combinazione di queste due caratteristiche producesse un effetto maggiore o minore rispetto alla semplice somma dei loro impatti individuali. Il coefficiente dell'interazione è risultato statisticamente significativo ($p < 0.001$), ma numericamente praticamente nullo (~0). Questo significa che non emergono né amplificazioni né attenuazioni: ogni capacità avanzata (CoT, multimodalità) ha un costo prestazionale autonomo e quando i modelli integrano più capacità, i costi si accumulano linearmente, senza compensazioni.

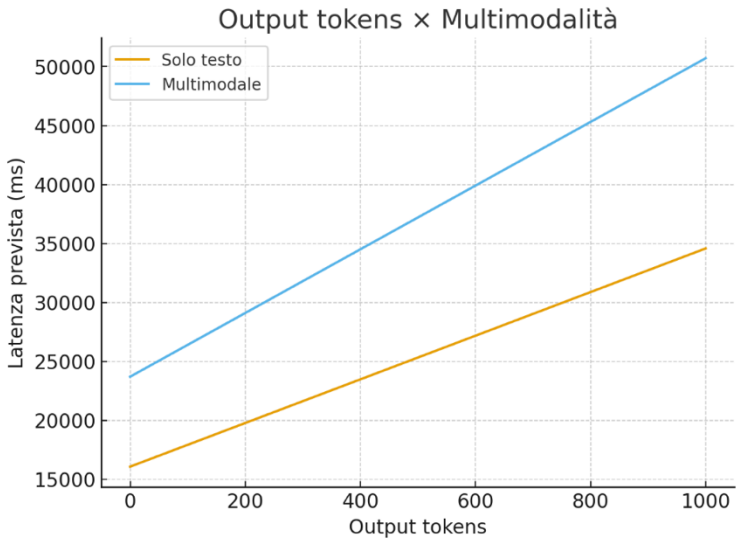
- Output tokens × Chain-of-Thought (CoT)

Abbiamo analizzato l'interazione tra output tokens e Chain-of-Thought per verificare se la presenza del CoT modificasse la relazione tra lunghezza dell'output e latenza. Il coefficiente stimato (+12,6 ms per token, $p < 0.001$) mostra che, nei modelli con CoT, ogni token prodotto è più costoso rispetto a un modello standard di 12,6 millisecondi. Questo significa che l'effetto del CoT non si limita a introdurre un ritardo "fisso", ma aumenta anche la pendenza della curva che lega numero di token e latenza. Questo rende i modelli con CoT particolarmente onerosi in scenari di output esteso, come la scrittura di testi complessi o la risoluzione di problemi multi-step (spiega perché sistemi come o3 e DeepSeek-R1 risultano molto più energivori). In altre parole, più lunga è la sequenza di output, più si amplifica il tempo complessivo.



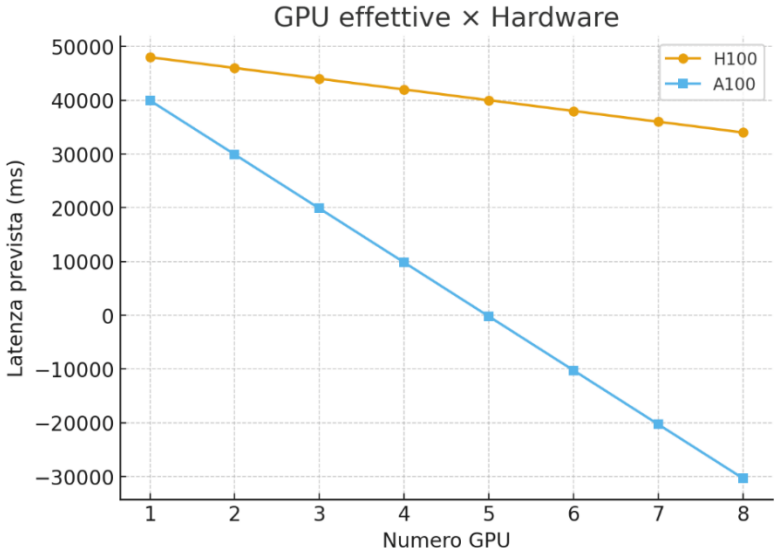
3. Output tokens × Multimodalità

L'interazione tra token in output e multimodalità indica che anche la multimodalità non si limita a un incremento iniziale, ma incide proporzionalmente sul numero di token generati. Ogni token aggiuntivo generato da modelli che sfruttano la Multimodalità richiede 8,5 millisecondi aggiuntivi rispetto ad un modello che generano solo testo (+8,5 ms per token, $p < 0.001$).
L'onerosità, quindi, non è solo strutturale (supportare più modalità), ma anche incrementale. Questo penalizza in particolare applicazioni di generazione massiva multimodale (es. analisi video).



- GPU effettive × Hardware (A100 vs H100)

Infine, abbiamo verificato l'interazione tra GPU effettive e classe hardware (A100 vs H100) per valutare se l'impatto delle GPU aggiuntive fosse uniforme. Il risultato mostra che su A100 l'aggiunta di una GPU riduce la latenza di circa 8 secondi in più rispetto agli H100 (−8041 ms per GPU su A100, $p < 0.01$).
La motivazione di questa analisi è da ricercare nel fatto che le GPU di vecchia generazione (A100) beneficiano in modo maggiore dall'aggiunta di parallelismo, mentre le GPU di nuova generazione (H100) sono già ottimizzate, quindi presentano ritorni marginali più bassi.
Un modello piccolo ma su hardware meno efficiente può consumare più energia di un modello grande su hardware ottimizzato.



4.1.1 ANOVA

L'analisi dell'ANOVA (Analysis of Variance) L'ANOVA (Analysis of Variance) serve a confrontare le medie di più gruppi per capire se le differenze osservate sono reali o se possono essere dovute al caso. Di seguito i risultati dell'analisi svolta sulle principali variabili prese in considerazione in questo studio:

- **Architettura (Dense vs MoE)**
  Abbiamo analizzato se l'architettura del modello (Dense o Mixture of Experts) influenzasse la latenza. Il risultato mostra una differenza significativa ($F = 9{,}43$; $p = 0{,}002$): i modelli MoE hanno una latenza mediamente inferiore rispetto ai Dense, a conferma che l'attivazione parziale dei parametri a runtime rende più efficiente il processo di inferenza.
- **CoT (Chain-of-Thought)**
  Abbiamo analizzato l'effetto della Chain-of-Thought sulla latenza media. I risultati mostrano la differenza più netta di tutte ($F = 51{,}41$; $p < 0{,}001$): i modelli che generano passaggi intermedi di ragionamento esplicito hanno tempi di inferenza sensibilmente più lunghi. Questo dato conferma

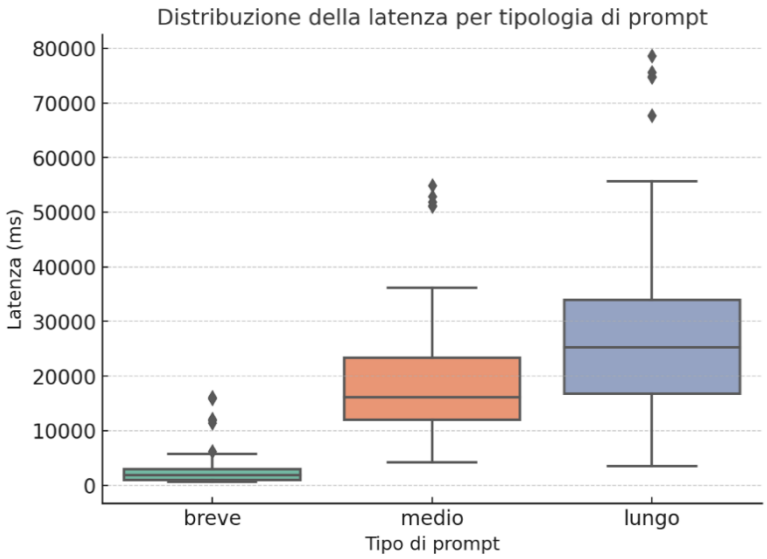
quanto emerso anche nei modelli di regressione, dove CoT risultava tra i principali determinanti della latenza.

- **Multimodalità (Sì/No)**
  Abbiamo verificato se la multimodalità incidesse in modo significativo sulla latenza media. I risultati evidenziano differenze statisticamente significative (F = 31,5; p = 0,008): la latenza media dei modelli multimodali è diversa rispetto ai modelli solo testuali.
- **Hardware (H100, A100, L4)**
  Abbiamo analizzato le differenze di latenza tra le classi di GPU (H100, A100, L4). L'ANOVA conferma differenze significative (F = 12,57; p < 0,001): le performance cambiano in modo rilevante in funzione dell'hardware. In particolare, gli A100 mostrano latenze più alte degli H100, mentre gli L4 risultano in alcuni casi più veloci degli H100, a dimostrazione che l'efficienza computazionale non dipende solo dalla potenza teorica della GPU ma anche dal tipo di ottimizzazione adottata.

Oltre alle variabili considerate nel modello di regressione, l'analisi ANOVA rappresenta uno strumento utile per verificare l'impatto di altri elementi oggetto di questo studio e per i quali sono stati raccolti dati empirici.

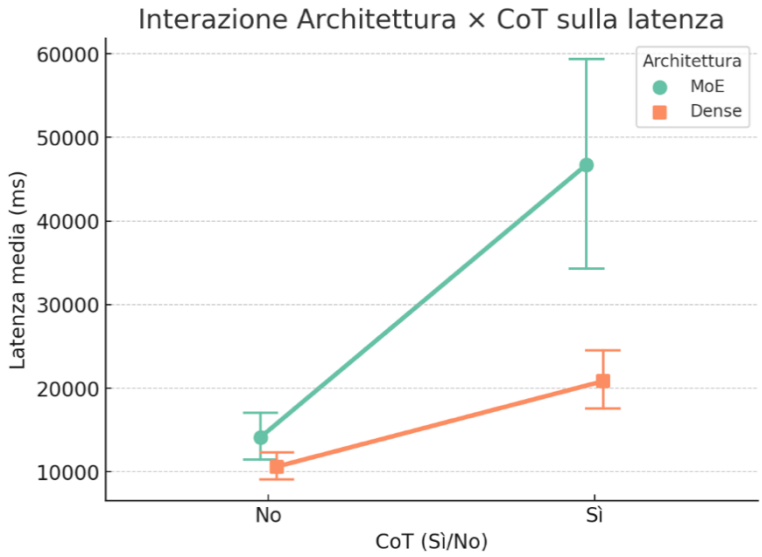
In particolare, è stato analizzato il ruolo della lunghezza del prompt (breve, medio, lungo) sulla latenza. L'ANOVA ha evidenziato una differenza altamente significativa (F = 97,58; p < 0,001), indicando che la latenza cresce sistematicamente al variare della lunghezza del prompt.

Per approfondire queste differenze è stato applicato il test post-hoc di Tukey HSD**,** che consente di confrontare le medie a coppie e stabilire tra quali gruppi le differenze siano statisticamente robuste. I risultati mostrano che tutti i confronti (breve vs medio, breve vs lungo, medio vs lungo) sono significativi (p < 0,001), confermando che la latenza aumenta progressivamente al crescere della complessità e della lunghezza del prompt.



Per rispondere alla domanda *"L'effetto del CoT sulla latenza è lo stesso nei modelli Dense e nei MoE?"* è stato utilizzato l'analisi Two-way ANOVA.

L'analisi a due vie ha valutato l'effetto congiunto dell'architettura del modello (Dense vs MoE) e della presenza di Chain-of-Thought sulla latenza. I risultati mostrano che entrambi i fattori incidono in modo significativo, ma soprattutto che l'interazione tra di essi è anch'essa significativa ($F = 26{,}66$; $p < 0{,}001$). Questo significa che l'impatto del CoT non è uniforme: nei modelli Dense la latenza cresce molto più rapidamente rispetto ai modelli MoE, in cui l'attivazione parziale dei parametri riduce parzialmente il costo prestazionale del ragionamento esplicito. In termini applicativi, il costo energetico del CoT non dipende soltanto dal suo utilizzo, ma anche da come il modello è progettato: le architetture MoE si dimostrano più resilienti, mentre i Dense soffrono maggiormente l'aumento di complessità introdotto dal CoT.



Per evitare di penalizzare i modelli che generano un numero maggiore di token o che utilizzano più GPU, è stata condotta un'analisi sulla latenza normalizzata. Sono stati costruiti due indici: la latenza per token, che misura il tempo medio di generazione di un singolo token, e la latenza per GPU, che misura il tempo medio di generazione rapportato al numero di GPU utilizzate. Questi indicatori rappresentano una proxy dell'efficienza computazionale, in quanto consentono di confrontare i modelli a parità di carico o di risorse.

Una votla individuate queste variabili, è possibile rispondere alla domanda *"A parità di token o di GPU, quali categorie di modelli consumano di più?"*

L'analisi ANOVA sulla latenza per token mostra che i modelli con Chain-of-Thought risultano i meno efficienti ($F = 85{,}87$; $p < 0{,}001$): ogni token generato da modelli con CoT è significativamente più oneroso

in termini di tempo. Anche la multimodalità incide negativamente ($F = 8,97$; $p = 0,003$), con tempi di generazione per token più elevati.

L'analisi ANOVA per la latenza per GPU mostra effetti significativi sia del CoT ($F = 9,41$; $p = 0,002$) che della multimodalità ($F = 23,19$; $p < 0,001$), a conferma che queste caratteristiche generano un overhead consistente anche quando vengono allocate più risorse. Inoltre, la variabile hardware ha un impatto rilevante ($F = 15,02$; $p < 0,001$): le GPU A100 risultano mediamente meno efficienti, le H100 più performanti, mentre le L4 ottengono in alcuni scenari risultati competitivi. Al contrario, l'architettura (Dense vs MoE) non mostra differenze significative ($F = 0,00024$; $p = 0,99$).

Finora in questo capitolo è stato analizzato come l'output, le variabili strutturali (parametri del modello, architettura) e infrastrutturali (GPU e hardware) possano influire sulla latenza (espressa in millisecondi), considerata punto di partenza per determinare il fabbisogno energetico e, di conseguenza, l'impatto ambientale.
Nei paragrafi successivi l'analisi si concentrerà su come la localizzazione dei data center possa influire sull'impatto ambientale, in particolare in termini di emissioni di CO_2 e di consumo idrico necessario per il raffreddamento delle infrastrutture.

4.2 Localizzazione dei Data Center

I principali provider (AWS, Google, Azure, ecc) offrono la possibilità di scegliere la regione geografica di elaborazione e archiviazione dei dati. La localizzazione dei data center assume rilevanza in tema di modelli di generative AI su due piani distinti. In primo luogo, essa influisce sulla latenza end-to-end percepita dall'utente finale: maggiore è la distanza geografica, più elevato è il tempo di trasmissione dei dati, sebbene ciò non modifichi la latenza di inferenza generata a livello di GPU, che rimane l'oggetto dell'analisi. In secondo luogo, la posizione geografica dei data center è determinante ai fini dell'impatto ambientale, in quanto il mix energetico locale (carbone, gas, nucleare, rinnovabili) condiziona in misura significativa l'impronta di carbonio associata all'esecuzione dei modelli.

Per indagare in che modo la localizzazione geografica dei data center influenzi il consumo energetico e l'impatto emissivo, è stato costruito un dataset originale a risoluzione oraria basato su dati open-source di ElectricityMaps. La raccolta ha interessato dodici Paesi e zone elettriche individuati come sedi dei principali poli di calcolo a livello globale e delle infrastrutture dei maggiori provider di servizi cloud – Amazon Web Services (AWS), Microsoft Azure e Google Cloud: USA–Virginia (US-MIDA-PJM), USA–California (US-CAL-CISO), Canada (CA-ON), Brasile (BR-CS), Germania (DE), Francia (FR), Irlanda (IE), Paesi Bassi (NL), Svezia (SE-SE1), Italia (IT), India (IN-NE) e Giappone (JP).

Area	Paese / Regione chiave	Provider cloud presenti
America del Nord	**USA – Virginia**	AWS, Azure, Google
	USA – California	AWS, Azure, Google
	Canada	AWS, Azure, Google
America del Sud	**Brasile**	AWS, Azure, Google
Europa	**Germania**	AWS, Azure, Google
	Francia	AWS, Azure, Google
	Irlanda	AWS, Azure, Google
	Italia	AWS (in arrivo), Azure, Google
	Svezia	AWS, Azure, Google
	Paesi Bassi	AWS, Google
Asia	**India**	AWS, Azure, Google
	Giappone	AWS, Azure, Google

Per rappresentare in modo robusto le differenti condizioni di domanda elettrica, sono stati estratti i dati giornalieri relativi all'anno 2024. Per ciascun paese, ciascun mese dell'anno e per ogni ora delle giornate, sono stati raccolti dati relativi a un insieme di variabili energetiche e ambientali, che permettono di caratterizzare in dettaglio l'impatto emissivo e il mix di generazione:

- Intensità carbonica diretta, espressa in gCO_2eq/kWh, che misura le emissioni legate alla sola produzione elettrica.
- Intensità carbonica a ciclo di vita, che considera l'intero ciclo di vita della generazione, dalla costruzione degli impianti alla dismissione.
- Quota di energia rinnovabile (RE%)[41], ossia la percentuale di elettricità proveniente da fonti rinnovabili come eolico, solare, idroelettrico e biomassa.
- Quota di energia carbon-free (CFE%)[42], che comprende tutte le fonti a emissioni nette zero, incluse rinnovabili e nucleare.
- Composizione oraria del mix di generazione, con la produzione (in MW) imputabile a ciascuna fonte: wind, solar, hydro, nuclear, gas, coal, biomass e other.
- Variabili temporali (Month, Day, Hour). La variabile Ora è stata scomposta in Orario in UTC e Ora Locale.[43]
- Variabile categoriale che sintetizza le stagioni locali, in modo da rispecchiare le effettive condizioni climatiche nei paesi presi in esame e permettere di analizzare correttamente l'impatto stagionale sulla produzione energetica e sulle emissioni di carbonio.[44]

Questa architettura consente non solo di effettuare analisi comparative tra Paesi, ma anche di valutare l'influenza della stagionalità, della dimensione mensile e della fascia oraria sulla disponibilità di energia a basse emissioni. In tal modo è possibile esplorare correlazioni tra localizzazione dei data center, orario di funzionamento e impatto energetico, individuando opportunità di ottimizzazione spaziale e temporale dei carichi di calcolo e contribuendo alle strategie di decarbonizzazione dell'intelligenza artificiale generativa e delle infrastrutture digitali.

---

41 Nel caso in cui una rete non utilizza energia nucleare la percentuale RE coincide con CFE, mentre quando il nucleare contribuisce alla produzione elettrica la quota carbon-free risulta più alta della sola quota rinnovabile.

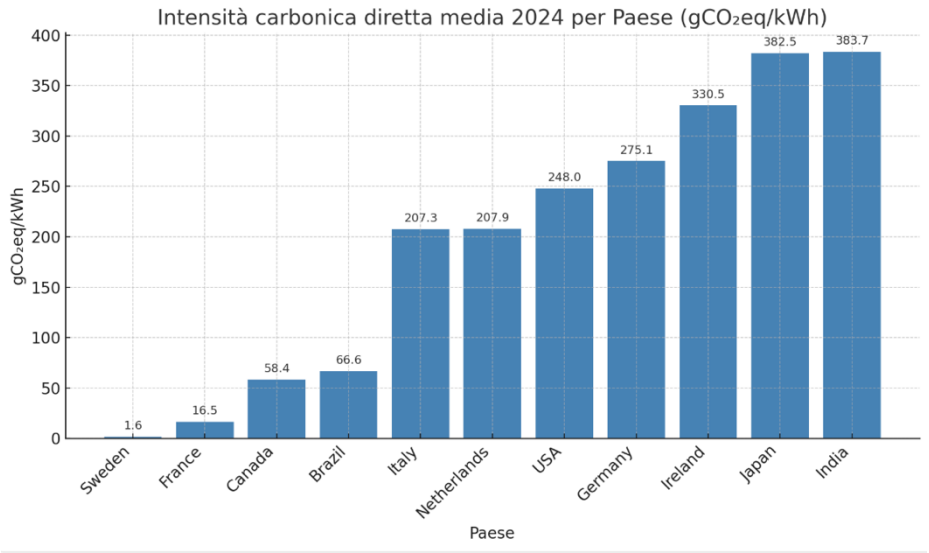
42 Nel caso in cui una rete non utilizza energia nucleare la percentuale RE coincide con CFE, mentre quando il nucleare contribuisce alla produzione elettrica la quota carbon-free risulta più alta della sola quota rinnovabile.

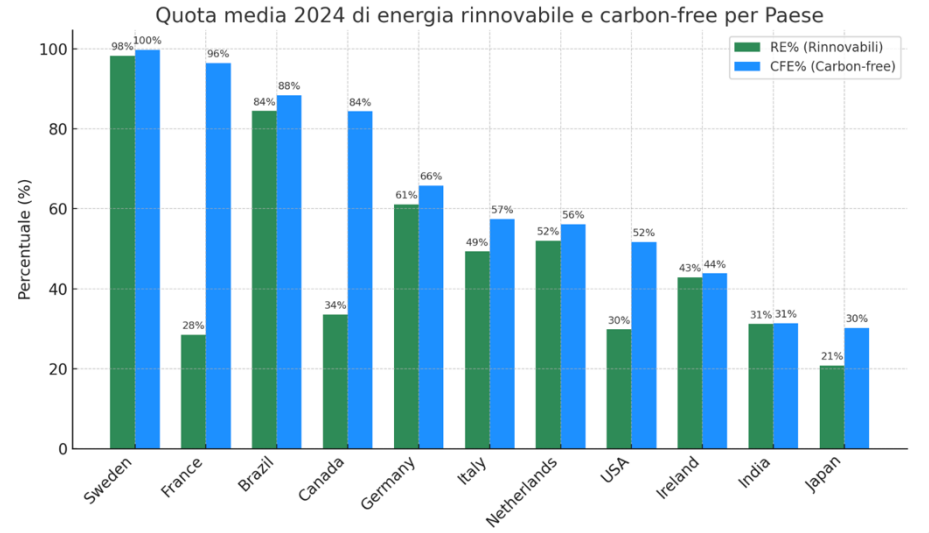
43 I dati sono stati raccolti utilizzando l'orario in UTC (tempo universale coordinato), in modo da evitare l'influenza del fuso orario nell'analisi. Per avere un'analisi maggiormente accurata, che possa rispecchiare il reale utilizzo degli strumenti di GenAI, è stato calcolato l'orario locale, corrispondente all'orario in UTC. In questo modo è possibile individuare best-practice di utilizzo.

44 La stagione meteorologica tipica dell'emisfero nord (inverno: dicembre–febbraio, primavera: marzo–maggio, estate: giugno–agosto, autunno: settembre–novembre). Tale classificazione, pur essendo utile per fornire un riferimento uniforme, non rispecchia le effettive condizioni climatiche nei Paesi situati nell'emisfero sud. Per questo motivo si è scelto di definire la stagione locale, che tiene conto dell'emisfero in cui si colloca ciascun Paese. In particolare: per i Paesi dell'emisfero nord (es. Stati Uniti, Canada, Italia, Germania, Francia, Irlanda, Paesi Bassi, Svezia, India, Giappone) la classificazione delle stagioni coincide con quella standard; per i Paesi dell'emisfero sud (nel nostro caso il Brasile) le stagioni sono invertite, così che i mesi di giugno–agosto siano considerati inverno e non estate, e analogamente dicembre–febbraio siano considerati estate e non inverno.

4.2.1 Intensità carbonica e composizione mix elettrico (fonti rinnovabili e fonti carbon-free) per paese

Per avere un quadro immediato delle differenze geografiche, la prima analisi confronta la media annua dell'intensità carbonica diretta (gCO₂eq/kWh) dei dodici Paesi considerati. Questo indicatore sintetizza quanta CO_2 viene emessa per ogni kWh di elettricità prodotta nei sistemi che alimentano i data center.





L'analisi congiunta della carbon intensity media e delle quote medie di energia rinnovabile e carbon-free fornisce un quadro completo del grado di decarbonizzazione dei sistemi elettrici nei dodici Paesi esaminati. La Svezia si conferma il contesto più virtuoso: con appena 1,6 gCO₂eq/kWh e una quota quasi totale di energia carbon-free (≈100%), di cui circa il 98% da fonti rinnovabili, rappresenta un benchmark per i data center a basse emissioni. Anche la Francia mostra un'elevata sostenibilità (≈16 gCO₂eq/kWh), seppure fondata soprattutto sull'energia nucleare (≈68%), come evidenzia l'ampia differenza tra la quota CFE% (≈96%) e quella RE% (≈28%).

Un gruppo intermedio è formato da Canada e Brasile, entrambi caratterizzati da una combinazione di idroelettrico e altre rinnovabili che consente di mantenere l'intensità carbonica tra 60 e 70 gCO_2eq/kWh, con CFE% intorno all'85 %.

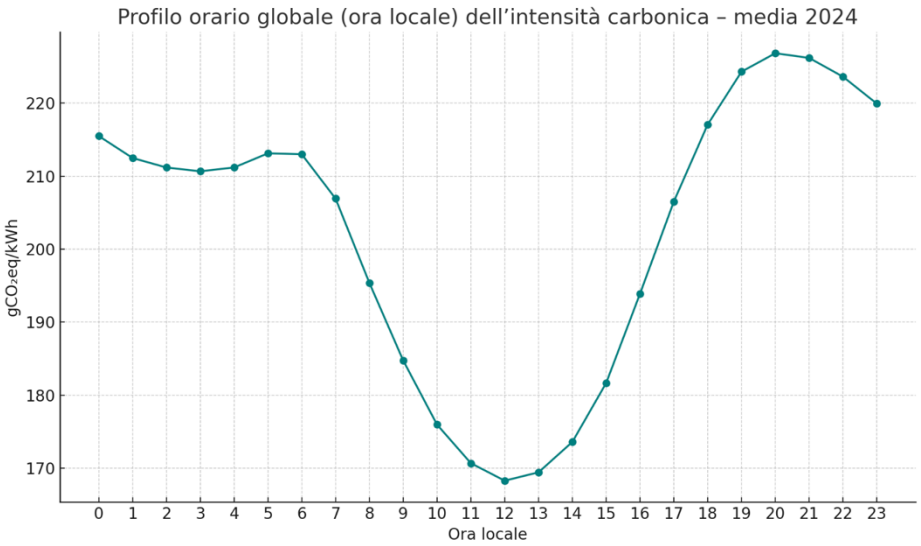
Italia, Paesi Bassi, Stati Uniti e Germania mostrano valori compresi tra 200 e 280 gCO_2eq/kWh, segno di una decarbonizzazione parziale: la quota di rinnovabili si aggira tra il 50 e il 60 %, ma la componente fossile resta significativa.

Infine, Irlanda, Giappone e India registrano le performance peggiori, con intensità carboniche superiori a 330 gCO_2eq/kWh e una quota carbon-free inferiore al 45 %, indicativa di una persistente dipendenza da carbone e gas.

Nel complesso, l'incrocio tra intensità delle emissioni e composizione del mix energetico dimostra che la localizzazione dei data center influisce in modo decisivo sull'impronta di carbonio: scegliere Paesi come Svezia, Francia, Canada o Brasile potrebbe ridurre drasticamente le emissioni associate all'esecuzione di modelli di intelligenza artificiale generativa.

4.2.2 Analisi orario globale dell'intensità carbonica

L'analisi oraria condotta sull'intero campione dei dodici Paesi, con conversione dall'ora UTC all'ora locale di ciascuna area, evidenzia una marcata dinamica giornaliera dell'intensità carbonica diretta (gCO_2eq/kWh).



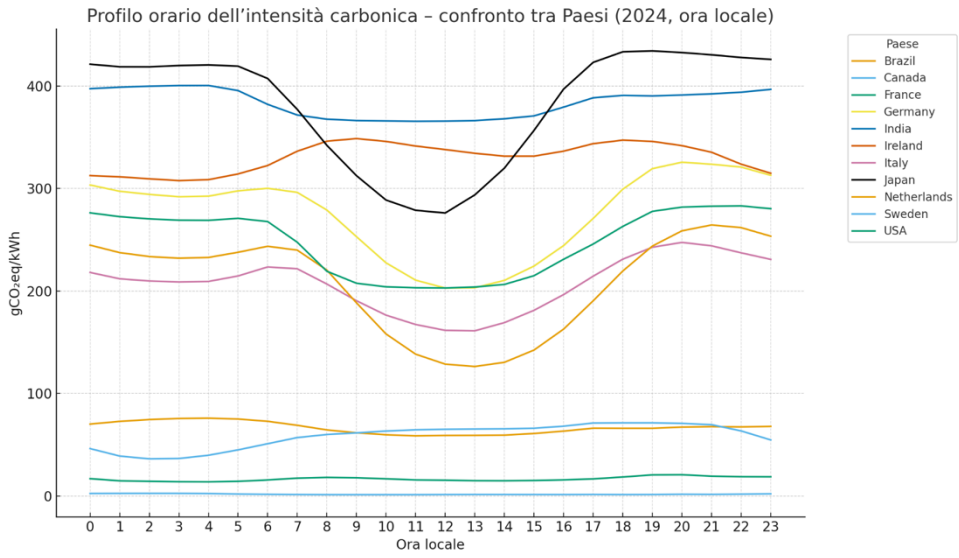
La minima intensità si registra mediamente alle 12:00 locali (fascia oraria tra 11:00 e 13:00), con un valore di circa 168 gCO_2eq/kWh, mentre la massima si osserva verso le 20:00 (fascia oraria tra 19:00 e 21:00),, con circa 227 gCO_2eq/kWh. Questa distribuzione trova spiegazione nella fisiologia del sistema elettrico:

- a mezzogiorno la combinazione di massima produzione fotovoltaica e domanda moderata riduce il ricorso a combustibili fossili;
- nelle prime ore serali il calo del solare, unito al picco della domanda domestica, comporta un maggiore utilizzo di gas e carbone.

L'inclusione dell'ora locale consente di allineare il profilo emissivo al reale ciclo solare di ciascuna area geografica. Ne deriva un'indicazione operativa rilevante: programmare i carichi di calcolo ad alta intensità energetica nelle ore centrali della giornata locale può contribuire in modo significativo alla riduzione delle emissioni complessive dei data center.

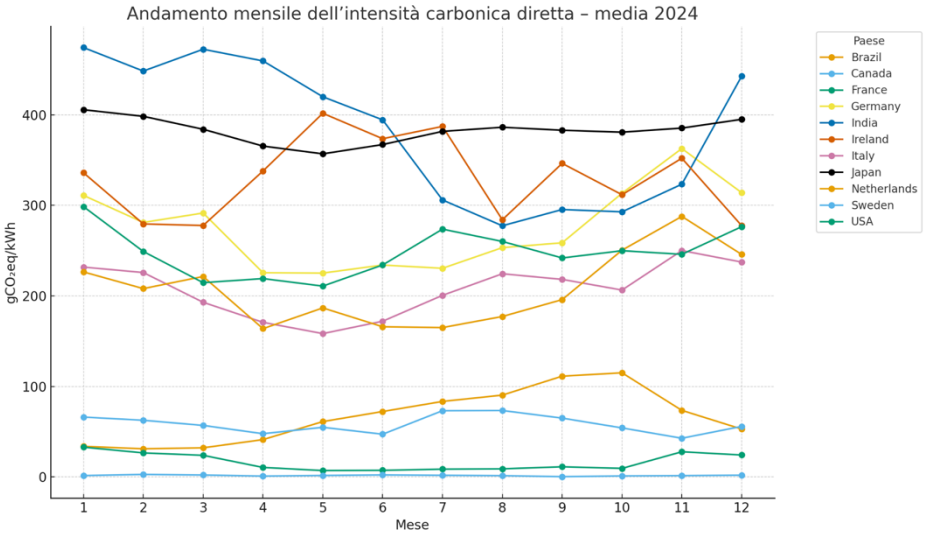
Le variazioni nei picchi massimi e minimi tra Paese e Paese risultano contenute, ma meritano comunque di essere approfondite:

- la Svezia evidenzia i valori assoluti più bassi dell'intero campione, grazie a un mix dominato da idroelettrico e nucleare che garantisce stabilità.
- Brasile presenta uno slittamento del minimo verso le prime ore del pomeriggio (13:00–16:00), riconducibile alla diversa distribuzione giornaliera della generazione idroelettrica.
- India si distingue per un picco serale più esteso, spesso protratto oltre le 21:00, in parte dovuto all'elevata domanda che caratterizza le ore notturne.
- California mostra una riduzione più marcata a mezzogiorno, segno della forte penetrazione del fotovoltaico, mentre la Virginia (East Coast USA) evidenzia un profilo più simile a quello europeo.
- Italia, Francia, Germania, Paesi Bassi seguono un andamento pressoché identico, con minimi compatti attorno a mezzogiorno e picchi serali concentrati tra le 19:00 e le 21:00.
- L'Irlanda mostra un comportamento controcorrente, si registra un picco tra le 7:00 e le 12:00, superiori all'aumento nelle ore serali.



4.2.3 Andamento mensile dell'intensità carbonica diretta per paese

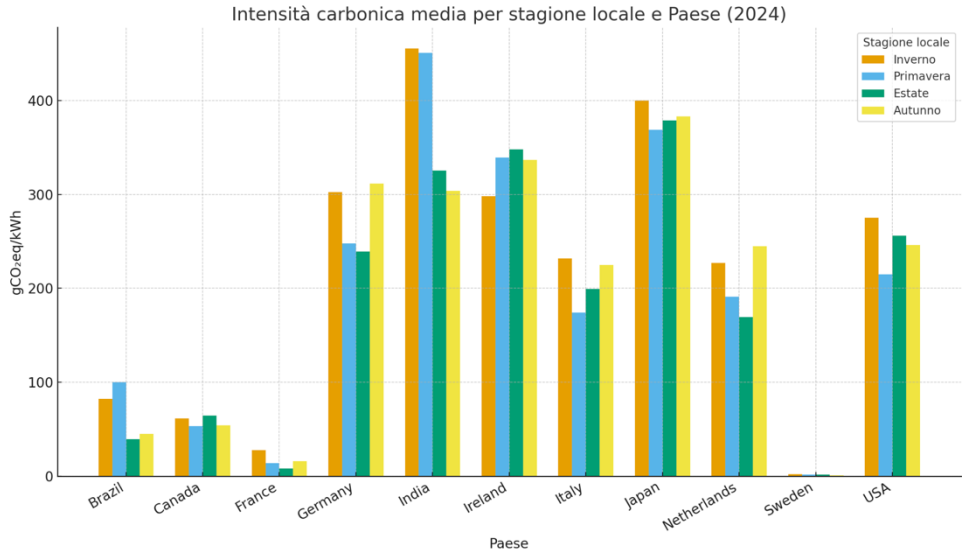
L'analisi di seguito mostra, per ciascun Paese, l'andamento mensile dell'intensità carbonica diretta (gCO_2eq/kWh). L'obiettivo è evidenziare come le caratteristiche climatiche e il mix energetico influenzino la disponibilità di energia a basse emissioni nel corso dell'anno.



La maggior parte dei Paesi europei (Italia, Francia, Germania, Paesi Bassi, Irlanda, Svezia) evidenzia valori più elevati in gennaio-febbraio, una progressiva discesa verso luglio-agosto e una lieve ripresa autunnale. Canada e Stati Uniti riproducono un andamento simile, evidenziando il classico profilo "alto-basso-alto": valori elevati in inverno, minimi a luglio–agosto, nuova risalita verso dicembre.
Il Brasile si distingue per un massimo nei mesi settembre-novembre, periodo di ridotta produzione idroelettrica, mentre India e Giappone mostrano oscillazioni più contenute ma con lievi incrementi nei mesi più caldi per l'uso del condizionamento.
Nel complesso emerge che l'inverno è il periodo mediamente più emissivo, mentre l'estate rappresenta una finestra favorevole per ridurre l'impronta carbonica delle infrastrutture digitali.

4.2.4 Analisi stagionale dell'intensità carbonica

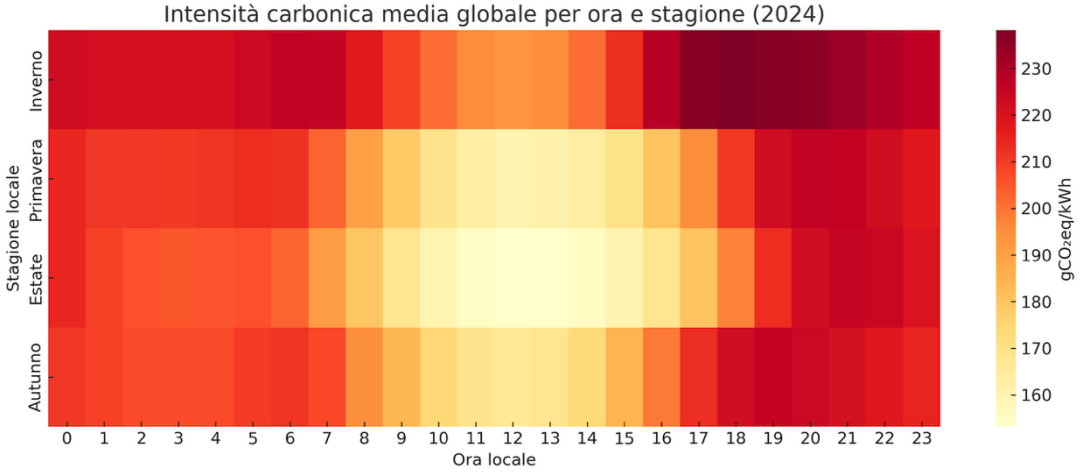
Dopo aver analizzato l'intensità carbonica a livello mensile, per evitare la distorsione legata alle differenze emisferiche (ad esempio inverno in Europa e estate in Brasile nello stesso mese), si procede con l'analisi stagionale. L'aggregazione per stagione locale permette di sintetizzare le differenze strutturali fra Paesi, mettendo in relazione l'intensità carbonica media con le principali dinamiche climatiche e di produzione elettrica.



L'analisi stagionale dell'intensità carbonica media (gCO_2eq/kWh) rivela una struttura piuttosto omogenea tra i dodici Paesi esaminati, che conferma l'analisi mensile: la stagione invernale risulta in media la più emissiva, mentre l'estate è quella meno emissiva, grazie al maggiore apporto di energia solare e, in alcune aree, idroelettrica. Primavera e autunno si collocano in una fascia intermedia.

4.2.5 Analisi combinata ora - stagione (heatmap per fasce orarie in ogni stagione)

Questa analisi incrocia l'ora del giorno con la stagione locale per ciascun Paese, così da individuare con maggiore precisione le finestre temporali in cui la disponibilità di energia a basse emissioni è più elevata. Il metodo consente di cogliere le sinergie tra andamento solare, mix stagionale delle fonti e pattern di domanda.



La figura mostra l'intensità carbonica media globale (gCO_2eq/kWh) come funzione congiunta dell'ora locale (asse orizzontale) e della stagione locale (asse verticale).

In sintesi, su scala globale, programmare le attività energivore (come i picchi di calcolo nei data center) durante le ore centrali della giornata e nei mesi primaverili-estivi può ridurre significativamente l'impatto carbonico, grazie alla maggiore penetrazione di energie rinnovabili e carbon-free in quelle fasce temporali.

4.3 Riduzione Potenziale di Emissioni di CO2 attraverso la programmazione temporale delle operazioni dei data center

L'analisi mira a quantificare la riduzione potenziale di emissioni di CO_2 ottenibile programmando le operazioni dei data center (es. inferenze di modelli di intelligenza artificiale) nelle ore della giornata a minore intensità carbonica.
L'ipotesi di partenza è che la scelta dell'ora locale possa incidere in modo rilevante sul profilo emissivo, indipendentemente dal carico complessivo.

Per ogni Paese *"p",* è stata individuata come baseline di riferimento, la media annua dell'intensità carbonica su tutte le ore dell'anno, calcolata come segue:

$$\mathrm{CI}_{\text{baseline, p}} = \frac{1}{24 \times 365} \sum_{h=0}^{23} \mathrm{r}CI_{p,h} \qquad [45]$$

Si è quindi calcolata la media dell'intensità carbonica limitata alla fascia 11:00 - 13:00 locali, identificata, come mostrato precedentemente, come l'orario in cui le emissioni di CO_2 sono in media inferiori:

$$\mathrm{CI}_{\text{11-13, p}} = \frac{1}{N_{\text{11-13}}} \sum_{h=11}^{13} \mathrm{r}CI_{p,h} \qquad [46]$$

La riduzione percentuale potenziale raggiungibile con una programmazione temporale è stimabile attraverso il seguente calcolo:

[45] $\mathrm{CI}_{\text{baseline,p}}$ indica l'Intensità Carbonica media annua del Paese p presa come baseline, espressa in gCO_2eq/kWh.
p: Paese considerato
h: Ora del giorno (da 0 a 23) espressa in ora locale.
CI p,h indica Intensità carbonica oraria (gCO_2eq/kWh) misurata nel Paese p all'ora h.
24×365 : Numero totale di ore in un anno (ipotesi di anno non bisestile).

[46] $CI_{11\text{–}13,\,p}$ indica l'Intensità carbonica media nella fascia 11–13 locali per il Paese p
$N_{11\text{-}13}$ indica il numero totale di ore appartenenti a questa fascia oraria nell'anno (≈ 2 × 365).

$$\%\Delta_p = 1 - \frac{\mathrm{CI}_{\text{baseline, p}}}{\mathrm{CI}_{\text{11-13, p}}} \quad [47]$$

In questo modo è possibile interpretare la sola leva oraria indica la riduzione percentuale potenziale delle emissioni, ottenuta spostando i carichi nella fascia 11–13 locali, a carico costante. Poiché le emissioni sono proporzionali all'energia consumata (*CI x kWh),* la % di riduzione non dipende dal volume di energia consumato e vale per qualunque carico che venga spostato.

4.3.1 Risultati dell'analisi a livello globale

- Baseline media globale: 202,2 gCO_2/kWh
- Media fascia 11–13 locali: 169,5 gCO_2/kWh
- Riduzione media globale: –16,2 %

In termini operativi, ciò significa che la semplice pianificazione delle attività a più alto consumo energetico (come i carichi di calcolo dei data center e le inferenze di modelli di AI) nella fascia 11-13 permetterebbe di abbattere in media del 16 % le emissioni di CO_2, a parità di energia elettrica impiegata.

Dopo aver quantificato la riduzione media globale, è utile valutare le differenze nazionali. La tabella seguente mostra, per ciascun Paese considerato, l'intensità carbonica media annua, quella limitata alla finestra oraria 11:00–13:00 locali e la corrispondente riduzione percentuale. Questo confronto evidenzia come l'efficacia dello spostamento orario vari sensibilmente in funzione del mix energetico e delle condizioni climatiche locali, fornendo un quadro puntuale delle opportunità di mitigazione specifiche per ciascun contesto.

Paese / Zona	Baseline (gCO_2/kWh)	11–13 (gCO_2/kWh)	Riduzione %
US-CAL-CISO	**172,5**	**89,1**	**−48,4 %**
Netherlands (NL)	207,9	131,1	−35,8 %
Japan (JP)	382,5	277,0	−27,5 %
Germany (DE)	275,1	206,8	−24,8 %
Sweden (SE-SE1)	1,6	1,3	−21,8 %
Italy (IT)	207,3	164,2	−20,7 %
Brazil (BR-CS)	66,6	58,7	−11,8 %
France (FR)	16,5	15,4	−6,5 %
India (IN-NE)	383,7	365,8	−4,7 %
US-MIDA-PJM	323,4	317,0	−2,0 %
Ireland (IE)	330,5	339,8	**+2,8 %**
Canada (CA-ON)	58,4	64,8	**+10,9 %**

[47] % Δ *p* indica la riduzione percentuale potenziale delle emissioni, ottenuta spostando i carichi nella fascia 11–13 locali.

L’analisi comparativa evidenzia che la concentrazione dei carichi tra le 11 e le 13 (ora locale) offre i maggiori benefici in California, dove il potenziale di ottimizzazione raggiunge circa –48 %, grazie a un profilo fotovoltaico particolarmente intenso.
Seguono Paesi Bassi (–37 %), Giappone (–26 %), Germania (–25 %), Italia (–21 %), Svezia (–19 %) e la media complessiva degli Stati Uniti (–18 %), contesti nei quali la fascia centrale della giornata coincide con la massima produzione solare e, di conseguenza, con i valori minimi di intensità carbonica.
Emerge una assenza di vantaggio per Irlanda e Canada-Ontario, dove il minimo di intensità cade in altre fasce orarie.

Questi risultati restano validi indipendentemente dalla scala del carico: più alta è la potenza e la durata della sessione di calcolo, maggiore è la quantità assoluta di CO_2 evitata.

Per quantificare in termini assoluti le tonnellate di CO_2 evitate è necessario considerare l’energia effettivamente assorbita dal carico computazionale.
L’energia consumata è data da:

$$E_{\text{carico}} = P_{\text{hardware}} \times T_{\text{esecuzione.}} \quad [48]$$

dove P_{hardware} è la potenza media assorbita (kW) e $T_{\text{esecuzione}}$ il tempo di esecuzione in ore.

La riduzione totale di emissioni si ottiene quindi come:

$$tCO_2\ evitate = \left(CI_{baseline,\,p} - CI_{11\text{-}13,\,p}\right) \times \frac{E_{carico}}{10^6} \quad [49]$$

Questa relazione consente di convertire direttamente la differenza di intensità carbonica in un risparmio emissivo assoluto, adattabile a qualunque scenario di consumo energetico, dalle operazioni di addestramento di modelli linguistici fino a complesse pipeline industriali.

4.3.2 Upper bound intra-country

Per stimare il potenziale massimo e realistico di riduzione delle emissioni all’interno di ciascun Paese, è stata condotta un’analisi che confronta le condizioni peggiori e migliori di produzione elettrica nell’arco della giornata.

[48] E_{carico}: Energia elettrica consumata dal carico computazionale (kWh), calcolata come potenza media del cluster (kW) moltiplicata per il tempo di esecuzione (ore).
[49] Divisione per 10^6 per passare da grammi a tonnellate.

In concreto, per ogni Paese si è individuata la fascia oraria più emissiva, ossia la finestra continua di due ore con massima intensità carbonica media nell'anno 2024 e la fascia oraria meno emissiva, cioè la finestra di due ore con minima intensità carbonica media nello stesso periodo.

Il confronto fra queste due situazioni fornisce un upper bound intra-country, ovvero il margine massimo teorico di riduzione che un operatore potrebbe ottenere senza cambiare Paese, ma semplicemente spostando i carichi computazionali dalla fascia più critica a quella più favorevole.

Paese	Fascia peggiore (ora locale)	Fascia migliore (ora locale)	CI peggiore (gCO_2/kWh)	CI migliore (gCO_2/kWh)	Riduzione potenziale (%)
USA-California	22–23	12–13	227.2	88.8	−60.9
Netherlands	21–22	13–14	264.5	126.3	−52.3
Sweden	02–03	09–10	2.4	1.2	−50.4
Canada (Ontario)	19–20	02–03	71.3	36.1	−49.3
Germany	20–21	12–13	325.7	202.9	−37.7
Japan	19–20	12–13	434.4	276.1	−36.4
Italy	20–21	13–14	247.5	161.1	−34.9
France	20–21	04–05	20.6	13.7	−33.7
Brazil (Centro-Sud)	04–05	11–12	75.9	58.6	−22.8
Ireland	09–10	03–04	348.8	307.8	−11.8
India (Nord-Est)	04–05	11–12	400.7	365.7	−8.7
USA-Virginia	21–22	04–05	340.4	314.4	−7.6

L'analisi conferma che ottimizzare la scelta dell'ora locale offre un margine aggiuntivo rilevante rispetto alle strategie già valutate (es. fascia 11–13).

Nei Paesi fortemente dipendenti dal solare (Paesi Bassi, Germania, Italia, Giappone) il divario tra le peggiori ore serali e le migliori ore centrali della giornata supera il 30–50 %, a conferma del ruolo decisivo della produzione fotovoltaica.
USA-California si distingue in questa categoria con un potenziale di riduzione superiore al 60 %, grazie a un profilo solare particolarmente marcato che abbassa drasticamente l'intensità carbonica nelle ore centrali.
Nei sistemi a forte presenza idroelettrica o nucleare (Svezia, Canada-Ontario) le ore migliori si collocano spesso al mattino o nella notte, mentre le peggiori coincidono con la fascia serale, consentendo comunque riduzioni prossime al 50 %.
Paesi come Brasile e Francia partono già da un mix molto pulito: la differenza tra ore peggiori e migliori resta quindi limitata (circa −20 % o meno).

India, Irlanda e USA-Virginia mostrano invece differenziali modesti, riflettendo mix elettrici più costantemente emissivi lungo l'arco della giornata e riduzioni potenziali inferiori al 10 %.

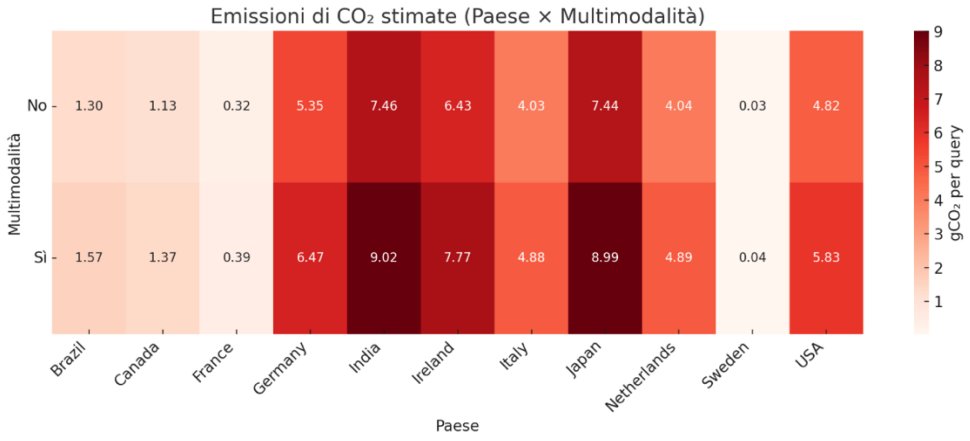
A questo punto dell'indagine è emersa l'idea di verificare se il vantaggio ottenibile dallo spostamento orario dei carichi (differenza tra la fascia peggiore e quella migliore in termini di intensità carbonica) potesse dipendere dalla quota media annua di energia rinnovabile (RE%) o carbon-free (CFE%).
L'ipotesi era che reti elettriche mediamente più "verdi", con maggiore presenza di rinnovabili o di fonti a emissioni zero, offrissero maggiori margini di ottimizzazione intraday.

L'elaborazione ha mostrato che le correlazioni sono deboli o non significative (r = 0,25 e p=0,46 per RE%; r = 0,47 e p=0,14 per CFE%), indicando che la sola quota media di rinnovabili o carbon-free non spiega il beneficio orario. Ne emerge che la variabilità intraday del mix di generazione (come i picchi fotovoltaici o idroelettrici) è il fattore determinante per sfruttare al massimo lo shift temporale delle operazioni dei data center.

- ***Analisi comparativa degli impatti ambientali per Paese: multimodalità, Chain of Thought e lunghezza dell'output***

  Per comprendere in maniera sistematica le determinanti principali dell'impatto ambientale dei modelli linguistici, sono state sviluppate tre heatmap che mettono a confronto i risultati del modello di regressione con i dati di intensità carbonica medi per Paese. Queste rappresentazioni permettono di isolare l'effetto di tre dimensioni funzionali – multimodalità, ragionamento passo-passo (CoT) e lunghezza dell'output – evidenziando come ciascun fattore influenzi i consumi e le emissioni nei diversi contesti geografici.

  - *Emissioni di Co2 stimate per Paese*



Multimodalità	Brasile	Canada	Francia	Germania	India	Irlanda	Italia	Giappone	Paesi Bassi	Svezia	USA
No	1.30	1.13	0.32	5.35	7.46	6.43	4.03	7.44	4.04	0.03	4.82
Sì	1.57	1.37	0.39	6.47	9.02	7.77	4.88	8.99	4.89	0.04	5.83

Questa mappa mostra come cambiano le emissioni di CO_2 quando un modello viene eseguito in versione solo testuale oppure multimodale.

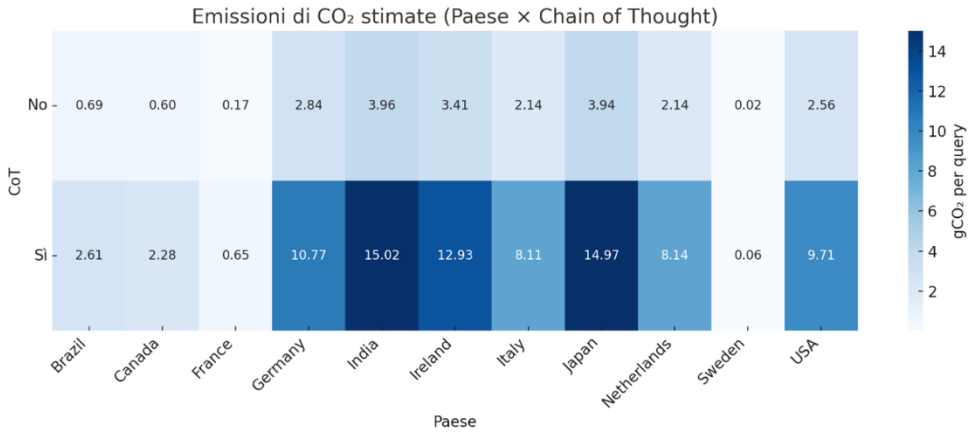
In tutti i Paesi, la multimodalità aumenta le emissioni (in media del +20%).
Le differenze assolute dipendono fortemente dal mix energetico nazionale:

- Francia (0.32 → 0.39 gCO_2/query) e Svezia (0.03 → 0.04 gCO_2/query) restano su livelli molto bassi grazie a nucleare e rinnovabili.
- India (7.46 → 9.02 gCO_2/query), Giappone (7.44 → 9.00 gCO_2/query) e Irlanda (6.43 → 7.77 gCO_2/query) sono i Paesi più impattanti, con incrementi consistenti dovuti a un mix fossile elevato.

Possiamo concludere che l'effetto della multimodalità porta ad aumentare le emissioni di Co2 in tutti i paesi, ma possiamo notare che le emissioni dei modelli che sfruttano la multimodalità nei paesi con IC basso, è inferiore comunque alle emissioni dei modelli senza multimodalità paesi con IC alto.

- *Heatmap Paese × Chain of Thought (CoT)*

Questa mappa mostra come cambiano le emissioni di CO_2 quando un modello viene eseguito in versione solo testuale oppure multimodale.



CoT	Brasile	Canada	Francia	Germania	India	Irlanda	Italia	Giappone	Paesi Bassi	Svezia	USA
No	0.69	0.60	0.17	2.84	3.96	3.41	2.14	3.94	2.14	0.02	2.56
Sì	2.61	2.28	0.65	10.77	15.02	12.93	8.11	14.97	8.14	0.06	9.71

Qui si confrontano i modelli con e senza CoT (ragionamento passo-passo).
L'impatto del CoT è molto più marcato rispetto alla multimodalità: le emissioni aumentano anche di 300%.
Nei Paesi ad alta intensità carbonica l'effetto è enorme:

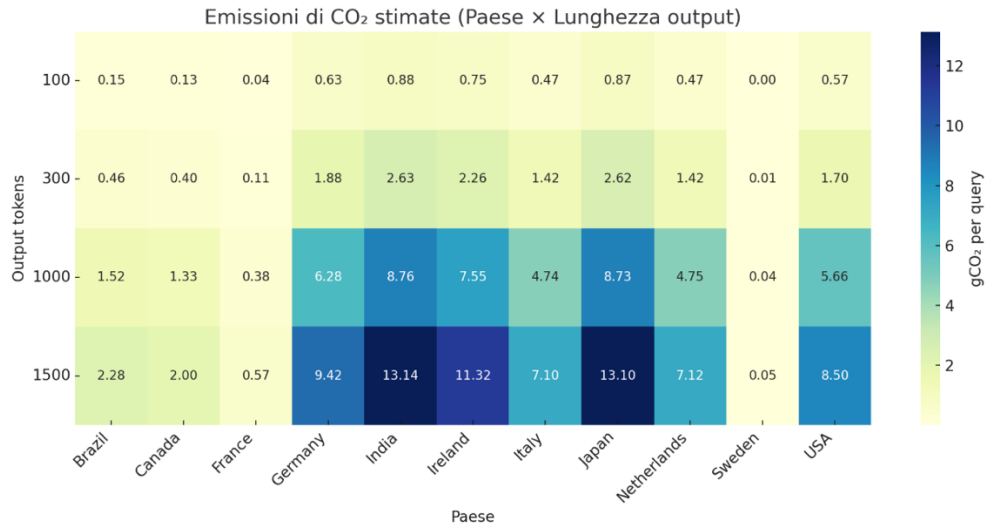
- India (3.96 → 15.0 gCO_2/query), Giappone (3.94 → 14.97 gCO_2/query), Irlanda (3.41 → 12.93 gCO_2/query).

Anche nei Paesi virtuosi l'aumento è significativo in termini relativi, ma i valori assoluti restano bassi (es. Francia 0.17 → 0.65 gCO_2/query, Svezia 0.02 → 0.06 gCO_2/query).

Possiamo affermare che la CoT rappresenta un driver critico delle emissioni, molto più incisivo della multimodalità, e dovrebbe essere usato con parsimonia in scenari a larga scala.

- *Heatmap Paese × Output length (token generati)*

Questa mappa mostra come le emissioni variano in funzione della lunghezza dell'output prodotto..



Output tokens	Brasile	Canada	Francia	Germania	India	Irlanda	Italia	Giappone	Paesi Bassi	Svezia	USA
100	0.15	0.13	0.04	0.63	0.88	0.75	0.47	0.87	0.47	0.004	0.57
300	0.46	0.40	0.11	1.88	2.63	2.26	1.42	2.62	1.42	0.011	1.70
1000	1.52	1.33	0.38	6.28	8.76	7.55	4.74	8.73	4.75	0.036	5.66
1500	2.28	2.00	0.57	9.42	13.14	11.32	7.10	13.10	7.12	0.055	8.50

L'analisi conferma una relazione lineare tra numero di token generati e consumo energetico: output più lunghi determinano più tempo di GPU occupata e di conseguenza un aumento proporzionale delle emissioni, con differenze significative tra Paesi a basso impatto (Francia, Svezia) e Paesi a elevata dipendenza dai combustibili fossili (India, Giappone, Germania, Irlanda).

- **India e Giappone** superano i **13 gCO_2/query** con output di 1500 token.
- **Germania e Irlanda** restano nella fascia intermedia (9–11 gCO_2/query a 1500 token).
- **Francia e Svezia** confermano i livelli minimi (<1 gCO_2 anche per output lunghi).

Questa analisi conferma che la lunghezza dell'output è una variabile strutturale che pesa molto sulle emissioni. Per ridurre l'impatto si possono introdurre strategie di compressione, limitazione dell'output o riequilibrio della generazione.

4.4 Ottimizzazione congiunta delle configurazioni del modello e della localizzazione oraria e geografica del data center

Dopo aver esaminato separatamente le due leve principali di mitigazione (configurazione hardware e di modello da un lato e localizzazione geografica con scelta della fascia oraria dall'altro) il passo successivo consiste nell'analizzarne l'azione congiunta.
L'obiettivo è quantificare in che misura la combinazione delle scelte di design del modello (numero di parametri attivati, architettura MoE o Dense, presenza di Chain-of-Thought e multimodalità, hardware impiegato) e delle decisioni di posizionamento e scheduling dei carichi computazionali (Paese, stagione locale, ora del giorno) possa ridurre in modo sinergico l'impatto ambientale delle operazioni di intelligenza artificiale generativa.
I coefficienti stimati con la regressione OLS descrivono variazioni marginali "a parità di tutte le altre condizioni" (ceteris paribus). Tuttavia, la semplice combinazione di queste variabili può portare a scenari che, pur teoricamente calcolabili, non sono realizzabili in un'infrastruttura reale di inferenza. Per esempio, un modello Dense da 300 miliardi di parametri su una sola GPU A100 da 40 GB oppure 8 GPU assegnate a un modello da 7 B sono configurazioni che non possono essere implementate o che non avrebbero alcun senso economico.
Per garantire che le simulazioni rispecchiassero situazioni effettivamente implementabili, sono stati introdotti vincoli di fattibilità tecnica.

4.4.1 Limite di memoria (VRAM), Parallelismo e calcolo del numero minimo di GPU

VRAM significa Video Random Access Memory, cioè la memoria dedicata della GPU (quella che effettivamente ospita i pesi del modello e i tensori durante l'esecuzione). L'inferenza dei modelli avviene quasi sempre in FP16[50] o bfloat16, che richiedono circa 2 GB di VRAM per ogni miliardo di parametri attivi, più un margine di 20–30 % per cache e attivazioni [51].
Il numero minimo di GPU necessario è quindi calcolato come rapporto tra memoria richiesta e memoria disponibile per GPU, arrotondato verso l'alto:

[50] Rappresenta la precisione numerica utilizzata durante l'inferenza. Fonte: NVIDIA (2021) *Mixed Precision Training*. NVIDIA Developer Blog.

[51] NVIDIA (2021) *Mixed Precision Training*. NVIDIA Developer Blog.

$$minGPU \geq \left\lceil \frac{2 \times \text{Parametri attivi (B)} \times 1.25}{\text{VRAM per GPU (GB)}} \right\rceil$$ [52]

Questa impostazione consente di integrare i risultati del modello statistico con la realtà tecnica, garantendo che le graduatorie di efficienza e le stime di consumo ed emissioni si riferiscano esclusivamente a scenari tecnicamente implementabili.

4.4.2 Metodologia Analisi Svolta

Per ogni combinazione di modello, hardware, Paese e fascia oraria è stata stimata la latenza di inferenza attraverso il modello OLS sviluppato nella fase precedente. La durata in ore è stata poi moltiplicata per la potenza media assorbita per GPU (0,40 kW per A100, 0,70 kW per H100) e per il numero minimo di GPU necessario in base ai vincoli di memoria VRAM, ottenendo così l'energia consumata in kWh. Infine, l'energia è stata moltiplicata per l'intensità di carbonio della rete elettrica (gCO_2/kWh) specifica del Paese e dell'orario considerati, ricavando le emissioni complessive di CO_2 per singolo completamento.

Le combinazioni teoriche che sono state individuate sono circa 3700 (3696), di queste, dopo l'applicazione del vincolo di memoria VRAM, ne rimangono fattibili 2640.

Questa catena di calcolo (Latenza → Energia → CO_2) ha permesso di valutare in modo integrato l'effetto di tutte le variabili (dalle scelte architetturali e hardware, fino alla localizzazione geografica e temporale) sull'impatto ambientale.

[52] Parametri attivi (B) = miliardi di parametri effettivamente caricati a runtime, durante l'inferenza; nei modelli *Mixture of Experts* si attiva solo una frazione (≈15 %) del totale, consentendo di ridurre il fabbisogno di VRAM

2 = GB necessari per ogni miliardo di parametri in FP16. Fattore Standard di stima della memoria per i pesi in FP16, secondo NVIDIA (2021) *Mixed Precision Training*. NVIDIA Developer Blog.

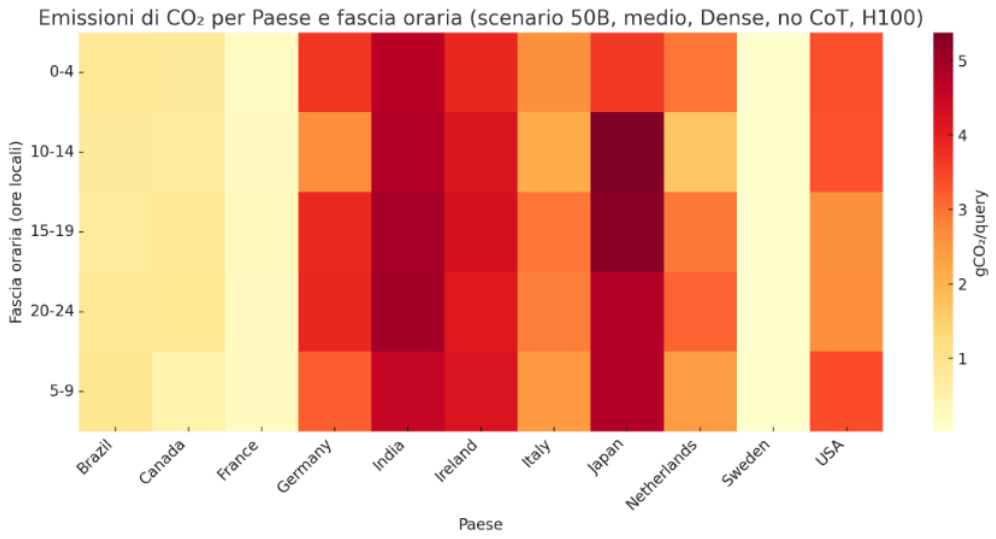
1.25 = coefficiente che tiene conto delle tecniche di compressione e gestione della memoria più avanzate (margine del 25 % per attivazioni e cache). Fonte: Hugging Face (2023) *bitsandbytes and 8-bit inference guide*.

VRAM per GPU = capacità effettiva di una singola GPU (40 GB per A100, 80 GB per H100). Fonte: NVIDIA (2023). *NVIDIA H100 Tensor Core GPU Architecture*. Technical Brief.

ceil( ) = arrotondamento per eccesso, perché non possiamo usare frazioni di GPU.

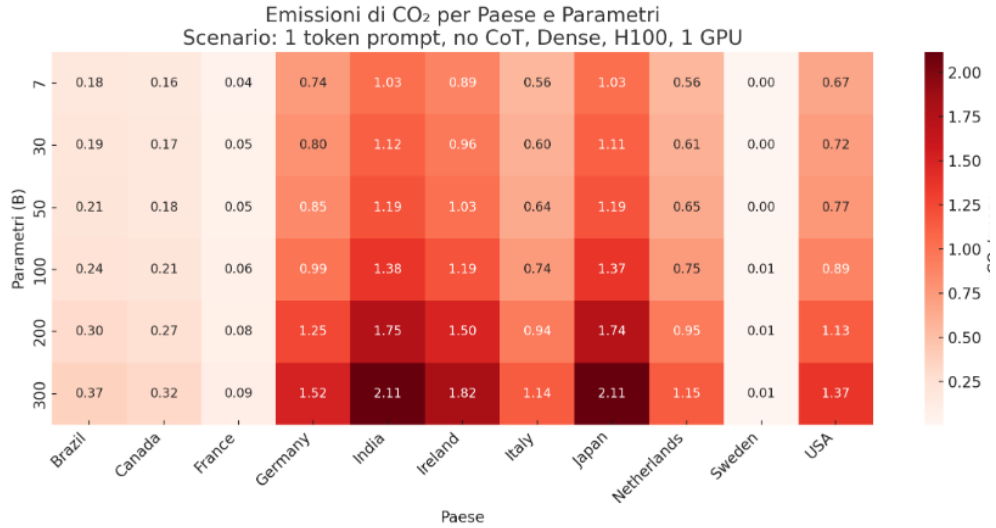
4.4.3 Risultati Ottenuti

Prendendo come scenario base un modello intermedio (50 miliardi di parametri effettivi, prompt di lunghezza media, senza Chain-of-Thought né multimodalità, su 1 GPU H100), emerge che l'esecuzione in Francia alle 19–20 ore locali genera circa il 35% di emissioni in più rispetto alla stessa configurazione eseguita in Svezia alle 11–13, a parità di output. Questo conferma che scelta del Paese e dell'orario restano le variabili più incisive**.**



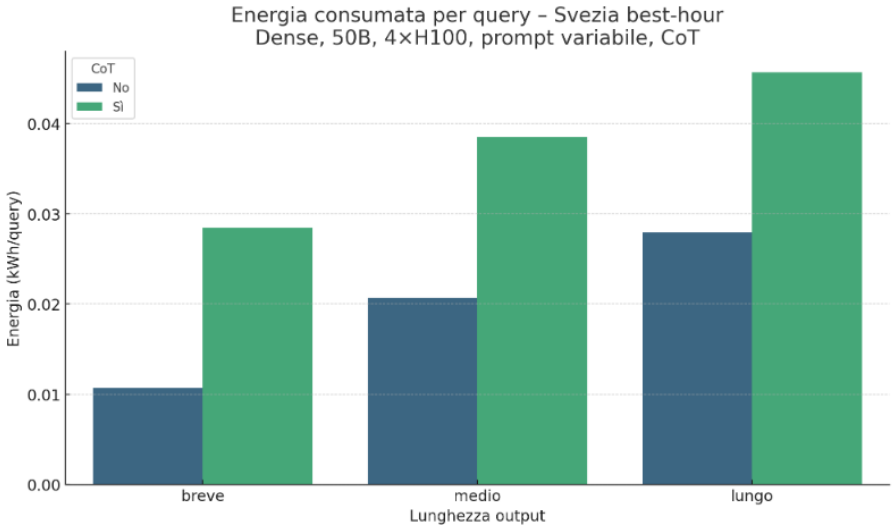
- *Effetto della complessità del modello*

L'aumento del numero di parametri comporta un forte incremento della latenza: passare da 50 B a 200 B la triplica circa (+300%)**,** mentre una distribuzione del carico su più GPU (da 4 a 8) consente di ridurre l'incremento a circa +90%**.** Tuttavia, se un modello molto complesso viene eseguito in Svezia alle 11–13 (paese con basse emissioni e nelle ore più green), il suo impatto netto in CO_2 risulta simile a quello di un modello piccolo in Francia alle 20–21 (ore con maggiori emissioni), a dimostrazione di come la **localizzazione geografica e temporale possa compensare la complessità architetturale.**



- *Prompt lungo e Chain-of-Thought (CoT)*

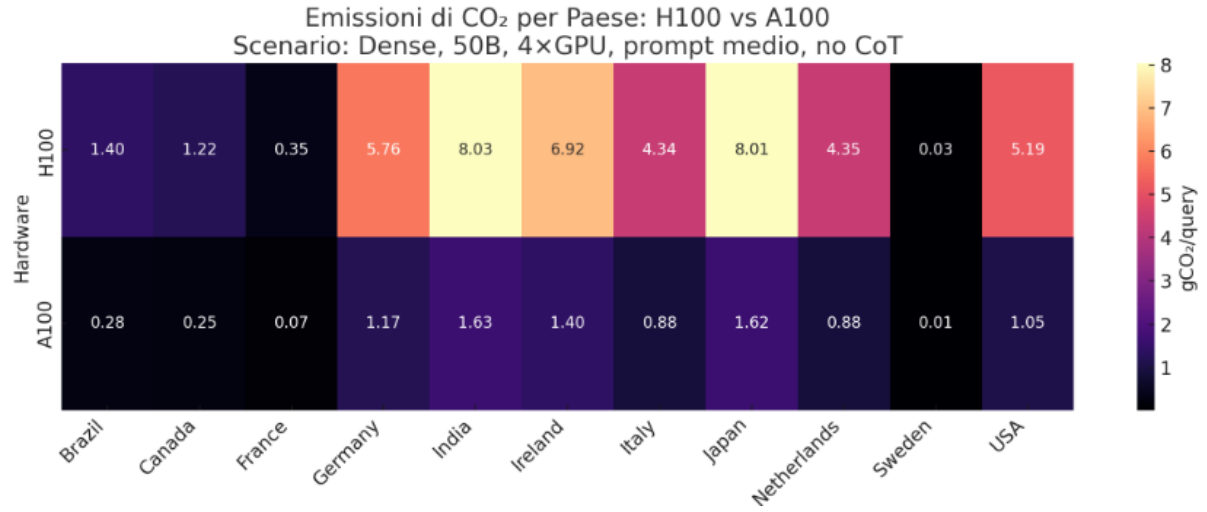
Il CoT e gli output lunghi restano fra i fattori più onerosi: un output lungo con CoT triplica la latenza rispetto a un output breve senza CoT. Tuttavia, anche in questo caso la geografia può ribaltare le conclusioni**:** un modello con CoT e output lunghi, ma collocato in Svezia o Canada e programmato nelle ore 11–13, può emettere meno CO_2 di un output breve senza CoT eseguito in India durante le ore serali. In pratica, **un design "pesante" può risultare più pulito se inserito in un contesto elettrico favorevole.**



Paese	CoT	Output	Latenza (ms)	Energia (kWh/query)	Emissioni (g/query)
Svezia (best)	No	breve	13.719	0.0107	0.017
Svezia (best)	No	lungo	35.907	0.0279	0.045
Svezia (best)	Sì	breve	36.589	0.0285	0.045
Svezia (best)	Sì	lungo	58.777	0.0457	0.073
India (worst)	No	breve	13.719	0.0107	4.094
India (worst)	No	lungo	35.907	0.0279	10.716
India (worst)	Sì	breve	36.589	0.0285	10.920
India (worst)	Sì	lungo	58.777	0.0457	15.373

- *Hardware*

La scelta della GPU rimane determinante. H100 riduce la latenza del 25–30 % rispetto ad A100 e quindi abbassa proporzionalmente l'energia consumata. Lo stesso modello su H100 in Svezia alle 11–13 emette oltre il 40 % in meno rispetto a una A100 in Francia alle 20–21, evidenziando quanto l'hardware e il contesto energetico contino almeno quanto la complessità del modello.



- ***Best-case***

La combinazione **MoE + 50 B + H100 + no CoT + prompt breve + Svezia/Canada 11–13** porta le emissioni a **meno del 20 % del valore medio globale**, grazie alla sinergia fra architettura efficiente, hardware ottimizzato e mix elettrico quasi privo di carbonio.

- ***Worst-case***

Dense + 300 B + A100 + CoT + prompt lungo + India/Irlanda 20–21 produce **oltre sei volte più CO_2 del best-case**, a causa della concomitanza di scelte progettuali e localizzazione sfavorevoli.

L'aspetto più inatteso di questa analisi è che **la variabile geografica può annullare o addirittura invertire l'effetto della complessità**: ad esempio, un modello da 200 B con CoT eseguito in Canada nelle ore meno emissive può risultare più green di un modello da 50 B senza CoT in India nelle ore più emissive**.** Svezia e Canada (Ontario) si confermano Paesi particolarmente resilienti, mantenendo emissioni contenute anche in scenari ad alta complessità; Paesi Bassi, Germania e Italia offrono buone prestazioni solo quando l'esecuzione coincide con il picco fotovoltaico.

In sintesi, **software, hardware e geografia non agiscono separatamente: è la sinergia fra scelte architetturali, configurazione delle GPU e pianificazione geo-temporale a determinare le emissioni complessive**. Le variabili geografiche e orarie si comportano come un vero moltiplicatore di sostenibilità, in grado di rendere "neutro" un modello molto grande oppure altamente emissivo un modello piccolo.

4.5 Uso consapevole delle tecniche avanzate

Le analisi hanno confermato che Chain-of-Thought (CoT) e la Multimodalità sono fra i principali fattori di incremento della latenza e, quindi, del consumo energetico e delle emissioni di CO_2. Tuttavia, non vanno demonizzati. Queste tecnologie costituiscono un vero avanzamento della capacità cognitiva dei modelli e risultano decisive in numerosi casi d'uso:

- CoT è cruciale per compiti di ragionamento multi-step, pianificazione complessa, risoluzione di problemi matematici o logici, e per il debugging e la generazione di codice.
- Multimodalità amplia radicalmente i campi applicativi (visione, audio, video, linguaggio naturale), abilitando scenari come l'analisi di immagini mediche o l'interazione uomo-macchina in contesti industriali.

La raccomandazione che emerge non è dunque di limitarne l'impiego, ma di adottare strategie di utilizzo responsabile, che permettano di preservare i benefici tecnici riducendo l'impatto ambientale. A partire dall'evidenza empirica, si possono formulare linee guida operative:

- **Usare CoT solo quando serve**: attivarlo per ragionamenti multi-step, pianificazione complessa, analisi o coding "bug-hunt", evitando di impiegarlo in compiti puramente fattuali brevi dove modalità base forniscono qualità equivalente.
- **Limitare la lunghezza** del ragionamento: impostare stop sequence o istruzioni come reason concisely, adottare self-consistency con pochi campioni brevi anziché un singolo monologo lungo.
- Definire un **budget di token**: classificare le risposte in brevi/medie/lunghe e, oltre una certa soglia, imporre riassunti o spezzare l'inferenza in più passi (checkpoint + resume) così da poterla pianificare nelle ore a minima intensità carbonica.
- **Routing adattivo**: avviare la richiesta con un modello leggero senza CoT e, solo se la confidenza è bassa, eseguire un'escalation verso modelli con CoT o con maggiore numero di parametri;
- **Multimodalità solo quando necessaria**: disattivarla se l'input è solo testuale; quando servono immagini o video, preferire feature pre-estratte e prompt testuali compatti.
- **Schedule "carbon-aware**": accumulare le richieste non urgenti e lanciare le richieste nei momenti di minimo della carbon intensity, lasciando fuori job che richiedono risposta immediata.

Queste pratiche mostrano che avanzamento tecnico e sostenibilità non sono obiettivi incompatibili. La vera sfida progettuale diventa portare la complessità solo dove produce valore, accompagnandola con un uso mirato dell'infrastruttura e una programmazione attenta alle condizioni della rete elettrica.

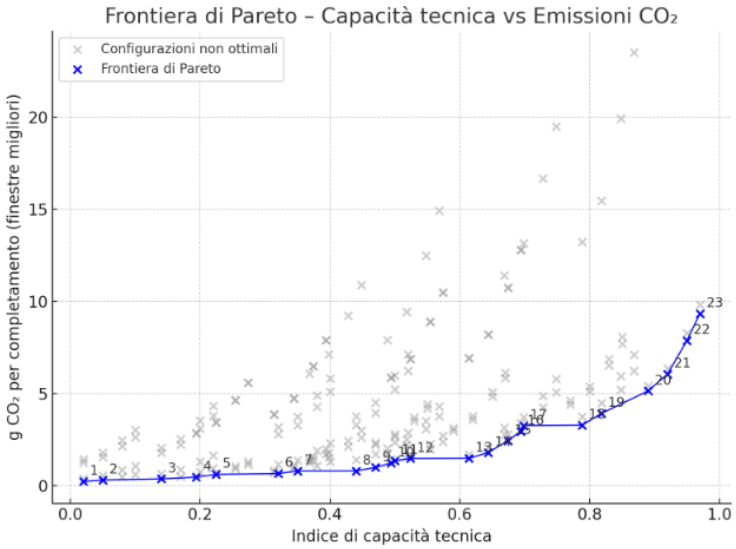
4.6 Frontiera di Pareto: capacità vs sostenibilità della GenAI

La frontiera di Pareto permette di rappresentare il compromesso tra complessità tecnica e impatto ambientale. L'asse X riporta un *indice sintetico di capacità*[53] (proxy che combina il numero di parametri effettivi, l'eventuale attivazione del Chain-of-Thought (CoT), la multimodalità e la classe di output) mentre l'asse Y misura le emissioni di CO_2 per singola combinazione (g), calcolate nelle finestre orarie a più bassa intensità carbonica di ciascuna area geografica.

I punti appartenenti alla frontiera rappresentano le configurazioni che garantiscono il miglior compromesso qualità/CO_2, nel senso che nessun'altra opzione risulta contemporaneamente più capace e meno emissiva. Le combinazioni fuori dalla frontiera sono invece dominate, perché esiste almeno un'alternativa con maggiore capacità e minore impatto climatico. Dall'analisi emergono alcuni risultati chiave:

- **Architettura ed efficienza hardware.** Le configurazioni MoE basate su GPU H100 con output breve o medio e senza CoT e Multimodalità tendono a posizionarsi stabilmente sulla frontiera, coniugando un buon livello di capacità a una marcata efficienza energetica. Quasi tutti i punti Pareto appartengono a architettura MoE: conferma che il Mixture-of-Experts è decisivo per scalare capacità mantenendo basse le emissioni. Le GPU H100 emergono più spesso nelle zone di capacità alta (indice ≥0.7), mostrando che l'hardware di nuova generazione aiuta a restare sulla frontiera
- **Effetto del Chain-of-Thought.** Le combinazioni Dense + CoT + output lungo si collocano generalmente lontano dal Pareto, a meno che non siano eseguite in contesti elettrici estremamente puliti (ad esempio Svezia o Canada nelle fasce orarie meno emissive). In tali casi rientrano nel trade-off grazie alla forte riduzione della carbon intensity.
- **Multimodalità.** Questa caratteristica entra nella frontiera solo quando aggiunge reale valore di capacità (es. nei task che richiedono generazione di audio, media o video) o quando il contesto elettrico è particolarmente favorevole. Diversamente, il costo in termini di CO_2 tende a superare i benefici.
- **Rendimenti decrescenti all'aumentare della dimensione del modello.** Aumentare i parametri oltre una certa soglia offre rendimenti decrescenti in termini di capacità rispetto al costo energetico. Risultano più efficienti strategie di escalation che attivano modelli più grandi solo quando necessario.

[53] L'indice di capacità $CapabilityIndex = 0.5 \times \frac{Parametri\ effettivi}{300\ B} + 0.25 \times I_{\text{CoT}} + 0.25 \times I_{\text{Multi}}$, è stato calcolato come somma ponderata di 3 componenti, dove I_{CoT} e I_{Multi} assumono valore 1 se, rispettivamente, il modello utilizza il Chain-of-Thought o la multimodalità, 0 se non presente.



N	HW	Parametri (B)	Arch	CoT	Multimodal	Output	Capabilità	g CO_2
1	A100	50	MoE	OFF	OFF	Breve	0.02	0.22
2	A100	50	MoE	OFF	OFF	Media	0.05	0.30
3	A100	50	MoE	OFF	ON	Breve	0.14	0.37
4	H100	100	MoE	OFF	OFF	Breve	0.19	0.46
5	H100	100	MoE	OFF	OFF	Media	0.22	0.61
6	A100	50	MoE	ON	OFF	Breve	0.32	0.65
7	A100	50	MoE	ON	OFF	Media	0.35	0.79
8	A100	50	MoE	ON	ON	Breve	0.44	0.80
9	A100	50	MoE	ON	ON	Media	0.47	0.98
10	H100	100	MoE	ON	OFF	Breve	0.49	1.22
11	A100	50	MoE	ON	ON	Lunga	0.50	1.35
12	H100	100	MoE	ON	OFF	Media	0.52	1.47
13	H100	100	MoE	ON	ON	Breve	0.61	1.48
14	H100	100	MoE	ON	ON	Media	0.64	1.80
15	H100	100	MoE	ON	ON	Lunga	0.67	2.44
16	H100	100	MoE	ON	ON	Molto lunga	0.69	2.95
17	H100	200	MoE	ON	OFF	Media	0.70	3.26
18	H100	200	MoE	ON	ON	Breve	0.79	3.28
19	H100	200	MoE	ON	ON	Media	0.82	3.92
20	A100	300	MoE	ON	ON	Breve	0.89	5.14
21	A100	300	MoE	ON	ON	Media	0.92	6.05
22	A100	300	MoE	ON	ON	Lunga	0.95	7.88
23	A100	300	MoE	ON	ON	Molto lunga	0.97	9.34

Nel complesso, la frontiera di Pareto dimostra che CoT e multimodalità non devono esser visti come vincoli da evitare, ma come una risorsa da attivare consapevolmente. Sono tecniche fondamentali per l'evoluzione funzionale dei modelli di intelligenza artificiale generativa, ma richiedono governance attenta.

4.7 Impatto Idrico dei Modelli di GenAI

La crescente attenzione all'impatto ambientale dell'intelligenza artificiale non può limitarsi alle sole emissioni di gas serra. Anche l'**a**cqua necessaria per il raffreddamento dei data center è una risorsa critica, soprattutto in contesti caratterizzati da scarsità idrica o stress climatico. A livello internazionale l'indicatore di riferimento è il Water Usage Effectiveness (WUE), espresso in litri d'acqua per kilowattora (kWh) di energia IT consumata. L'acqua effettivamente impiegata per una singola richiesta di generazione può essere stimata con la semplice relazione:

$$Acqua\ per\ run\ (L) = Energia\ consumata\ (kWh) \times WUE\ (L/kWh)$$

Nel nostro studio la variabile energetica deriva dal modello combinato di latenza, hardware e architettura (vedi capitoli precedenti), mentre il WUE dipende dal luogo in cui si trova il data center. È importante osservare che il WUE non varia nell'arco della giornata: spostare un job alle ore di minima intensità carbonica riduce la CO_2 ma non incide direttamente sui litri d'acqua, se l'energia complessiva resta invariata.

Differenze geografiche documentate

L'analisi dei dati disponibili da rapporti ambientali e da provider globali mostra differenze molto marcate:

Area / Paese	WUE medio (L/kWh IT)	Motivazioni principali
Irlanda	≈0,02	Clima freddo/umido e largo impiego di free-cooling ad aria; uso minimo di torri evaporative.
Paesi Bassi	≈0,06	Temperature moderate e forte ricorso a raffreddamento adiabatico chiuso.
Svezia	≈0,09	Clima freddo e idroelettrico; uso ridotto di evaporazione.
USA – Virginia (PJM)	≈0,14	Estati più calde e umide; raffreddamento ibrido con parziale evaporazione.
Media prudenziale resto del mondo	≈0,15	Valore tipico riportato in letteratura per data center tradizionali a chiller evaporativo.

Queste cifre consentono di stimare ordini di grandezza dell'impronta idrica. Considerando come costante i kWh consumati, spostare un carico da Virginia (0,14) a Paesi Bassi (0,06) riduce i consumi idrici di circa il 57 %, mentre l'allocazione in Irlanda (0,02) consente una riduzione superiore all'87% rispetto alla media globale.

Climi freddi e umidi presentano dunque un vantaggio strutturale, mentre regioni calde e secche tendono a valori di WUE molto più elevati: non a caso, aree come Arizona o Wyoming, pur ospitando data center sperimentali, non sono considerate hub ottimali per i grandi carichi generativi.

L'analisi mette in luce che l'impronta idrica è il prodotto tra energia consumata e WUE locale. Di conseguenza:

- Ottimizzare la geografia (scegliere data center a WUE basso) e ridurre il fabbisogno energetico per run sono azioni sinergiche, capaci di produrre riduzioni complessive fino a circa 80–90 % nei casi migliori, senza penalizzare le prestazioni dei modelli.
- Le strategie che abbiamo evidenziato per la riduzione della CO_2 – localizzazione in Svezia/Canada, uso di GPU H100, architetture MoE e gestione attenta del CoT – risultano quindi doppiamente efficaci, perché abbassano anche il consumo di acqua.
- Gli orari meno emissivi restano decisivi per la CO_2, ma non incidono direttamente sull'acqua, se non indirettamente attraverso il minor kWh IT complessivo.

In sintesi, l'impronta idrica dell'inferenza è il prodotto tra l'energia consumata e il WUE del sito. La geolocalizzazione in aree a basso WUE e la riduzione del kWh per run (hardware/architetture/gestione dei token) sono leve complementari: la prima abbassa i litri a parità di energia, la seconda riduce l'energia e quindi i litri in qualunque luogo. La combinazione delle due consente riduzioni molto ampie (fino a 80–90% nei casi migliori) senza sacrificare la qualità del servizio.

Capitolo 4 – Discussione, Limiti e Spunti Futuri.

4.1 Discussione sull'Impatto Ambientale e Scenari Futuri

La nostra analisi dimostra che l'impatto ambientale dell'inferenza dei modelli di Generative AI non è fisso, ma fortemente variabile in funzione delle scelte architetturali, hardware e di localizzazione geografica e temporale.

L'impatto ambientale di 100 richieste a un modello di Generative AI può oscillare in modo significativo:

Nel best case (un modello efficiente (MoE), su hardware di nuova generazione (H100), eseguito in paesi green come la Svezia nelle ore centrali della giornata) 100 prompt consumano appena 25 kWh e generano appena 2,5 kg di CO_2. L'impatto equivale a guidare un'auto per 15 km o un consumo idrico paragonabile ad una semplice bottiglietta d'acqua per quanto riguarda il raffreddamento del data center.

In uno scenario intermedio "moderato", come un modello medio su prompt brevi, ma eseguito in Francia in orario serale, il consumo sale a 120 kWh e oltre 20 kg di CO_2. È come bruciare due bombole di gas da cucina o riempire un secchio d'acqua solo per raffreddare i server.

Se si aggiungono funzionalità avanzate, come il Chain-of-Thought su modelli di grandi dimensioni e hardware meno efficiente (A100), la curva si impenna: un batch di 100 prompt lunghi può richiedere 650 kWh, producendo 165 kg di CO_2. In pratica, equivale a un volo Roma–Milano andata e ritorno e a consumare l'acqua pari di una doccia intera solo per raffreddare i sistemi.

Nel worst case assoluto analizzato in questo studio, con un modello Dense da 300 miliardi di parametri, su A100, con CoT e multimodalità attivi, eseguito in India o Irlanda in orario serale, l'impatto diventa insostenibile: oltre 1.100 kWh e 420 kg di CO_2. È come percorrere 3.000 km in auto o fare 3 voli Roma–Parigi, mentre il raffreddamento richiede 140 litri d'acqua, l'equivalente di una vasca da bagno piena.[54]

Alla domanda di ricerca: *Quali sono i fattori che influenzano l'impatto ambientale nella fase di inferenza?"* possiamo rispondere dicendo che sono di 4 tipologie principali:

- Scelte architetturali nella fase di design del modello. Il numero dei parametri attivati effettivamente influenza il consumo energetico e di conseguenza la scelta tra modelli Dense e Mixture of Experts, nella media secondo questo studio ogni query che sfrutta MoE costa circa 7,7 secondi in più di uno Dense, a parità di condizioni.
- Lunghezza e complessità dell'output, nello studio è emerso come ogni token aggiuntivo in media comporta un aumento di latenza di 18ms. Inoltre, le Attività di Prompt Engineering realizzate dall'utente finale (come l'utilizzo di Chain of Thoughts), pur permettendo di ottenere risultati prestazionalmente superiori, comportano impatti ambientali non trascurabili,
- Hardware e scelte di GPU, modelli più recenti ed efficienti (come H100) garantiscono pari prestazioni ma con consumo ridotto rispetto a soluzioni meno ottimizzate (come A100),

[54] Nella sezione Appendice, è presente una tabella riepilogativa dell'impatto ambientale delle varie configurazioni analizzate.

- Scelte di localizzazione dei data center, sia in termini di intensità carbonica sia in termini di acqua utilizzata per le tecniche di raffreddamento dei data center. Anche la finestra temporale nella quale programmare le attività energivore può ridurre significativamente l'impatto carbonico, grazie alla penetrazione di energie rinnovabili e carbon-free.

Il messaggio chiave è che **le emissioni e i consumi non dipendono solo dalla dimensione del modello, ma dalla combinazione tra design architetturale e contesto di esecuzione.**
Con scelte consapevoli (fattori architetturali, fattori funzionali, pianificazione nelle ore a bassa carbon intensity e localizzazione in paesi con mix elettrico pulito), si può **ridurre l'impatto fino al 90%, trasformando un modello "energivoro" in una soluzione sostenibile.**

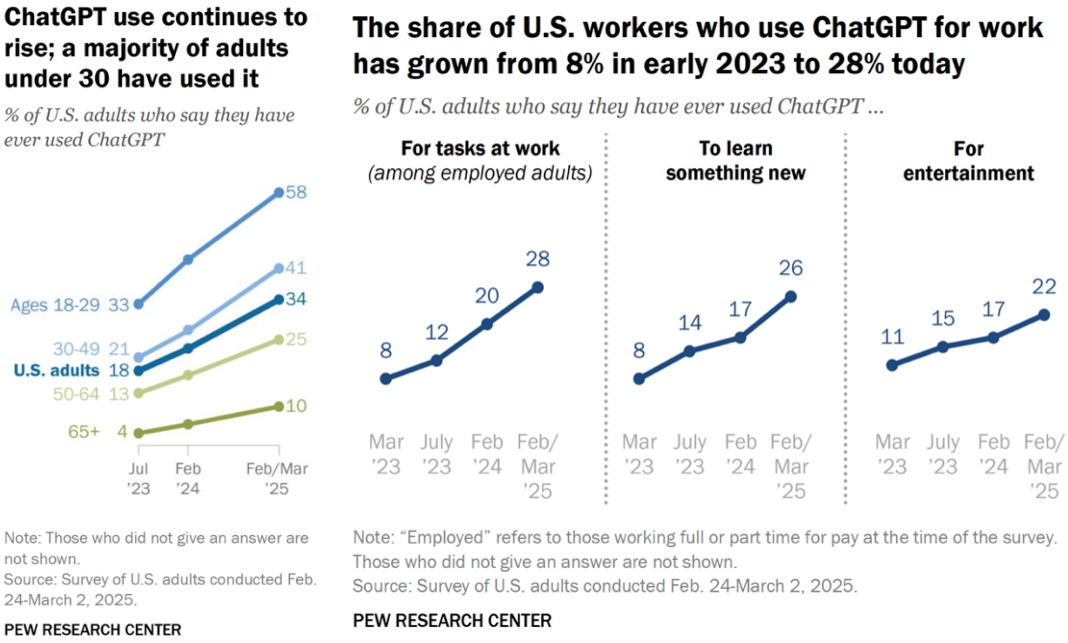
4.1.1 Diffusione e consumi

Se da un lato i risultati mostrano che l'impatto ambientale di una singola inferenza può essere drasticamente ridotto attraverso scelte consapevoli sul piano architetturale, funzionale e infrastrutturale, dall'altro non si può ignorare la dimensione della diffusione su larga scala dei modelli di Generative AI. È infatti questa la variabile che trasforma un problema tecnico in una questione sistemica.

Oggi non si tratta più di strumenti riservati a sviluppatori o ricercatori, ma componenti integrate nelle attività quotidiane di milioni di persone. La loro presenza nei motori di ricerca, nelle suite di produttività, nelle piattaforme di collaborazione e nei servizi consumer fa sì che l'interazione con un modello generativo diventi parte ordinaria della vita digitale. Si stima che centinaia di milioni di utenti accedano regolarmente a tali sistemi e che una quota crescente di imprese, piccole e grandi, stia incorporando l'AI generativa nei propri processi. In questo senso, la sostenibilità dell'inferenza non può più essere valutata solo a livello micro, sul singolo prompt, ma deve essere considerata alla luce dei volumi aggregati di utilizzo.
L'adozione delle generative AI è in forte crescita sia tra i consumatori sia nelle imprese. Google ha integrato i propri modelli direttamente nei servizi core: gli AI Overviews compaiono nei risultati della Ricerca Google, raggiungendo circa 2 miliardi di utenti mensili, mentre l'app Gemini, che sostituisce l'assistente Google e offre funzioni generative avanzate, conta oltre 450 milioni di utenti attivi mensili (Pichai, 2025, Alphabet Q2 2025 earnings call, Alphabet Inc., transcript).
OpenAI, con ChatGPT, mantiene una posizione di vertice tra i servizi digitali globali, con più di 400 milioni di utenti attivi ogni settimana (Kant, 2025[55]).

[55] Kant, 2025, OpenAI's weekly active users surpass 400 million, Reuters

Le statistiche ufficiali confermano l'ampiezza dell'adozione: nel Regno Unito il 31 % degli adulti dichiara di aver usato strumenti di IA generativa come ChatGPT o Gemini (Ofcom, 2025[56]); negli Stati Uniti la quota sale al 34 % (Sidoti & McClain, 2025[57]).



Dal lato delle imprese, nell'Unione Europea il 13,5 % delle aziende con almeno dieci addetti utilizza IA, con un picco oltre il 41 % tra le grandi (Eurostat, 2025[58]).

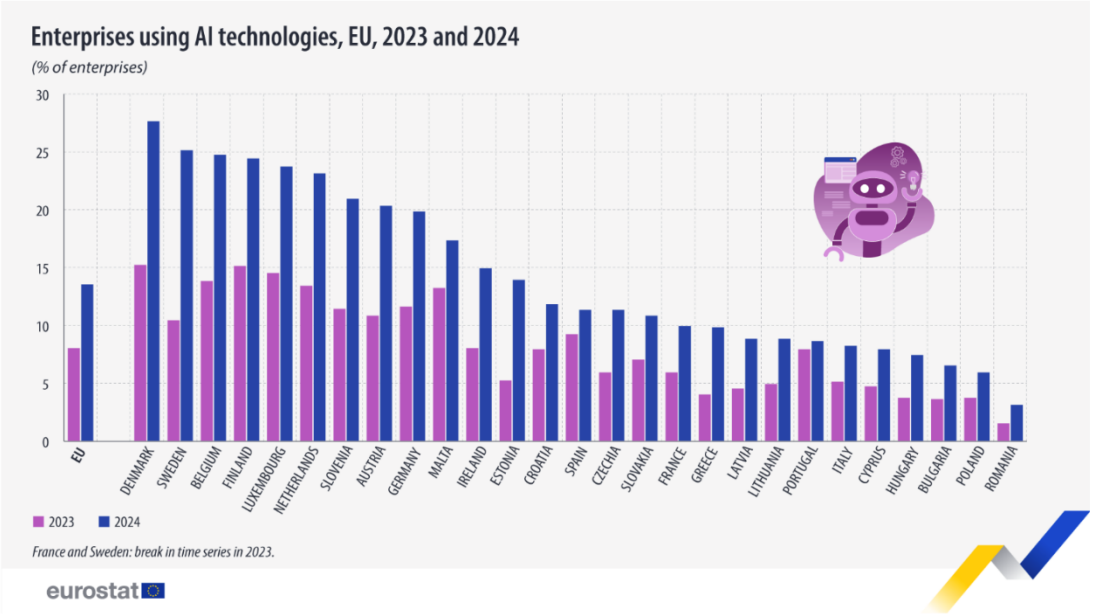
A livello globale, circa tre organizzazioni su quattro impiegano l'IA in almeno una funzione, segnalando il passaggio dai progetti pilota a implementazioni capaci di generare valore economico (McKinsey & Company, 2025[59]).

[56] Ofcom, 2025, Adults' media use and attitudes 2025, Ofcom

[57] Sidoti & McClain, 2025, 34% of U.S. adults have used ChatGPT, Pew Research Center

[58] Eurostat, 2025, Usage of AI technologies increasing in EU enterprises, Eurostat News; Eurostat, 2025, Use of artificial intelligence in enterprises – Statistics Explained, Eurostat

[59] McKinsey & Company, 2025, The state of AI: How organizations are rewiring to capture value, McKinsey & Company, 12 March; Stanford HAI, 2025, Artificial Intelligence Index Report 2025, Stanford University Human-Centered AI Institute



Secondo l'International Energy Agency (2025), la domanda elettrica dei data center globali potrebbe quasi raddoppiare entro il 2030, raggiungendo i 900–1.000 TWh annui, pari al 3–4% della domanda elettrica mondiale. L'IA generativa è identificata come uno dei principali driver di questa crescita, insieme allo streaming video e ai servizi cloud tradizionali (IEA, 2025).

4.1.2 Teorie Economiche sull'uso dell'energia: Paradossi di Jevons, Rebound-Effect e postulato di Khazzoom–Brookes

Il tema dell'efficienza e dei consumi aggregati trova un importante riferimento teorico nel cosiddetto paradosso di Jevons, formulato per la prima volta dall'economista britannico William Stanley Jevons nel 1865, nell'opera The Coal Question. Jevons osservava che i miglioramenti tecnologici nell'uso del carbone, ad esempio l'introduzione di macchine a vapore più efficienti, non portavano a una riduzione dei consumi complessivi di questa risorsa, ma al contrario a un loro incremento. Il motivo risiedeva nel fatto che l'aumento di efficienza riduceva i costi marginali di utilizzo, stimolando un'espansione della domanda e l'applicazione del carbone in un numero sempre maggiore di settori produttivi.
Applicato all'intelligenza artificiale generativa, ciò significa che rendere più efficienti i modelli abbassa il costo per singola inferenza, ma incentiva una diffusione d'uso sempre più ampia, che può compensare o superare i risparmi iniziali.
Questa dinamica si inserisce all'interno di un quadro più ampio di teorie economiche sull'uso dell'energia. Una delle estensioni più note è il concetto di rebound effect, elaborato nella letteratura energetica a partire dal XX secolo e sistematizzato negli studi di Khazzoom e Brookes (anni '80)[60].

[60] Il postulato di Khazzoom–Brookes (Saunders, 1992) interpreta il fenomeno su scala macroeconomica: l'aumento di efficienza energetica stimola la crescita economica e l'espansione delle attività produttive, incrementando i consumi aggregati a livello di

Il rebound effect descrive il fenomeno per cui i miglioramenti di efficienza energetica riducono il costo per unità di servizio erogato, stimolando un incremento della domanda che compensa in parte o del tutto i risparmi energetici iniziali *(Greening, Lorna A.; Greene, David L.; Difiglio, Carmen, Energy efficiency and consumption — the rebound effect — a survey, 2000, Energy Policy, 28(6-7), p. 389–401).* In relazione ai modelli di generative AI, ciò si traduce in un uso esteso dello strumento per attività a basso valore aggiunto, come la redazione in ambito aziendale di memo interni o l'elaborazione di e-mail di routine, che prima non sarebbero state delegate a sistemi di AI o l'integrazione di modelli leggeri in motori di ricerca, applicazioni di messaggistica e suite di produttività: azioni quotidiane che generano un volume di richieste senza precedenti e che fino a pochi anni fa non implicavano consumi energetici aggiuntivi.

Il paradosso di Jevons e il rebound effect applicati all'AI generativa mostrano chiaramente che la sostenibilità non può essere garantita dal solo progresso tecnologico. Anche se modelli e hardware diventano più efficienti, l'aumento della scala di utilizzo rischia di superare i benefici ottenuti. L'impatto ambientale non dipende quindi solo dalla loro efficienza tecnica, ma dal volume globale di query e dalla diffusione su scala consumer ed enterprise.

La crescita di queste applicazioni – dai motori di ricerca agli assistenti virtuali – moltiplica gli impatti ambientali per ordini di grandezza: anche un consumo unitario ridotto diventa significativo quando milioni di aziende e centinaia di milioni di utenti interagiscono quotidianamente con tali sistemi.

Le implicazioni sono rilevanti e richiedono un approccio integrato che unisca dimensione tecnologica, regolatoria e comportamentale.

In primis, la governance dei provider e le politiche pubbliche devono orientare la domanda verso un uso responsabile e consapevole. Sul piano industriale, diverse iniziative internazionali indicano la necessità di metriche standardizzate per misurare consumo energetico, emissioni e uso idrico lungo tutto il ciclo di vita del modello, dall'addestramento all'inferenza. Le linee guida dell'OECD raccomandano un monitoraggio comparabile e obblighi di disclosure sui consumi (OECD, 2024), mentre il Programma Ambiente delle Nazioni Unite (UNEP) sottolinea l'urgenza di reporting trasparente e di standard globali per valutare l'impatto ambientale dell'AI (UNEP, 2024).

Sul piano normativo, alcuni ordinamenti iniziano a muoversi. Negli Stati Uniti l'Artificial Intelligence Environmental Impacts Act prevede la creazione di un consorzio pubblico-privato e l'adozione di sistemi di rendicontazione volontaria per l'impatto ambientale dell'AI (United States Congress, 2024). Nell'Unione Europea, l'AI Act, pur concentrandosi sulla sicurezza, introduce obblighi di trasparenza e documentazione che possono essere estesi agli impatti energetici e climatici. Questi primi passi indicano una direzione verso regole vincolanti di disclosure e audit indipendenti, ma mostrano anche lacune significative per la fase di inferenza, oggi responsabile di gran parte della crescita nei consumi.

---

sistema. Traslato sull'AI generativa, questo postulato suggerisce che modelli sempre più efficienti non si limitano a sostituire processi esistenti, ma abilitano nuovi mercati e modelli di business, generando una domanda aggiuntiva che accresce l'impatto complessivo.

Accanto alla regolazione, le politiche pubbliche possono orientare la domanda attraverso strumenti economici:

- incentivi fiscali o crediti per data center alimentati da fonti rinnovabili;
- tassazione sul carbonio o price signals che internalizzino il costo delle emissioni;
- standard minimi di efficienza energetica per infrastrutture e modelli.

A queste misure si affiancano le strategie di governance interna dei provider e strategie di gestione della domanda, che agiscano sul lato dell'utilizzo. Esempi significativi sono:

- la pianificazione dei carichi di lavoro in funzione della carbon intensity della rete elettrica, privilegiando fasce orarie o aree geografiche con mix energetico più pulito;
- la definizione di budget di token o di crediti di calcolo per applicazioni non critiche, così da evitare consumi incontrollati;
- l'uso selettivo delle funzionalità avanzate, come ragionamenti multi-step o capacità multimodali, che, pur offrendo valore aggiunto, comportano un costo energetico sensibilmente superiore.

In quest'ottica, la sostenibilità non può essere garantita dal solo progresso tecnologico. Efficienza tecnica e consapevolezza d'uso devono procedere insieme: l'adozione massiva di modelli sempre più potenti rischia altrimenti di annullare i progressi compiuti, trasformando un potenziale beneficio ambientale in una nuova fonte di pressione sulle risorse. Una governance efficace, sostenuta da politiche pubbliche lungimiranti, diventa quindi la condizione per trasformare l'innovazione in vantaggio ambientale concreto e duraturo.

4.2 Limiti dell’analisi

Questa ricerca, pur offrendo un contributo originale allo studio dell’impatto ambientale dei modelli di Generative AI, presenta alcune limitazioni che è necessario considerare. In primo luogo, la raccolta dei dati è avvenuta tramite API pubbliche. Questa modalità ha consentito di ottenere osservazioni dirette e comparabili tra modelli, ma ha al contempo limitato l’accesso a informazioni hardware di dettaglio (ad esempio, consumo elettrico istantaneo delle GPU, parallelismo interno, tecniche di batching). Di conseguenza, le determinanti dei consumi sono colte in modo indiretto, senza la possibilità di isolarne sempre l’effetto puntuale.

La latenza e i consumi dipendono anche da fattori esterni (es. traffico di rete, condizioni climatiche) che non è stato possibile isolare, ma è stato possibile gestire mediante un dataset molto ampio. I dati sono stati raccolti in giornate differenti, con strumenti differenti, ma utilizzando sempre la stessa connessione. In fase di pulizia dei dati sono stati eliminati i run di warm-up per ogni tipologia di prompt per ogni modello e sono stati eliminati gli outliers, in quanto non rappresentativi.

Questo lavoro ha posto il proprio focus sulla fase di inferenza, escludendo il training e i cicli di vita completi dell’hardware. Tale scelta comporta il limite di non considerare lo Scope 3, ossia le emissioni indirette lungo la catena del valore (produzione e smaltimento delle GPU, costruzione dei data center, logistica dei componenti, dispositivi finali). Un approccio di Life Cycle Assessment (LCA) sarebbe necessario per restituire una fotografia completa dell’impatto ambientale dei modelli di Generative AI.

Per quanto riguarda la copertura temporale e geografica, seppur ampia (12 Paesi, dati orari per il 2024), rimane circoscritta. Non è stato possibile includere tutte le regioni globali né verificare con dati primari le effettive strategie di bilanciamento energetico dei provider. Di conseguenza, la generalizzazione dei risultati richiede cautela, soprattutto in scenari extra-europei o in contesti energetici in rapido mutamento.

Un ulteriore limite riguarda la dimensione idrica: il WUE è stato considerato come indicatore standard, ma i dati disponibili restano aggregati e, in molti casi, non verificabili in modo indipendente. La quantificazione precisa degli impatti idrici richiederebbe studi caso per caso sui singoli data center, oggi ancora scarsamente accessibili.

Il settore dei modelli generativi è in rapida evoluzione. L’hardware, gli algoritmi e le pratiche di ottimizzazione cambiano a ritmi accelerati, per cui alcune delle conclusioni potrebbero diventare rapidamente obsolete. I risultati vanno quindi interpretati come una fotografia del contesto attuale più che come un punto di arrivo definitivo.

4.3 Spunti per Ricerche Future

Alla luce dei limiti evidenziati, questo lavoro apre a diverse possibili direttrici di ricerca futura.
Un primo sviluppo potrebbe consistere in un'analisi più granulare della latenza, distinguendo in modo esplicito il time-to-first-token dal tempo di decoding, così da cogliere con maggiore precisione il contributo delle due componenti all'impatto energetico, idrico e in termini di emissioni di CO_2.
Un altro fronte di ricerca riguarda l'inclusione di variabili operative oggi non osservabili, come il batching dinamico, la caching delle richieste o le strategie di ottimizzazione e di raffreddamento adottate dai provider. Questi fattori incidono in modo significativo sia sulla latenza sia sull'efficienza energetica, ma non sono rilevabili tramite interazioni con le API pubbliche. Per ridurre questa lacuna, studi futuri potrebbero svilupparsi attraverso collaborazioni dirette con i fornitori di servizi cloud, così da accedere a dati di telemetria in tempo reale e a informazioni più dettagliate sulla configurazione hardware e sul mix energetico effettivamente utilizzato.
Infine, ulteriori estensioni potrebbero riguardare l'integrazione di una valutazione del ciclo di vita dell'hardware (Life Cycle Assessment), capace di stimare l'impatto ambientale della produzione e della dismissione delle GPU. Futuri sviluppi di ricerca potrebbero integrare questa prospettiva, includendo lo Scope 3 e adottando metodologie di LCA per restituire una fotografia più completa e multilivello dell'impatto ambientale dei Large Language Models.
Lo sviluppo di scenari prospettici basati su configurazioni di calcolo più avanzate, ad esempio modelli con un trilione di parametri, output fino a 10.000 token e cluster di GPU di nuova generazione, potrà infine fornire indicazioni preziose sui possibili trend futuri e sui rischi di scala connessi all'evoluzione tecnologica.

Queste linee di ricerca, nel loro insieme, permetterebbero di ampliare e rendere più accurata la comprensione dell'impatto ambientale dell'inferenza nei modelli di Generative AI, rafforzando la solidità e l'attualità delle analisi proposte in questa tesi.

Capitolo 5 - Conclusioni

L'analisi condotta in questo lavoro si è posta come obiettivo quello di rispondere alla domanda di ricerca – "Quali fattori influenzano l'impatto ambientale, in termini di consumo energetico, consumo idrico ed emissioni di CO_2, dei modelli di Generative AI nella fase di inferenza?". I risultati ottenuti mostrano che l'impatto non può essere spiegato da una singola variabile, ma è il prodotto dell'interazione di fattori architetturali, operativi e infrastrutturali.
Muovendo da un gap della letteratura, che ha storicamente privilegiato lo studio della fase di training e analisi aggregate a scapito dell'inferenza e delle variabili micro, la ricerca ha costruito un dataset originale basato su query reali a 16 modelli eterogenei, testati tramite API pubbliche. Tale approccio ha consentito di stimare empiricamente la latenza come proxy del consumo energetico, di collegarla a fattori architetturali, funzionali e infrastrutturali, e di estendere l'analisi a emissioni e impatto idrico.

I risultati ottenuti mostrano, in primo luogo, che la latenza, e dunque l'energia consumata, non dipende soltanto dalla dimensione del modello, ma è sensibile a variabili architetturali (Dense vs MoE), funzionali (multimodalità e reasoning, es. chain-of-thought,) e operative (lunghezza di input e output). I dati empirici raccolti evidenziano che la latenza cresce in maniera statisticamente significativa al crescere dei token in output, con effetti diretti sul fabbisogno di calcolo e quindi sull'energia consumata. Innovazioni funzionali come la Chain-of-Thought e la multimodalità, pur rappresentando un avanzamento in termini di capacità cognitiva e versatilità applicativa, comportano incrementi sostanziali di latenza e consumi. L'analisi empirica ha mostrato che queste scelte progettuali possono moltiplicare l'energia richiesta fino a raddoppiare l'impatto per 100 richieste, trasformando un consumo paragonabile a una lampadina accesa per alcune ore in uno equivalente a un forno elettrico acceso per un'ora.

Lo studio ha permesso di analizzare inoltre le emissioni di CO2 dei modelli di Generative AI. L'analisi basata su dati orari di carbon intensity ha dimostrato che il luogo e l'orario di esecuzione delle query possono modificare sensibilmente l'impatto ambientale. A parità di modello, inferenze eseguite in fasce orarie ad alta penetrazione di rinnovabili o in regioni con mix energetici meno carbon-intensive comportano riduzioni delle emissioni fino al 90%, portando le emissioni di CO2 da quelle di 3 voli Roma-Parigi a quelle di un'auto in funzione per 15km. Questo risultato sottolinea come la scelta del data center e delle finestre temporali di utilizzo non sia neutrale, ma costituisca una leva strategica di sostenibilità.

Infine, il lavoro ha esteso l'analisi alla dimensione idrica, spesso trascurata dalla letteratura. L'inclusione del Water Usage Effectiveness (WUE) ha mostrato che il raffreddamento dei data center rappresenta un ulteriore vincolo materiale, con impatti che possono diventare critici in contesti di scarsità idrica. Come abbiamo visto, i provider offrono data center localizzati in paesi differenti e in fase di scelta è importante tenere a mente le tecniche di raffreddamento utilizzate dai data center e gli impatti a livello idrico che questi possono

avere. Integrare questa prospettiva permette di superare la visione ristretta alla sola carbon footprint e di restituire un quadro più completo delle esternalità ambientali associate ai modelli di GenAI.

Complessivamente, il lavoro dimostra che l'impatto ambientale dei modelli di generative AI in inferenza è il risultato di una combinazione di scelte tecniche e infrastrutturali: le caratteristiche dei modelli, l'hardware impiegato, la localizzazione e l'orario di utilizzo interagiscono nel determinare consumi, emissioni e impatto idrico. Rispetto al gap iniziale, questo studio offre un contributo duplice: da un lato, introduce un'analisi sistematica e comparabile basata su dati reali; dall'altro, propone un approccio multidimensionale che integra energia, CO_2 e acqua. In tal modo, fornisce strumenti concreti per orientare la progettazione e la governance della Generative AI, mostrando che le scelte consapevoli possono ridurre l'impatto ambientale fino al 90% senza sacrificare le capacità dei modelli.
Solo un approccio integrato, che combini efficienza tecnica e consapevolezza d'uso, può evitare che l'adozione massiva dei modelli annulli i progressi compiuti sul fronte della sostenibilità.

BIBLIOGRAFIA

174 Power Global. 2025.How AI Changes Data Center Design.

Alder, N., Ebert, K., Herbrich, R. and Hacker, P. (2024) *AI, Climate, and Regulation: From Data Centers to the AI Act*. arXiv preprint arXiv:2410.06681.

Amazon Web Services, Optimizing AI Responsiveness: A Practical Guide to Amazon Bedrock Latency-Optimized Inference, AWS Machine Learning Blog, 2024, disponibile su: https://aws.amazon.com/blogs/machine-learning/optimizing-ai-responsiveness-a-practical-guide-to-amazon-bedrock-latency-optimized-inference

Amazon Web Services (2023) Sustainability in the Cloud: AWS Data Center Efficiency. Seattle, WA: AWS.

AMD (2023) AMD Instinct MI300 Accelerator Technical Overview. Sunnyvale, CA: AMD.

Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav Gulavani, Alexey Tumanov, and Ramachandran Ramjee. 2024.Taming Throughput-Latency tradeoff in LLM inference with Sarathi-Serve. In 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24). 117–134.

Anthropic, Series F update, 2025;

Anthropic (2023) Anthropic and AWS Partnership Announcement. San Francisco, CA: Anthropic.

Anthropic (2024). Claude 3: Next-generation language models from Anthropic. Disponibile su: https://www.anthropic.com/news/claude-3-family [Accesso: aprile 2025].

Artificial Analysis, Performance Benchmarking Methodology, 2025, disponibile su: https://artificialanalysis.ai/methodology/performance-benchmarking

Bansal, P., Demystifying LLM Benchmarks: Tokens, Quality, Latency, Throughput, Medium, 2023, disponibile su: https://medium.com/@prateekbansalind/demystifying-llm-benchmarks-tokens-quality-latency-throughput-b61965dd93e0

Barroso, L., Hölzle, U., Clidaras, P., The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, 2013, Morgan & Claypool, p. 33.

ChatBase (2025). Is ChatGPT accurate? Disponibile su: https://www.chatbase.co/blog/is-chatgpt-accurate

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C. e Tafjord, O. (2018) Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv preprint arXiv:1803.05457. Disponibile su: https://arxiv.org/abs/1803.05457

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, Ł. e Miller, J. (2021) Training Verifiers to Solve Math Word Problems. arXiv preprint arXiv:2110.14168. Disponibile su: https://arxiv.org/abs/2110.14168

CodeCarbon (2022) CodeCarbon: A Tool for Tracking the Carbon Emissions of Machine Learning Workloads. Montreal: Mila – Quebec AI Institute.

Databricks, LLM Inference Performance Engineering: Best Practices, 2024, disponibile su: https://www.databricks.com/blog/llm-inference-performance-engineering-best-practices

DataCamp (2024–2025). Claude-4 Overview. Disponibile su: https://www.datacamp.com/blog/claude-4

DeepSeek AI (2024). DeepSeek V3: Open-source LLMs for multilingual and multimodal tasks. Disponibile su: https://deepseek.com

Dell Technologies, Optimizing Throughput in Large Language Models and Understanding Token Dynamics, InfoHub, 2024, disponibile su: https://infohub.delltechnologies.com/it-it/p/optimizing-throughput-in-large-language-models-and-understanding-token-dynamics/

Deng, Y., Xu, M., Yang, Z., Chen, T. and Xu, C. (2025) How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference. arXiv preprint arXiv:2501.01234.

Digital Realty (2024) *EU sustainability regulations: CSRD and beyond*.

European Parliament (2021) *Non-Financial Reporting Directive: Implementation and challenges*.

European Commission (2024) *EU Artificial Intelligence Act: Implementation timeline*.

European Union (2024) *Regulation (EU) 2024/1689 laying down harmonised rules on Artificial Intelligence (AI Act)*. Official Journal of the European Union, 12 July.

Eurostat, Usage of AI technologies increasing in EU enterprises, 2025;

Eurostat, Use of artificial intelligence in enterprises, 2025;

Financial Times (2025). AI start-up Anthropic valued at $170bn

Greening, L.A., Greene, D.L. & Difiglio, C. (2000) Energy efficiency and consumption — the rebound effect — a survey. *Energy Policy*, 28(6–7), pp.389–401.

GRESB (2024) *Sustainable data centers: ESG compliance and futureproofing for success*.

Glazer, E., Choi, J., Steinhardt, J., Wei, J. e Zhou, D. (2024) FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI. arXiv preprint arXiv:2406.14615.

Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep Learning*. Cambridge, MA: MIT Press.

Google (2023). How Google is integrating generative AI into Search. Google Blog. Disponibile su: https://blog.google/products/search/generative-ai-search-update/

Google (2024). Google Environmental Report 2024. Disponibile su: https://sustainability.google/reports/google-2024-environmental-report/

Google LLC (2023) Environmental Report 2023: Efficiency and Carbon Neutrality in Google Data Centers. Mountain View, CA: Google.

Greening, L. A., Greene, D. L. & Difiglio, C. (2000). Energy efficiency and consumption — the rebound effect — a survey. *Energy Policy*, 28(6–7), pp. 389–401.

Guo, G., et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv preprint arXiv:2501.12948.

Gupta, V., Pantoja, D., Ross, C., Williams, A. e Ung, M. (2024) Changing Answer Order Can Decrease MMLU Accuracy. arXiv preprint arXiv:2406.19470.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D. and Pineau, J. (2020) Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. Journal of Machine Learning Research, 21(248), pp. 1–43.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. e Steinhardt, J. (2020) Measuring Massive Multitask Language Understanding. arXiv preprint arXiv:2009.03300.

Hugging Face (2024–2025). Open LLM Leaderboard discussions. Disponibile su: https://huggingface.co/spaces/open-llm-leaderboard

Hugging Face (2025). Announcing AI Energy Score. huggingface.co.

Hugging Face Model Card Llama-3.2-90B Vision Instruct → huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct

International Energy Agency (IEA), CO_2 Emissions from Fuel Combustion Highlights, 2022, OECD/IEA, n.p.

International Energy Agency (IEA) (2024) Global Electricity Review: Carbon Intensity by Region. Paris: IEA.

Jegham, N., Abdelatti, M., Elmoubarki, L., Hendawi, A., How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference, 2025, arXiv

Jevons, W. S. (1865). *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-mines*. London: Macmillan.

Jiang, A.Q. et al. (2023). Mistral 7B. arXiv:2310.06825.

Jiang, A.Q. et al. (2024). Mixtral of Experts. arXiv:2401.04088.

Kant, Rishi, OpenAI's weekly active users surpass 400 million, 2025, Reuters;

Klu.ai (2024–2025). Mixtral-8×7B Benchmark Note. Disponibile su: https://klu.ai/glossary/mistral-8x7b

Kreyszig, E., Advanced Engineering Mathematics, 2011, Wiley, p. 902.

Lawrence Berkeley National Laboratory (2022) Water Usage Effectiveness (WUE) and Power Usage Effectiveness (PUE) in Data Centers. Berkeley, CA: U.S. Department of Energy.

LeCun, Y., Bengio, Y. & Hinton, G. (2015) 'Deep learning', *Nature*, 521(7553), pp. 436–444.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M. et al. (2022) Holistic Evaluation of Language Models. arXiv preprint arXiv:2211.09110.

Lin, S., Hilton, J. e Evans, O. (2021) TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv preprint arXiv:2109.07958. Disponibile su: https://arxiv.org/abs/2109.07958

Luccioni, S., Viguier, A., Ligozat, A., Estimating the Carbon Footprint of Machine Learning, 2023, arXiv, p. 3.

Mas-Colell, Andreu, Whinston, Michael, Green, Jerry, Microeconomic Theory, 1995

McKinsey & Company, The State of AI, 2025; Stanford HAI, AI Index 2025, 2025).

Menlo Ventures (2025). 2025 Mid-Year LLM Market Update. Disponibile su: https://menlovc.com/perspective/2025-mid-year-llm-market-update/

Meta (2025). LLaMA-3.2-90B Vision – Model Card. Hugging Face. Disponibile su: https://huggingface.co/meta-llama/Llama-3.2-90B-Vision

Meta AI (2023) LLaMA 2 Technical Report. Menlo Park, CA: Meta.

Meta AI Blog, Introducing Llama 3 and Llama 3.2, 2024 → ai.meta.com/llama

Microsoft (2024). Environmental Sustainability Report 2024. Disponibile su: https://www.microsoft.com/en-us/corporate-responsibility/sustainability/report

Microsoft Corporation (2023) Microsoft Azure Sustainability Report. Redmond, WA: Microsoft.

Mistral AI (2023) Mistral 7B and Mixtral Model Release Notes. Paris: Mistral AI.

Mistral AI (2023). Announcing Mistral-7B. Disponibile su: https://mistral.ai/news/announcing-mistral-7b

MLCommons (2023) MLPerf Inference Benchmark Results. Available at: https://mlcommons.org (Accessed: 25 August 2025).

MLCommons (2024). MLPerf Inference v4.0 Results. mlcommons.org.

Montgomery, D.C., Design and Analysis of Experiments, 2017, Wiley, p. 69.

Mukherjee, Supantha, French startup Mistral launches chatbot for companies, 2025, Reuters;

National Institute of Standards and Technology (2021) *Artificial Intelligence Risk Management Framework (AI RMF)*. Gaithersburg, MD: NIST.

Nadella, Satya, Microsoft FY25 Q4 Earnings Conference Call, 2025;

NVIDIA Corporation (2020) NVIDIA A100 Tensor Core GPU Architecture Whitepaper. Santa Clara, CA: NVIDIA.

NVIDIA Corporation (2022) NVIDIA Management Library (NVML) Documentation. Santa Clara, CA: NVIDIA.

NVIDIA Corporation (2023a) NVIDIA H100 Tensor Core GPU Whitepaper. Santa Clara, CA: NVIDIA.

NVIDIA Corporation (2023b) NVIDIA L4 Tensor Core GPU Product Datasheet. Santa Clara, CA: NVIDIA.

OECD (2024) *Measuring the environmental impacts of artificial intelligence: compute and applications*. OECD Publishing.

Ofcom, Adults' Media Use and Attitudes 2025, 2025;

OpenAI (2023) OpenAI and Microsoft Azure Partnership Documentation. San Francisco, CA: OpenAI.

OpenAI (2024). GPT-4o: OpenAI's multimodal flagship model. Disponibile su: https://openai.com/index/gpt-4o

OpenAI (2024). Introducing GPT-4.1 in the API. Disponibile su: https://openai.com/index/gpt-4-1

OpenAI (2025). GPT-o3 and o3-mini: Multimodal instruction-tuned models. Disponibile su: https://openai.com/index/openai-o3-mini/

OpenAI (2025a). Introducing GPT-5 for Developers. Disponibile su: https://openai.com/index/gpt-5

OpenAI (2025b). GPT-5 System Card. Disponibile su: https://openai.com/research/gpt-5

OpenMetal, Measuring AI Model Performance: Tokens per Second, Model Sizes, and Inferencing Tools, Blog, 2025.

Pareto, Manuale di economia politica, 1906) (Pareto, Vilfredo, Manuale di economia politica, 1906).

Pew Research Center, 34% of U.S. adults have used ChatGPT, 2025;

Piastou, M. (2025) A Review of Benchmarks for Evaluating the Mathematical Abilities of Modern LLM in Advanced Problem Solving. Eurasian Union Scientists. Disponibile su: https://www.researchgate.net/publication/388556839_A_Review_of_Benchmarks_for_Evaluating_the_Mathematical_Abilities_of_Modern_LLM_in_Advanced_Problem_Solving

Pichai, Sundar, Alphabet Q2 2025 earnings call, 2025;

Polimeni, J. M., Mayumi, K., Giampietro, M. & Alcott, B. (2008). *The Jevons Paradox and the Myth of Resource Efficiency Improvements*. London: Earthscan.

Polo, F.M., Weber, L., Choshen, L., Sun, Y., Xu, G. e Yurochkin, M. (2024) tinyBenchmarks: Evaluating LLMs with Fewer Examples. arXiv preprint arXiv:2402.14992. Disponibile su: https://arxiv.org/abs/2402.14992

Reuters (2025). OpenAI's weekly active users surpass 400 million.

Russell, S. & Norvig, P. (2021) *Artificial Intelligence: A Modern Approach*. 4th edn. Harlow: Pearson.

Saunders, H. D. (1992). The Khazzoom–Brookes postulate and neoclassical growth. *The Energy Journal*, 13(4), pp. 131–148.

Schwartz, Roy, Dodge, Jesse, Smith, Noah, Etzioni, Oren, Green AI, 2020;

Schwartz, R., Dodge, J., Smith, N., Etzioni, O. (2020). Green AI. Communications of the ACM.

Sensor Tower, State of AI Apps 2025, 2025;

Silberling, Amanda, Meta says Llama models have been downloaded 1.2B times, 2025;

Sorrell, S. (2009). Jevons' Paradox revisited: The evidence for backfire from improved energy efficiency. *Energy Policy*, 37(4), pp. 1456–1469.

Strubell, E., Ganesh, A. and McCallum, A. (2019) Energy and Policy Considerations for Deep Learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3645–3650.

Strubell, E., Ganesh, A., McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. arXiv:1906.02243.

Suzgun, M., Scao, T.L., Shuster, K., Tay, Y., Wu, Y., Zoph, B. e Wei, J. (2022) Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. arXiv preprint arXiv:2210.09261. Disponibile su: https://arxiv.org/abs/2210.09261

Symbl.ai, A Guide to LLM Inference Performance Monitoring, 2024, disponibile su: https://symbl.ai/developers/blog/a-guide-to-llm-inference-performance-monitoring

Taylor, J.R., An Introduction to Error Analysis, 1997, University Science Books, p. 27.

TechCrunch (2025). Google's AI Overviews have 2B monthly users; Gemini 450M MAU.

TensorWave (2025) LLM Benchmarks: How to Read AI's Report Card. TensorWave Blog [online]. Disponibile su: https://tensorwave.com/blog/llm-benchmarks

The Green Grid, PUE: A Comprehensive Examination of the Metric, 2012, White Paper 49

The Green Grid, Water Usage Effectiveness (WUE): A Green Grid Data Center Sustainability Metric, 2011, White Paper,

Together.ai Models → api.together.ai/models

Touvron, H., Martin, L., Stone, K. et al. (2023) LLaMA 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288.

UNEP (2024) *AI has an environmental problem: Here's what the world can do about it*. United Nations Environment Programme.

United States Congress (2024) *Artificial Intelligence Environmental Impacts Act of 2024*. 118th Congress.

Uptime Institute (2023) Global Data Center Survey 2023. Uptime Institute Intelligence Report.

Utterback, J.M. and Abernathy, W.J. (1975) A Dynamic Model of Process and Product Innovation. Omega, 3(6), pp. 639–656.

Vals.ai (2025). MMLU-Pro Leaderboard. Disponibile su: https://www.vals.ai/benchmarks/mmlu_pro

Vendrow, J., Vendrow, E., Beery, S. e Mądry, A. (2025) Do Large Language Model Benchmarks Test Reliability? NeurIPS 2024 Workshop on SafeGenAi. Disponibile su: https://arxiv.org/html/2502.03461v1

Wu, J. et al. (2024). The Carbon Footprint of Large Language Models. Nature

Zellers, R., Bisk, Y., Farhadi, A. e Choi, Y. (2019) HellaSwag: Can a Machine Really Finish Your Sentence?. arXiv preprint arXiv:1905.07830.

Zhou, S. et al. (2024) *AI, Climate, and Regulation: From Data Centers to the AI Act*. *arXiv preprint* arXiv:2410.06681.

Zuckerberg, Mark, Celebrating 1 Billion Downloads of Llama, 2025;

APPENDICE

ALLEGATO 1: Tabella Riepilogativa – Impatto Ambientale Configurazioni

Prompt type	Configurazione	Energia (kWh)	CO_2 min (Svezia)	CO_2 max (India)	Acqua min (Irlanda)	Acqua max (Virginia)	Equivalente quotidiano
Breve	Base	0,34	0,5 g	130 g	0,007 L	0,048 L	lampadina LED 6h
	+ Multimodalità	~0,42	0,7 g	161 g	0,008 L	0,059 L	lampadina LED 8h
	+ CoT	~0,58	0,9 g	222 g	0,012 L	0,081 L	asciugacapelli 10 min
	+ CoT + Multi	~0,70	1,1 g	268 g	0,014 L	0,098 L	phon 15 min
Medio	Base	2,25	3,6 g	863 g	0,045 L	0,315 L	lavatrice 40°
	+ Multimodalità	~2,8	4,5 g	1.077 g	0,056 L	0,392 L	aria condizionata 2h
	+ CoT	~3,8	6,1 g	1.455 g	0,076 L	0,532 L	forno 45 min
	+ CoT + Multi	~4,5	7,2 g	1.720 g	0,090 L	0,630 L	forno 1h
Lungo	Base	3,46	5,5 g	1.326 g	0,069 L	0,484 L	TV 20–30 min
	+ Multimodalità	~4,3	6,9 g	1.648 g	0,086 L	0,602 L	microonde 30 min
	+ CoT	~5,9	9,4 g	2.260 g	0,118 L	0,826 L	auto elettrica 10 km
	+ CoT + Multi	~7,1	11,4 g	2.720 g	0,142 L	0,994 L	auto elettrica 15 km

ALLEGATO 2:**Costruzione degli indici di sostenibilità: Impatto ambientale (energia, CO_2, acqua) dei modelli LLM frutto dell'analisi empirica**

Il primo passo dell'analisi riguarda la costruzione di una classifica dei modelli di linguaggio di grandi dimensioni (LLM) in base al loro impatto ambientale, valutato attraverso tre dimensioni: consumo energetico, emissioni di anidride carbonica equivalente e consumo idrico.

Si è scelto di esprimere tutti i valori per 100 token generati, così da ottenere un indicatore di intensità comparabile indipendentemente dalla lunghezza delle risposte prodotte.

Le tre metriche ambientali sono state calcolate come segue:

- **Energia (Wh/100 tok)**

$$Energia(Wh/100tok) = \frac{Energia_{(tot\_kWh\times)}1000}{Output_{tokens}} \times 100$$

- **Emissioni di CO_2 (gCO_2e/100 tok)**

$$\mathrm{CO2(gCO2e/100tok)} = \frac{\mathrm{CO2_{kg}} \times 1000}{\mathrm{Output_{tokens}}} \times 100$$

- **Acqua (mL/100 tok)**

$$\mathrm{Acqua(mL/100tok)} = \frac{\mathrm{Acqua_{L}} \times 1000}{\mathrm{Output_{tokens}}} \times 100$$

Tali indicatori rappresentano l'intensità ambientale associata a un'unità standard di output linguistico, rendendo i modelli direttamente confrontabili.

1. Normalizzazione e aggregazione dei dati

Una volta definite le formule di calcolo delle intensità ambientali per 100 token, si è proceduto all'applicazione ai dati sperimentali raccolti, in modo tale da passare per ciascun modello da valori assoluti (Energia totale in kWh, CO_2 in kg, Acqua in litri) a valori normalizzati e comparabili (Wh/100 tok, gCO_2e/100 tok, mL/100 tok).

Questa operazione consente di svincolare i risultati dalla lunghezza delle risposte generate dai modelli, che nei file variava da poche decine fino a oltre un migliaio di token. Senza tale standardizzazione, i modelli che tendono a produrre output più lunghi sarebbero penalizzati da consumi assoluti maggiori.

Successivamente, per ciascun modello sono stati aggregati i valori normalizzati provenienti da tutti i run disponibili. L'aggregazione è stata condotta calcolando la **media aritmetica** dei valori di ciascuna metrica (Wh/100 tok, gCO_2e/100 tok, mL/100 tok). Questa scelta metodologica garantisce la disponibilità di un indicatore sintetico e rappresentativo.

2. Costruzione dei ranking

Una volta completata la normalizzazione e l'aggregazione per modello, procediamo alla costruzione di tre classifiche distinte, ordinate dal valore più alto (impatto maggiore) al più basso (impatto minore):

- ranking energetico (Wh/100 tok),
- ranking di emissioni (gCO_2e/100 tok),
- ranking idrico (mL/100 tok).

Nella costruzione delle classifiche si è scelto di distinguere due scenari: uno **europeo** e uno **globale**. La ragione di questa distinzione risiede nella necessità di contestualizzare l'analisi rispetto ai probabili luoghi di esecuzione delle richieste.

Lo scenario europeo è stato considerato prioritario poiché, con elevata probabilità, i prompt inviati vengono processati in data center collocati in Europa, in particolare in hub strategici come **Paesi Bassi e Irlanda**. Queste aree rappresentano i principali nodi infrastrutturali delle grandi piattaforme cloud (Microsoft Azure, AWS, Google Cloud) e costituiscono quindi il contesto più realistico per valutare l'impatto ambientale delle interrogazioni effettuate.

Parallelamente, è stato introdotto anche uno scenario globale con l'obiettivo di conferire **maggior respiro internazionale all'analisi** e di permettere confronti più ampi. In questo caso, oltre ai data center europei, sono stati inclusi i principali poli statunitensi, come la **Virginia (USA),** che rappresenta uno dei cluster più rilevanti a livello mondiale per l'hosting di servizi cloud.

- Ranking Energetico

Il ranking energetico riporta i modelli ordinati dal meno al più energivoro, mostrando in modo comparativo l'intensità media di consumo per 100 token. Questa rappresentazione consente di individuare con chiarezza i modelli più efficienti e quelli con il maggior impatto energetico nei due scenari analizzati.

Ranking energetico – Europa

La Tabella sottostante mostra il ranking dei modelli nel contesto europeo, ordinati dal meno al più energivoro. Questo scenario include i data center collocati nei principali poli in Europa (Paesi Bassi e Irlanda), così da fornire una rappresentazione più realistica dell'impatto energetico dei modelli. Per eccezione, DeepSeek viene sempre associato alla Cina come unico paese di riferimento, in quanto i suoi data center sono localizzati esclusivamente in quel contesto geografico.

Posizione	Modello	Energia (Wh/100 tok)
1 (meno energivoro)	GPT-4.1-nano	0,09
2	Mistral-7B	0,14
3	GPT-3.5-turbo	0,20
4	GPT-4.1-mini	0,36
5	Claude-3.7 Sonnet (2025)	0,41
6	Claude-Sonnet-4 (2025)	0,47
7	GPT-5-mini	0,51
8	o3	2,60
9	Mixtral-8×7B	2,92
10	GPT-5	3,05
11	GPT-4.1	3,20
12	GPT-4o	6,86
13	LLaMA-3.2-90B Vision Instruct Turbo	7,46
14	Claude-Opus-4.1 (2025)	9,04
15	Claude-Opus-4 (2025)	9,60

Posizione	Modello	Energia (Wh/100 tok)
16 (più energivoro)	DeepSeek-Reasoner	11,08

Nel contesto europeo, i modelli meno energivori sono quelli di piccola taglia o ottimizzati, come GPT-4.1 nano, Mistral-7B e GPT-3.5-turbo, tutti con valori inferiori a 0,3 Wh/100 token. Questi sistemi, pur con capacità più contenute rispetto ai grandi modelli, dimostrano un'elevata efficienza energetica e rappresentano benchmark di riferimento per un uso sostenibile.

In una fascia intermedia si collocano varianti compatte come GPT-4.1 mini e modelli ottimizzati della famiglia Claude (Sonnet-3.7 e Sonnet-4), con valori ancora sotto 0,5 Wh/100 token. A seguire, la curva dei consumi cresce sensibilmente: i modelli di fascia medio-alta (GPT-5, GPT-4.1, GPT-4o) arrivano a valori tra 3 e 7 Wh/100 token.

Nella parte alta della classifica, i modelli più energivori sono LLaMA-3.2-90B Vision Instruct Turbo e soprattutto i due Claude-Opus, fino ad arrivare a DeepSeek-Reasoner, che con oltre 11 Wh/100 token si conferma il sistema meno efficiente in termini energetici.

Ranking energetico – Globale

La Tabella sottostante mostra il ranking dei modelli nel contesto globale, ordinati dal meno al più energivoro. Questo scenario include i principali poli statunitensi (Virginia), così da fornire una rappresentazione più internazionale dell'impatto energetico dei modelli. Per eccezione, DeepSeek viene sempre associato alla Cina come unico paese di riferimento, in quanto i suoi data center sono localizzati esclusivamente in quel contesto geografico.

Posizione	Modello	Energia (Wh/100 tok)
1 (meno energivoro)	GPT-4.1-nano	0,09
2	Mistral-7B	0,14
3	GPT-3.5-turbo	0,20
4	GPT-4.1-mini	0,36
5	Claude-3.7 Sonnet (2025)	0,40
6	Claude-Sonnet-4 (2025)	0,46
7	GPT-5-mini	0,51
8	Mixtral-8×7B	0,63

Posizione	Modello	Energia (Wh/100 tok)
9	LLaMA-3.2-90B Vision Instruct Turbo	0,94
10	o3	2,61
11	GPT-5	3,06
12	GPT-4.1	3,22
13	GPT-4o	6,89
14	Claude-Opus-4.1 (2025)	8,89
15	Claude-Opus-4 (2025)	9,44
16 (più energivoro)	DeepSeek-Reasoner	11,08

Anche nello scenario globale, il quadro di fondo non cambia: GPT-4.1 nano, Mistral-7B e GPT-3.5-turbo si confermano i modelli meno energivori, tutti ben al di sotto di 0,3 Wh/100 token. Ciò dimostra come le varianti compatte mantengano un profilo di efficienza elevato anche in contesti infrastrutturali eterogenei. Nella fascia intermedia ritroviamo i modelli Claude in versione Sonnet, insieme a GPT-4.1 mini, mentre modelli come Mixtral-8×7B e LLaMA-3.2-90B Vision Instruct Turbo mostrano una posizione meno penalizzante rispetto allo scenario europeo, probabilmente per effetto di deployment su infrastrutture più recenti (GPU H100/H200 negli Stati Uniti).
Nella parte alta del ranking globale si confermano come più energivori Claude-Opus-4, Claude-Opus-4.1 e soprattutto DeepSeek-Reasoner, che rimane stabilmente il modello con il maggior consumo.

Il confronto tra scenario europeo e globale mostra una sostanziale stabilità agli estremi: i modelli meno energivori (GPT-4.1 nano, Mistral-7B, GPT-3.5-turbo) e i più energivori (Claude-Opus-4, Claude-Opus-4.1, DeepSeek-Reasoner) mantengono le stesse posizioni. Le differenze emergono nella fascia intermedia: ad esempio, LLaMA-3.2-90B Vision Instruct Turbo risulta molto più energivoro in Europa rispetto al contesto globale, probabilmente per l'impiego di infrastrutture meno efficienti nei data center europei. In generale, dunque, le variazioni riflettono differenze regionali di hosting, mentre l'efficienza intrinseca dei modelli resta invariata.

- Ranking di Emissioni

Il ranking delle emissioni riporta i modelli ordinati dal meno al più emissivo, sulla base dell'intensità media di CO_2 equivalente generata per 100 token. Questa rappresentazione consente di evidenziare i modelli più efficienti dal punto di vista delle emissioni e quelli con il maggiore impatto climatico nei due scenari considerati.

Ranking emissioni – Europa

La Tabella sottostante mostra il ranking dei modelli nel contesto europeo, ordinati dal meno al più emissivo (gCO_2e/100 token).

Posizione	Modello	CO_2 (gCO_2e/100 tok)
1 (meno emissivo)	GPT-4.1-nano	0,02
2	Mistral-7B	0,04
3	GPT-3.5-turbo	0,05
4	GPT-4.1-mini	0,09
5	Claude-3.7 Sonnet (2025)	0,10
6	Claude-Sonnet-4 (2025)	0,12
7	GPT-5-mini	0,13
8	o3	0,65
9	GPT-5	0,76
10	GPT-4.1	0,80
11	Mixtral-8×7B	0,82
12	GPT-4o	1,71
13	LLaMA-3.2-90B Vision Instruct Turbo	1,87
14	Claude-Opus-4.1 (2025)	2,26
15	Claude-Opus-4 (2025)	2,40
16 (più emissivo)	DeepSeek-Reasoner	6,10

Nel contesto europeo, i modelli meno emissivi sono GPT-4.1 nano (0,02 gCO_2e/100 token), Mistral-7B (0,04) e GPT-3.5-turbo (0,05). Tutti e tre si collocano sotto la soglia di 0,05 gCO_2e/100 token, confermando l'elevata efficienza delle architetture compatte e ottimizzate.

Subito dopo troviamo GPT-4.1 mini (0,09) e le varianti Claude Sonnet (3.7 e 4), con valori rispettivamente pari a 0,10 e 0,12 gCO_2e/100 token, seguiti da GPT-5 mini (0,13). Questi modelli mantengono un profilo emissivo molto contenuto, grazie a un miglior bilanciamento tra dimensioni architetturali e risorse di calcolo.

Nella fascia centrale della classifica emergono modelli più grandi e complessi come o3 (0,65), GPT-5 (0,76), GPT-4.1 (0,80) e Mixtral-8×7B (0,82), che si attestano tra 0,6 e 0,8 gCO_2e/100 token. Questi valori, pur non elevatissimi, mostrano un netto distacco rispetto ai modelli più compatti, segnalando l'aumento proporzionale di footprint al crescere della potenza computazionale.

Nella parte alta della classifica compaiono modelli multimodali e di ragionamento avanzato: GPT-4o (1,71) e LLaMA-3.2-90B Vision Instruct Turbo (1,87), che superano 1,5 gCO_2e/100 token. A seguire si collocano le due versioni Claude-Opus (2,26 e 2,40 gCO_2e/100 token), mentre il primato negativo spetta a DeepSeek-Reasoner, che con 6,10 gCO_2e/100 token rappresenta il sistema più emissivo nello scenario europeo.

Ranking emissioni – Globale

Posizione	Modello	CO_2 (gCO_2e/100 tok)
1 (meno emissivo)	GPT-4.1-nano	0,03
2	Mistral-7B	0,05
3	GPT-3.5-turbo	0,07
4	GPT-4.1-mini	0,14
5	Claude-3.7 Sonnet (2025)	0,15
6	Claude-Sonnet-4 (2025)	0,18
7	GPT-5-mini	0,19
8	Mixtral-8×7B	0,24
9	LLaMA-3.2-90B Vision Instruct Turbo	0,36
10	o3	0,99
11	GPT-5	1,16
12	GPT-4.1	1,22
13	GPT-4o	2,62
14	Claude-Opus-4.1 (2025)	3,38
15	Claude-Opus-4 (2025)	3,59
16 (più emissivo)	DeepSeek-Reasoner	6,10

Nel contesto globale, i modelli meno emissivi si confermano GPT-4.1 nano (0,03 gCO_2e/100 token), Mistral-7B (0,05) e GPT-3.5-turbo (0,07). Pur con valori leggermente più alti rispetto allo scenario europeo, restano nettamente i più efficienti, mantenendosi ben al di sotto della soglia di 0,1 gCO_2e/100 token.

A seguire troviamo GPT-4.1 mini (0,14) e le versioni Claude Sonnet (3.7 e 4), con valori compresi tra 0,15 e 0,18 gCO_2e/100 token, insieme a GPT-5 mini (0,19). Anche in questo caso si tratta di modelli che, grazie a ottimizzazioni architetturali e dimensioni più contenute, riescono a mantenere emissioni molto ridotte.

La parte centrale del ranking è occupata da Mixtral-8×7B (0,24) e LLaMA-3.2-90B Vision Instruct Turbo (0,36), che nello scenario globale risultano meno emissivi rispetto all'Europa, probabilmente grazie a

deployment in infrastrutture più recenti e a un mix energetico più favorevole. Più in alto compaiono o3 (0,99), GPT-5 (1,16) e GPT-4.1 (1,22), con valori attorno a 1 gCO_2e/100 token.

Nella parte alta si colloca GPT-4o (2,62), seguito dalle due varianti Claude-Opus (3,38 e 3,59 gCO_2e/100 token). Il modello più emissivo resta DeepSeek-Reasoner, che con 6,10 gCO_2e/100 token si conferma stabilmente in coda alla classifica anche nello scenario globale.

Il confronto tra scenario europeo e globale evidenzia una sostanziale stabilità agli estremi della classifica: i modelli meno emissivi restano le varianti compatte (GPT-4.1 nano, Mistral-7B, GPT-3.5-turbo), mentre i più emissivi si confermano Claude-Opus-4, Claude-Opus-4.1 e soprattutto DeepSeek-Reasoner, con valori superiori a 6 gCO_2e/100 token.

Le differenze emergono invece nelle posizioni intermedie. In particolare, LLaMA-3.2-90B Vision Instruct Turbo, tra i modelli più emissivi nello scenario europeo (1,87 gCO_2e/100 token), appare sensibilmente più efficiente a livello globale (0,36 gCO_2e/100 token). Un miglioramento analogo si osserva anche per Mixtral-8×7B, che scende da 0,82 a 0,24 gCO_2e/100 token. Queste variazioni suggeriscono che l'impatto emissivo non dipende soltanto dall'architettura del modello, ma anche dal contesto infrastrutturale e dal mix energetico dei data center in cui avviene l'esecuzione.

- Ranking Idrico

Il ranking idrico riporta i modelli ordinati dal meno al più idrovoro, sulla base dell'intensità media di acqua consumata per 100 token. Questa rappresentazione consente di individuare i modelli più efficienti nell'uso delle risorse idriche e quelli con il maggior impatto nei due scenari considerati.

Ranking idrico – Europa

La tabella sottostante mostra il ranking dei modelli nel contesto europeo, ordinati dal meno al più idrovoro (mL/100 token).

Posizione	Modello	Acqua (mL/100 tok)
1 (meno idrovoro)	GPT-4.1-nano	0,03
2	Mistral-7B	0,06
3	GPT-3.5-turbo	0,07
4	GPT-4.1-mini	0,12
5	Claude-3.7 Sonnet (2025)	0,16
6	GPT-5-mini	0,17

Posizione	Modello	Acqua (mL/100 tok)
7	Claude-Sonnet-4 (2025)	0,19
8	o3	0,88
9	GPT-5	1,04
10	GPT-4.1	1,09
11	Mixtral-8×7B	1,17
12	GPT-4o	2,33
13	LLaMA-3.2-90B Vision Instruct Turbo	2,54
14	Claude-Opus-4.1 (2025)	3,61
15	Claude-Opus-4 (2025)	3,84
16 (più idrovoro)	DeepSeek-Reasoner	6,65

Nello scenario europeo i modelli meno idrovori sono GPT-4.1 nano, Mistral-7B e GPT-3.5-turbo, tutti sotto 0,1 mL/100 token. Subito dopo seguono GPT-4.1 mini, Claude Sonnet (3.7 e 4) e GPT-5 mini, che restano al di sotto di 0,2 mL/100 token, mostrando un profilo idrico molto contenuto.

La fascia centrale della classifica è occupata da modelli più complessi come o3, GPT-5, GPT-4.1 e Mixtral-8×7B, con valori compresi tra 0,9 e 1,2 mL/100 token.

Nella parte alta si collocano invece modelli multimodali e di grandi dimensioni come GPT-4o (2,33) e LLaMA-3.2-90B Vision Instruct Turbo (2,54). I più idrovori restano le due versioni di Claude-Opus (3,6–3,8 mL/100 token) e soprattutto DeepSeek-Reasoner, che con 6,65 mL/100 token è di gran lunga il modello meno efficiente nell'uso della risorsa idrica.

Ranking idrico – Globale

La tabella sottostante riporta il ranking dei modelli nello scenario globale, dal meno al più idrovoro (mL/100 token).

Posizione	Modello	Acqua (mL/100 tok)
1 (meno idrovoro)	GPT-4.1-nano	0,03
2	Mistral-7B	0,06
3	GPT-3.5-turbo	0,07
4	GPT-4.1-mini	0,14
5	Claude-3.7 Sonnet (2025)	0,17
6	GPT-5-mini	0,19
7	Claude-Sonnet-4 (2025)	0,20

Posizione	Modello	Acqua (mL/100 tok)
8	Mixtral-8×7B	0,26
9	LLaMA-3.2-90B Vision Instruct Turbo	0,36
10	o3	0,99
11	GPT-5	1,16
12	GPT-4.1	1,22
13	GPT-4o	2,62
14	Claude-Opus-4.1 (2025)	3,73
15	Claude-Opus-4 (2025)	3,96
16 (più idrovoro)	DeepSeek-Reasoner	6,65

Anche nello scenario globale il quadro rimane stabile agli estremi: GPT-4.1 nano, Mistral-7B e GPT-3.5-turbo si confermano i modelli meno idrovori, mentre DeepSeek-Reasoner resta il più impattante con oltre 6,6 mL/100 token.

Nelle posizioni intermedie si osservano alcune variazioni rispetto all'Europa: Mixtral-8×7B e LLaMA-3.2-90B Vision Instruct Turbo migliorano sensibilmente, scendendo rispettivamente a 0,26 e 0,36 mL/100 token, a conferma del peso del contesto infrastrutturale. Al contrario, i modelli più complessi e multimodali come GPT-4o e i Claude-Opus restano nella parte alta della classifica, mostrando valori superiori ai 2–3,5 mL/100 token.

Il confronto tra i due scenari evidenzia un quadro complessivamente stabile agli estremi: i modelli compatti sono i meno idrovori, mentre DeepSeek e Claude-Opus restano i più impattanti. Le differenze emergono nella fascia intermedia, dove Mixtral-8×7B e LLaMA-3.2-90B Vision Instruct Turbo migliorano sensibilmente a livello globale. Ciò conferma che l'impatto idrico, come quello emissivo, è influenzato non solo dalle caratteristiche architetturali dei modelli, ma anche dal contesto di deployment e dalle condizioni dei data center.

3. Indice di Sintesi “Impatto Ambientale”

Per rendere più agevole la lettura dei risultati e facilitare un confronto trasversale tra le diverse dimensioni ambientali, è stato costruito un indice di sintesi ambientale. Questo indicatore integra in un’unica metrica le tre dimensioni considerate: consumo energetico (Wh/100 token), emissioni di CO_2 (gCO_2e/100 token) e consumo idrico (mL/100 token), offrendo una visione complessiva della sostenibilità dei modelli e consentendo di distinguere in maniera immediata i sistemi più efficienti da quelli con maggiore impronta ecologica.

Per garantire confrontabilità, ciascun indicatore è stato preliminarmente **normalizzato su scala 0–1** mediante la tecnica **min–max normalization**, comunemente impiegata negli indici compositi in ambito ambientale ed economico.
La formula adottata è la seguente:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

dove “x” rappresenta il valore osservato, mentre *min(x)* e *max(x)* sono rispettivamente i valori minimo e massimo riscontrati per la metrica in questione. In questo modo, il modello più efficiente assume valore 0, il meno efficiente valore 1, e gli altri si distribuiscono proporzionalmente tra questi estremi.

Una volta riportati su scala comune i tre indicatori, è stato calcolato l’indice sintetico come **media aritmetica semplice:**

$$\textit{Indice sintetico} = \frac{x'_{\text{energia}} + x'_{CO_2} + x'_{\text{acqua}}}{3}$$

Questo consente di ottenere un valore unico compreso tra 0 (modelli più sostenibili) e 1 (modelli meno sostenibili). Questo approccio ha il vantaggio di conservare le differenze reali tra i modelli, poiché riflette l’ampiezza delle distanze quantitative e non solo le posizioni relative ed inoltre rende le dimensioni confrontabili in quanto riportate sulla stessa scala, consentendo un’integrazione equilibrata di consumi energetici, emissioni di CO_2 e uso idrico.

- *Ranking Sintetico Normalizzato – Europa*

La Tabella seguente mostra il ranking sintetico normalizzato nello scenario europeo, calcolato come media dei valori normalizzati di energia, CO_2 e acqua. I modelli sono ordinati dal più efficiente al meno efficiente.

Posizione	Modello	Energia (Wh/100 tok)	CO_2 (gCO_2e/100 tok)	Acqua (mL/100 tok)	Indice sintetico
1 (più efficiente)	GPT-4.1 nano	0,09	0,02	0,03	0,000
2	Mistral-7B	0,14	0,04	0,06	0,004
3	GPT-3.5-turbo	0,20	0,05	0,07	0,007
4	GPT-4.1 mini	0,36	0,09	0,12	0,017
5	Claude-3.7 Sonnet (2025)	0,41	0,10	0,16	0,021
6	Claude-Sonnet-4 (2025)	0,47	0,12	0,19	0,025
7	GPT-5 mini	0,51	0,13	0,17	0,026
8	o3	2,60	0,65	0,88	0,153
9	GPT-5	3,05	0,76	1,04	0,181
10	Mixtral-8×7B	2,92	0,82	1,17	0,187
11	GPT-4.1	3,20	0,80	1,09	0,190
12	GPT-4o	6,86	1,71	2,33	0,414
13	LLaMA-3.2-90B Vision Instruct Turbo	7,46	1,87	2,54	0,451
14	Claude-Opus-4.1 (2025)	9,04	2,26	3,61	0,575
15	Claude-Opus-4 (2025)	9,60	2,40	3,84	0,611
16 (meno efficiente)	DeepSeek-Reasoner	11,08	6,10	6,65	1,000

- *Ranking Sintetico Normalizzato – Globale*

La Tabella seguente riporta il ranking sintetico normalizzato nello scenario globale, che considera i principali hub statunitensi (Virginia, USA). Anche in questo caso i modelli sono ordinati dal più efficiente al meno efficiente.

Posizione	Modello	Energia (Wh/100 tok)	CO_2 (gCO_2e/100 tok)	Acqua (mL/100 tok)	Indice sintetico
1 (più efficiente)	GPT-4.1 nano	0,09	0,03	0,03	0,000
2	Mistral-7B	0,14	0,05	0,06	0,004
3	GPT-3.5-turbo	0,20	0,07	0,07	0,008
4	GPT-4.1 mini	0,36	0,14	0,14	0,019

Posizione	Modello	Energia (Wh/100 tok)	CO_2 (gCO_2e/100 tok)	Acqua (mL/100 tok)	Indice sintetico
5	Claude-3.7 Sonnet (2025)	0,40	0,15	0,17	0,023
6	Claude-Sonnet-4 (2025)	0,46	0,18	0,20	0,027
7	GPT-5 mini	0,51	0,19	0,19	0,030
8	Mixtral-8×7B	0,63	0,24	0,26	0,039
9	LLaMA-3.2-90B Vision Instruct Turbo	0,94	0,36	0,36	0,060
10	o3	2,61	0,99	0,99	0,177
11	GPT-5	3,06	1,16	1,16	0,209
12	GPT-4.1	3,22	1,22	1,22	0,220
13	GPT-4o	6,89	2,62	2,62	0,478
14	Claude-Opus-4.1 (2025)	8,89	3,38	3,73	0,637
15	Claude-Opus-4 (2025)	9,44	3,59	3,96	0,677
16 (meno efficiente)	DeepSeek-Reasoner	11,08	6,10	6,65	1,000

Il confronto tra i ranking sintetici normalizzati europeo e globale evidenzia una sostanziale stabilità agli estremi. In entrambi gli scenari i modelli più efficienti risultano essere GPT-4.1 nano, Mistral-7B e GPT-3.5-turbo, con indici prossimi allo zero che ne confermano la sostenibilità trasversale rispetto a tutte le metriche ambientali. All'opposto, DeepSeek-Reasoner mantiene invariata la posizione di modello meno efficiente, seguito dalle due varianti Claude-Opus (4 e 4.1), a dimostrazione di un impatto sistematicamente elevato su energia, emissioni e consumo idrico.

Le differenze si osservano invece nelle posizioni intermedie, dove alcuni modelli mostrano miglioramenti significativi nello scenario globale. In particolare, Mixtral-8×7B e LLaMA-3.2-90B Vision Instruct Turbo presentano valori sensibilmente più bassi a livello globale (rispettivamente 0,039 e 0,060) rispetto allo scenario europeo (0,187 e 0,451). Questo risultato suggerisce che il deployment su infrastrutture più recenti o con mix energetici più favorevoli, come quelli statunitensi, possa incidere in maniera positiva sulla sostenibilità ambientale.

- *Efficienza Prestazionale dei Modelli*

Dopo aver analizzato l'impatto ambientale dei modelli, per una analisi più accurata, è fondamentale esaminare l'efficienza operativa dei modelli, intesa come la capacità di fornire risposte in tempi rapidi e con un livello di accuratezza adeguato. Questa dimensione è cruciale poiché consente di valutare un modello, non soltanto per quanto sia sostenibile dal punto di vista energetico, ma anche quanto sia efficiente in contesti reali, dove la velocità e la qualità delle risposte rappresentano fattori determinanti. Un modello può infatti risultare altamente efficiente dal punto di vista ambientale, ma se produce risposte con ritardi eccessivi o con bassa precisione, la sua applicabilità pratica viene fortemente ridimensionata.

Le metriche di efficienza nei modelli linguistici si articolano principalmente lungo due assi: la latenza e il throughput. La latenza, comunemente indicata con l'acronimo inglese TTFT (time-to-first-token), misura il tempo che intercorre tra l'invio del prompt e la produzione del primo token di risposta. In termini pratici, rappresenta il tempo di attesa percepito dall'utente prima che il modello "inizi a parlare". Una latenza bassa è quindi sinonimo di maggiore responsività, qualità particolarmente importante nelle applicazioni interattive, dove anche poche centinaia di millisecondi possono influenzare significativamente l'esperienza d'uso (*Artificial Analysis, 2025; Symbl.ai, 2024; Amazon Web Services, 2024).* Al contrario, latenze elevate compromettono la fluidità dell'interazione, riducendo l'efficacia del modello anche quando la qualità delle risposte è elevata.

Accanto alla latenza, la seconda metrica fondamentale è il throughput, solitamente espresso in token generati al secondo (token/s). Questo indicatore non si concentra sull'avvio della risposta, bensì sulla velocità con cui il modello continua a produrre testo una volta avviata la generazione. In altre parole, mentre la latenza cattura la rapidità iniziale, il throughput descrive la produttività sostenuta nel tempo. Nei casi d'uso in cui vengono generati output di grandi dimensioni, un throughput elevato è essenziale per contenere i tempi complessivi di esecuzione e garantire la scalabilità del servizio *(Bansal, 2023; OpenMetal, 2025; Dell Technologies, 2024).*

Le due metriche non sono alternative ma complementari. Un modello potrebbe avere una latenza molto ridotta ma un throughput basso, risultando ideale per risposte brevi ma inefficiente su testi lunghi. Viceversa, un modello con latenza elevata ma throughput elevato può essere meno adatto a scenari interattivi in tempo reale, ma efficace nella generazione di documenti complessi. La valutazione congiunta di latenza e throughput consente pertanto di ottenere un quadro più completo delle prestazioni.

Dal punto di vista metodologico, nei dataset raccolti per questa analisi sono disponibili entrambe le misure. È stata registrata ogni run, la latenza totale in millisecondi e il numero di token prodotti. Questo consente di calcolare la latenza media per token e, da essa, il throughput espresso in token/s.

4. Benchmark standardizzati per testare l'accuratezza dei modelli LLM

L'efficienza di un modello non può essere valutata unicamente in termini di velocità: un sistema che genera rapidamente ma commette errori frequenti non può essere considerato realmente efficiente. Per questa ragione, nell'analisi complessiva si integrano i valori di latenza e throughput con le metriche di accuratezza pubblicate nei principali studi di settore.
I benchmark costituiscono strumenti per la valutazione comparativa dei modelli linguistici. Essi sono definiti come insiemi di dati e protocolli di test concepiti per misurare in modo standardizzato la capacità dei sistemi di eseguire specifici compiti cognitivi, fornendo metriche oggettive e replicabili. La necessità di benchmark nasce dal fatto che le prestazioni dei modelli non possono essere adeguatamente comprese tramite valutazioni isolate o dataset proprietari: senza strumenti comuni, risulterebbe difficile stabilire criteri di confronto affidabili tra architetture, dataset di addestramento e approcci di ottimizzazione.

L'uso dei benchmark ha permesso di strutturare un linguaggio condiviso nella comunità scientifica, trasformando la valutazione delle capacità dei modelli in un processo sistematico e comparabile. Essi consentono di misurare in modo trasparente abilità diverse – dalla conoscenza enciclopedica al ragionamento matematico, fino alla robustezza logica – e di costruire classifiche che permettono di tracciare i progressi nel tempo e fra generazioni di modelli (Liang et al., 2022).

Un ulteriore aspetto di rilievo è che i benchmark mettono in luce non solo i punti di forza dei modelli, ma anche le loro debolezze strutturali. Studi recenti hanno mostrato che le prestazioni su alcuni dataset possono risultare sensibili a variazioni apparentemente marginali, come l'ordine delle risposte in MMLU, rivelando così limiti di robustezza (Gupta et al., 2024). Analogamente, la diffusione di "benchmark saturi", in cui i modelli raggiungono punteggi prossimi al 100%, ha sollevato interrogativi sulla capacità di tali strumenti di discriminare efficacemente tra sistemi di nuova generazione. Per rispondere a queste criticità, sono stati introdotti benchmark più complessi e difficilmente contaminabili, come FrontierMath per il ragionamento matematico avanzato (Glazer et al., 2024).

Da un punto di vista metodologico, i benchmark sono anche strumenti cruciali per integrare valutazioni quantitative (come velocità o consumo energetico) con valutazioni qualitative sulle capacità cognitive. Questo consente di superare approcci monodimensionali e costruire indicatori compositi di prestazione e sostenibilità. Come sottolineato da Liang et al. (2022), il valore dei benchmark risiede nella loro funzione di "valutazione olistica", che permette di considerare dimensioni diverse – accuratezza, robustezza, affidabilità – in un quadro unitario.

In sintesi, i benchmark rappresentano la base scientifica su cui poggia gran parte della ricerca sugli LLM. Essi non solo offrono strumenti comparativi per misurare le prestazioni, ma contribuiscono a identificare le aree di miglioramento e a guidare lo sviluppo di nuovi modelli più solidi ed efficienti. In un panorama in continua evoluzione, caratterizzato da modelli di scala crescente e da esigenze sempre più stringenti di efficienza e affidabilità, il ruolo dei benchmark è destinato a rimanere centrale per assicurare una valutazione trasparente, rigorosa e condivisa.

4.1 Principali benchmark per la valutazione degli LLM

I benchmark più utilizzati nella letteratura scientifica sono progettati per valutare dimensioni diverse delle capacità dei modelli linguistici: conoscenze enciclopediche, ragionamento matematico, inferenza scientifica, coerenza narrativa e veridicità delle risposte. Nella sezione seguente vengono descritti i principali strumenti di riferimento, con le loro caratteristiche e fonti ufficiali.

1. MMLU – Massive Multitask Language Understanding

Il benchmark MMLU è stato introdotto per valutare la **copertura enciclopedica e specialistica** dei modelli linguistici. Comprende oltre 15.000 domande a scelta multipla distribuite su 57 discipline accademiche, dalle scienze umane alla medicina. La metrica di riferimento è l'accuratezza percentuale. MMLU rappresenta uno dei primi tentativi sistematici di misurare la "conoscenza generale" dei modelli, ed è oggi considerato uno standard di riferimento nella letteratura. (Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. e Steinhardt, J. (2020) *Measuring Massive Multitask Language Understanding*. arXiv preprint arXiv:2009.03300.)

2. GSM8K – Grade School Math 8K

Il benchmark GSM8K è stato proposto da OpenAI per misurare la capacità dei modelli di risolvere **problemi matematici elementari** con ragionamento step-by-step. Contiene circa 8.500 problemi a testo libero, ciascuno con soluzione numerica finale. La peculiarità di GSM8K è che non richiede solo la risposta, ma la costruzione di un processo di calcolo corretto, diventando quindi un test robusto di ragionamento numerico. (Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, Ł. e Miller, J. (2021) *Training Verifiers to Solve Math Word Problems*. arXiv preprint arXiv:2110.14168.)

3. ARC – AI2 Reasoning Challenge

ARC è stato introdotto dall'Allen Institute for AI per valutare la capacità di **ragionamento scientifico di base**. Comprende circa 7.787 domande a scelta multipla, suddivise in due set: *Easy* e *Challenge*.

Quest'ultimo è progettato per includere quesiti che richiedono inferenze non banali e quindi rappresenta un banco di prova importante per la generalizzazione. La metrica principale è l'accuratezza. (Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C. e Tafjord, O. (2018) *Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge*. arXiv preprint arXiv:1803.05457.)

4. HellaSwag

HellaSwag è un benchmark pensato per valutare la capacità dei modelli di **ragionamento di buon senso e completamento narrativo**. I test presentano brevi testi che devono essere completati scegliendo la frase più plausibile fra opzioni generate in modo da apparire realistiche. Questo lo rende un test particolarmente difficile, perché richiede di distinguere tra alternative apparentemente corrette. (Zellers, R., Bisk, Y., Farhadi, A. e Choi, Y. (2019) *HellaSwag: Can a Machine Really Finish Your Sentence?*. arXiv preprint arXiv:1905.07830.)

5. TruthfulQA

TruthfulQA è stato sviluppato per misurare la capacità dei modelli di fornire risposte veritiere evitando di riprodurre errori o "allucinazioni". Il dataset include domande progettate per indurre risposte tipicamente scorrette o fuorvianti, così da testare il grado di affidabilità del modello. La metrica è il truthfulness score, che valuta quanto una risposta sia effettivamente corretta e non ingannevole. (Lin, S., Hilton, J. e Evans, O. (2021) *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. arXiv preprint arXiv:2109.07958.)

6. BIG-Bench Hard (BBH)

BIG-Bench Hard è un sottoinsieme del benchmark BIG-Bench, contenente 23 compiti selezionati perché particolarmente resistenti anche ai modelli più avanzati**.** Include prove di logica, ragionamento matematico complesso e compiti linguistici che richiedono *chain-of-thought reasoning*. È uno degli strumenti più utilizzati per testare i limiti superiori delle capacità di ragionamento. (Suzgun, M., Scao, T.L., Shuster, K., Tay, Y., Wu, Y., Zoph, B. e Wei, J. (2022) *Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them*. arXiv preprint arXiv:2210.09261.)

7. FrontierMath

FrontierMath è un benchmark di nuova generazione, sviluppato per valutare abilità matematiche avanzate che non risultano più adeguatamente testate da GSM8K. Comprende problemi originali, costruiti per evitare contaminazioni con i dataset di addestramento e verificabili automaticamente. È stato introdotto per superare la saturazione dei benchmark tradizionali e mantenere la capacità discriminante fra modelli di frontiera. (Glazer, E., Choi, J., Steinhardt, J., Wei, J. e Zhou, D. (2024) *FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI*. arXiv preprint arXiv:2406.14615.)

4.2 Individuazione dei punteggi MMLU per i modelli LLM analizzati

Dopo avere presentato le caratteristiche dei principali benchmark utilizzati per valutare l'accuratezza dei modelli linguistici di grandi dimensioni, si è scelto di concentrare l'analisi comparativa sul punteggio MMLU. Questo benchmark, infatti, rappresenta oggi lo standard più diffuso e riconosciuto per misurare la copertura enciclopedica e specialistica dei modelli, abbracciando oltre cinquanta discipline accademiche e offrendo un'indicazione sintetica e comparabile delle capacità di conoscenza e ragionamento generale (Hendrycks et al., 2020).

Al fine di integrare la dimensione di efficienza ambientale già calcolata con indicatori di prestazione cognitiva, è stato dunque necessario individuare un valore di MMLU per ciascuno dei modelli considerati nello studio. L'obiettivo non è soltanto di fornire un confronto delle prestazioni, ma anche di creare una base per la costruzione di un indicatore composito che tenga insieme accuratezza e sostenibilità computazionale, superando valutazioni parziali e consentendo una lettura più olistica dell'impatto dei modelli.

La tabella seguente riassume i punteggi MMLU raccolti per i modelli in esame, distinguendo fra valori ufficiali forniti dai produttori, risultati derivati da fonti tecniche indipendenti e stime metodologiche sviluppate ad hoc.

Modello	MMLU (%)	Fonte	Commenti
GPT-5	87,0	ChatBase blog	Valore su **MMLU-Pro**, benchmark più impegnativo dello standard
GPT-5-mini	85,0	Elaborazione su dati OpenAI (*GPT-5 System Card*, 2025)	Nessun dato ufficiale pubblicato; stima ottenuta scalando GPT-5 con il rapporto medio mini/main osservato su AIME '25, GPQA, HMMT 2025 e MMMU.
GPT-4.1	90,2	OpenAI (2024-11-20)	Fonte ufficiale
GPT-4.1 mini	87,5	OpenAI (2024-11-20)	Fonte ufficiale
GPT-4.1 nano	80,1	OpenAI (2024-11-20)	Fonte ufficiale
GPT-4o	85,7	OpenAI (2024-11-20)	Fonte ufficiale
O3	85,6	Vals.ai	Valore su **MMLU-Pro** (più difficile, ~12k domande, 10 opzioni)
GPT-3.5-turbo	70,0	HuggingFace leaderboard / community	Valore stimato, non ufficiale OpenAI
Claude-3.7 Sonnet (2025)	82,7	Vals.ai (MMLU-Pro)	Fonte indipendente, riferito a MMLU-Pro

Modello	MMLU (%)	Fonte	Commenti
Claude-Sonnet-4 (2025)	86,5	DataCamp	Sintesi divulgativa
Claude-Opus-4 (2025)	88,8	DataCamp	Sintesi divulgativa
Claude-Opus-4.1 (2025)	87,8	Vals.ai (vals.ai)	Valore su **MMLU-Pro**; molto elevato (87,8 non-thinking, 87,6 thinking)
DeepSeek-Reasoner (R1)	90,8 (std) / 84,0 (Pro)	DeepSeek arXiv (2025)	Fonte primaria; valori ufficiali su MMLU standard e MMLU-Pro
LLaMA-3.2-90B Vision Turbo	86,0	Meta (HuggingFace / Medium)	Fonte ufficiale Meta, MMLU 0-shot con chain-of-thought
Mistral-7B	62,6	Stima basata su dichiarazioni Mistral + community	Stima coerente con "on par con LLaMA-34B" (~63–65 %)
Mixtral-8×7B	71,3	Klu.ai / Openlaboratory.ai	MMLU 5-shot, dato indipendente

Per quanto riguarda la famiglia **GPT-4.1**, i punteggi MMLU sono stati ricavati direttamente dalle tabelle ufficiali OpenAI pubblicate nel novembre 2024. In particolare, GPT-4.1 raggiunge un punteggio del 90,2%, mentre le varianti più compatte, GPT-4.1 mini e GPT-4.1 nano, si attestano rispettivamente a 87,5% e 80,1%. Questi valori, forniti da fonte primaria, possono essere utilizzati senza alcun intervento di stima o adattamento (OpenAI, *Introducing GPT-4.1 in the API*, 2024, s.p.). Anche per GPT-4o, modello multimodale introdotto nello stesso periodo, è stato possibile fare riferimento al valore ufficiale di 85,7% pubblicato nella stessa tabella (OpenAI, *Introducing GPT-4.1 in the API*, 2024, s.p.).

Per quanto riguarda **O3**, non esistono al momento risultati ufficiali sul benchmark MMLU standard. È stato però rilevato un punteggio di 85,6% su MMLU-Pro, riportato in una leaderboard tecnica indipendente (Vals.ai, *MMLU-Pro Leaderboard*, 2025, s.p.). Tale valore è stato mantenuto in tabella, con la chiara indicazione che MMLU-Pro rappresenta una versione più difficile del test e pertanto non è perfettamente comparabile con MMLU standard, ma offre comunque un'indicazione utile per posizionare il modello.

Il caso di **GPT-5** è analogo: la letteratura tecnica secondaria riporta un valore pari a 87,0% su MMLU-Pro (ChatBase, *Is ChatGPT accurate?*, 2025, s.p.). Questo dato, pur non sostituendo un punteggio ufficiale standard, è stato conservato con l'opportuna etichetta, così da fornire almeno un'indicazione di ordine di grandezza.

Diversa è stata la procedura adottata per **GPT-5 mini,** per il quale OpenAI non ha pubblicato alcun punteggio MMLU. Per evitare un vuoto informativo, si è scelto di stimare il valore partendo dal confronto fra i punteggi ufficiali mini/main disponibili su altri benchmark (AIME '25, GPQA Diamond, HMMT 2025

e MMMU), dai quali si ricava un rapporto medio mini/main pari a circa 0,958. Applicando questo fattore a un intervallo plausibile di performance per GPT-5 main (90–92%), si ottiene una stima di GPT-5 mini compresa tra 86% e 88%. Tale intervallo è riportato in tabella come valore stimato, con nota metodologica che spiega il procedimento (OpenAI, *GPT-5 System Card*, 2025, s.p.; OpenAI, *Introducing GPT-5 for Developers*, 2025, s.p.).

Per **GPT-3.5-turbo**, OpenAI non ha mai pubblicato valori ufficiali. Le leaderboard comunitarie di Hugging Face riportano però un punteggio attorno al 70% su MMLU (5-shot). Questo valore è stato assunto come indicativo e inserito in tabella, con la chiara avvertenza che non si tratta di un dato ufficiale (Hugging Face, *Open LLM Leaderboard discussions*, 2024–2025, s.p.).

Per quanto riguarda la famiglia **Claude**, le fonti disponibili sono più eterogenee. Claude-3.7 Sonnet presenta un punteggio di 82,7% su MMLU-Pro, tratto da leaderboard indipendenti (Vals.ai, *MMLU-Pro Leaderboard*, 2025, s.p.), mentre Claude-Sonnet-4 e Claude-Opus-4 hanno rispettivamente 86,5% e 88,8% secondo sintesi divulgative (DataCamp, *Claude-4 Overview*, 2024–2025, s.p.). Infine, Claude-Opus-4.1 è riportato a 87,8% (87,6% in modalità thinking) sempre su MMLU-Pro in leaderboard indipendenti, valore utilizzato con l'avvertenza della non piena confrontabilità (Vals.ai, *MMLU-Pro Leaderboard*, 2025, s.p.).

Il caso di **DeepSeek-Reasoner (R1)** è invece chiaro: i punteggi ufficiali riportati nel paper accademico indicano un MMLU standard pari a 90,8% e un MMLU-Pro pari a 84,0%. Trattandosi di fonte primaria, i valori sono stati adottati senza modifiche (Guo, Guoqi et al., *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via RL*, 2025, s.p.).

Per **LLaMA-3.2-90B Vision Instruct Turbo**, il punteggio di 86,0% su MMLU è stato ricavato dalle risorse ufficiali Meta e Hugging Face. Il test è condotto in configurazione 0-shot con chain-of-thought; l'informazione è conservata in tabella, segnalando il setup adottato (Meta, *LLaMA-3.2-90B Vision – Model Card*, 2025, s.p.).

Infine, per **Mistral-7B e Mixtral-8×7B** non sono disponibili dati ufficiali completi. Mistral-7B viene dichiarato dall'azienda "superiore a LLaMA-2-13B" e "alla pari con LLaMA-34B". Dato che questi modelli presentano punteggi MMLU noti rispettivamente intorno a 54–55% e 63–65%, è stata stimata per Mistral-7B una performance di circa 62,6%, coerente con i dati comunitari reperibili (Mistral AI, *Announcing Mistral-7B*, 2023, s.p.). Per Mixtral-8×7B invece sono reperibili stime indipendenti su MMLU (5-shot) pari a 71,3% (Klu.ai, *Mixtral-8×7B Benchmark Note*, 2024–2025, s.p.), mantenute in tabella come indicazione di ordine di grandezza.

4.3 Costruzione dell'Indicatore Composito

L'analisi condotta finora ha messo in luce due dimensioni fondamentali: da un lato l'accuratezza dei modelli linguistici, misurata attraverso benchmark consolidati come MMLU, dall'altro l'impatto ambientale dell'inferenza, calcolato in termini di consumo energetico, emissioni di CO_2 equivalente e uso idrico. Entrambe queste dimensioni risultano centrali per una valutazione complessiva degli LLM, ma, considerate separatamente, restituiscono un quadro parziale: un modello può essere estremamente accurato ma molto impattante sul piano ambientale, oppure al contrario sostenibile ma con prestazioni cognitive limitate.

Per superare questa frammentazione, la letteratura sugli indici compositi suggerisce la costruzione di misure sintetiche capaci di integrare grandezze eterogenee in un unico indicatore, così da fornire un criterio comparativo immediato (OECD, *Handbook on Constructing Composite Indicators*, 2008, s.p.). Applicato al contesto degli LLM, questo approccio consente di rappresentare in maniera unitaria la tensione tra performance e sostenibilità, riconducendo il confronto dei modelli a una metrica condivisa.

La costruzione di un indicatore composito risponde dunque a due obiettivi principali. In primo luogo, consente di individuare i modelli che rappresentano il miglior compromesso tra accuratezza e impatto ambientale, ovvero quelli che riescono a massimizzare le capacità cognitive minimizzando al contempo l'impronta ecologica. In secondo luogo, permette di tradurre risultati complessi e multidimensionali in una forma interpretabile anche da decisori e stakeholder non specialisti, facilitando l'uso pratico dell'analisi (Nardo et al., *Handbook on Composite Indicators*, 2005, s.p.).

In questo lavoro, l'indicatore composito è stato concepito come strumento di sintesi per mettere in relazione i due pilastri della valutazione degli LLM: la capacità di fornire risposte accurate e l'efficienza ecologica delle operazioni di inferenza. Esso rappresenta quindi la fase conclusiva del percorso analitico, integrando i risultati precedentemente ottenuti sui ranking ambientali e sui benchmark di accuratezza in una misura unitaria e comparabile.

4.3.1 Variabili considerate

La costruzione dell'indicatore composito richiede l'individuazione preliminare delle variabili da includere. In questo lavoro si è scelto di integrare due dimensioni principali: accuratezza cognitiva ed impatto ambientale.

La prima dimensione è rappresentata dai punteggi ottenuti dai modelli al benchmark MMLU (Massive Multitask Language Understanding). MMLU è stato scelto come proxy dell'accuratezza complessiva poiché

fornisce una misura sintetica e comparabile tra modelli diversi, evitando di frammentare l'analisi in una molteplicità di benchmark parziali.

La seconda dimensione è rappresentata dall'indice ambientale sintetico sviluppato nella prima parte del lavoro. Questo indice integra tre componenti normalizzate:

- consumo energetico (Wh/100 token),
- emissioni di CO_2 equivalente (gCO_2e/100 token),
- uso idrico (mL/100 token).

Le tre metriche sono state armonizzate su scala comune (0–1) tramite normalizzazione min–max e successivamente aggregate come media aritmetica, in linea con la prassi metodologica consolidata negli indici compositi (OECD, Handbook on Constructing Composite Indicators, 2008, s.p.). L'indice risultante permette di rappresentare la sostenibilità ambientale in forma sintetica, conservando al tempo stesso le differenze relative tra modelli.
La rapidità di inferenza (latenza e throughput), pur avendo un ruolo rilevante nelle applicazioni pratiche, non è stata inclusa direttamente nell'indice composito. È stata invece riservata una sezione di analisi dedicata in un secondo momento, così da evidenziare eventuali trade-off tra velocità, accuratezza e impatto ambientale senza compromettere la coerenza metodologica del calcolo.

4.3.2 Normalizzazione della variabile Accuratezza Cognitiva

L'impatto ambientale è già stato sintetizzato e normalizzato attraverso la costruzione di un indice basato su consumo energetico, emissioni di CO_2 e uso idrico, riportati su scala comune 0–1 mediante min–max normalization. Di conseguenza, in questa fase l'unica variabile che necessita di normalizzazione è l'accuratezza, misurata tramite i punteggi MMLU.

Per rendere i valori comparabili e integrabili nell'indice composito, anche l'accuratezza è stata riportata su scala 0–1 utilizzando la seguente formula:

$$MMLU_i' = \frac{MMLU_i - \min(MMLU)}{\max(MMLU) - \min(MMLU)}$$

dove $MMLU_i$ rappresenta il punteggio osservato per il modello $_i$, mentre min (MMLU) e max (MMLU) sono rispettivamente il valore minimo e massimo riscontrati tra i modelli considerati.

La tabella seguente riporta i punteggi MMLU normalizzati per ciascun modello considerato. La normalizzazione è stata effettuata considerando come valore minimo corrispondente a Mistral-7B (62,6%) e valore massimo a DeepSeek-Reasoner (90,8%).

Modello	MMLU (%)	MMLU′ normalizzato
GPT-5	87,0	0,865
GPT-5-mini	87,0	0,865
GPT-4.1	90,2	0,979
GPT-4.1 mini	87,5	0,883
GPT-4.1 nano	80,1	0,621
GPT-4o	85,7	0,819
O3	85,6	0,816
GPT-3.5-turbo	70,0	0,262
Claude-3.7 Sonnet (2025)	82,7	0,717
Claude-Sonnet-4 (2025)	86,5	0,848
Claude-Opus-4 (2025)	88,8	0,928
Claude-Opus-4.1 (2025)	87,8	0,894
DeepSeek-Reasoner (R1)	90,8	1,000
LLaMA-3.2-90B Vision Turbo	86,0	0,830
Mistral-7B	62,6	0,000
Mixtral-8×7B	71,3	0,309

4.3.3 Trasformazione dell'indice ambientale in sostenibilità

L'impatto ambientale è già stato sintetizzato e normalizzato attraverso la costruzione di un indice basato su consumo energetico, emissioni di CO_2 e uso idrico. Per poter integrare questa dimensione con l'accuratezza normalizzata, è stato necessario invertire la direzione dell'indicatore, trasformandolo in una misura di **sostenibilità**. Questa operazione, raccomandata dalla letteratura sugli indici compositi (OECD, *Handbook on Constructing Composite Indicators*, 2008), permette di rendere coerenti le variabili: in entrambi i casi, valori elevati corrispondono a prestazioni migliori (maggiore accuratezza o maggiore sostenibilità).

La trasformazione non altera la distribuzione relativa dei valori, ma ne ribalta l'interpretazione, semplificando la lettura comparata e l'aggregazione nell'indice composito finale.

La formula utilizzata per ottenere la sostenibilità a partire dall'indice ambientale è la seguente:

$$Sostenibilita_i = 1 - IndiceAmbientale_i$$

dove:

- $IndiceAmbientale_i$ è il valore normalizzato (0–1) dell'indice sintetico per il modello i;
- $Sostenibilita_i$ è il nuovo valore armonizzato, che assume valori prossimi a 1 per i modelli più efficienti e prossimi a 0 per i meno sostenibili.

Questa trasformazione ha il vantaggio di conservare la distribuzione relativa dei valori, invertendone soltanto la direzione.

Applicando la trasformazione all'indice ambientale, si ottiene una nuova graduatoria di sostenibilità, distinta per scenario europeo e globale.

La tabella seguente riporta i valori calcolati per ciascun modello nello scenario europeo.

Modello	Indice ambientale	Sostenibilità
GPT-4.1 nano	0,000	1,000
Mistral-7B	0,004	0,996
GPT-3.5-turbo	0,007	0,993
GPT-4.1 mini	0,017	0,983
Claude-3.7 Sonnet (2025)	0,021	0,979
Claude-Sonnet-4 (2025)	0,025	0,975
GPT-5 mini	0,026	0,974
o3	0,153	0,847
GPT-5	0,181	0,819
Mixtral-8×7B	0,187	0,813
GPT-4.1	0,190	0,810
GPT-4o	0,414	0,586
LLaMA-3.2-90B Vision Turbo	0,451	0,549
Claude-Opus-4.1 (2025)	0,575	0,425
Claude-Opus-4 (2025)	0,611	0,389
DeepSeek-Reasoner	1,000	0,000

La tabella seguente riporta i valori calcolati per ciascun modello nello scenario globale.

Modello	Indice ambientale	Sostenibilità
GPT-4.1 nano	0,000	1,000
Mistral-7B	0,004	0,996

Modello	Indice ambientale	Sostenibilità
GPT-3.5-turbo	0,008	0,992
GPT-4.1 mini	0,019	0,981
Claude-3.7 Sonnet (2025)	0,023	0,977
Claude-Sonnet-4 (2025)	0,027	0,973
GPT-5 mini	0,030	0,970
Mixtral-8×7B	0,039	0,961
LLaMA-3.2-90B Vision Turbo	0,060	0,940
o3	0,177	0,823
GPT-5	0,209	0,791
GPT-4.1	0,220	0,780
GPT-4o	0,478	0,522
Claude-Opus-4.1 (2025)	0,637	0,363
Claude-Opus-4 (2025)	0,677	0,323
DeepSeek-Reasoner	1,000	0,000

4.4 Indicatore Composito Accuratezza – Sostenibilità

Completata la normalizzazione delle variabili considerate, è stato possibile integrare l'accuratezza cognitiva, misurata attraverso i punteggi MMLU′, e la sostenibilità ambientale, ottenuta dall'inversione dell'indice di impatto, in un indicatore composito.

Dal punto di vista tecnico, l'indicatore è stato calcolato come media aritmetica semplice dei due valori normalizzati, secondo la seguente formula:

$$IC_i = 0.5 \cdot MMLU_i' + 0.5 \cdot S_i$$

dove MMLU rappresenta il punteggio normalizzato di accuratezza del modello i e S_i il relativo indice di sostenibilità.

L'aggregazione è stata effettuata attribuendo lo stesso peso alle due variabili. In questo modo ciascuna delle due componenti contribuisce equamente alla valutazione finale, evitando di imporre arbitrariamente un peso maggiore all'una o all'altra. Il risultato, compreso tra 0 e 1, sintetizza il valore complessivo del modello i nella duplice dimensione considerata.

L'adozione di questo approccio consente di superare la frammentazione analitica dei singoli ranking, fornendo una graduatoria unica e immediatamente interpretabile. Nei paragrafi successivi vengono presentati

i risultati dello scenario europeo e globale, così da cogliere eventuali differenze legate al contesto infrastrutturale e alla localizzazione geografica dei data center.

Indicatore Accuratezza – Sostenibilità: Scenario Europeo

La tabella seguente riporta i risultati dell'indicatore composito **accuratezza–sostenibilità (50/50)** calcolato per lo scenario europeo. I valori sono ordinati dal modello più performante al meno performante.

Posizione	Modello	MMLU′	Sostenibilità	IC (50/50)
1	GPT-4.1 mini	0.883	0.983	0.933
2	GPT-4.1 nano	0.621	1.000	0.811
3	Claude-Sonnet-4 (2025)	0.848	0.975	0.912
4	Claude-3.7 Sonnet (2025)	0.717	0.979	0.848
5	GPT-5 mini	0.865	0.974	0.920
6	GPT-5	0.865	0.819	0.842
7	GPT-4.1	0.979	0.810	0.895
8	GPT-3.5-turbo	0.262	0.993	0.628
9	Mixtral-8×7B	0.309	0.813	0.561
10	LLaMA-3.2-90B Vision Turbo	0.830	0.549	0.690
11	GPT-4o	0.819	0.586	0.703
12	O3	0.816	0.847	0.832
13	Claude-Opus-4.1 (2025)	0.894	0.425	0.660
14	Claude-Opus-4 (2025)	0.928	0.389	0.659
15	DeepSeek-Reasoner (R1)	1.000	0.000	0.500
16	Mistral-7B	0.000	0.996	0.498

Indicatore Accuratezza – Sostenibilità: Scenario Globale

La tabella seguente riporta i risultati dell'indicatore composito **accuratezza–sostenibilità (50/50)** calcolato per lo scenario globale. I valori sono ordinati dal modello più performante al meno performante.

Posizione	Modello	MMLU′	Sostenibilità	IC (50/50)
1	GPT-4.1 mini	0.883	0.981	0.932
2	Claude-Sonnet-4 (2025)	0.848	0.973	0.911
3	GPT-5 mini	0.865	0.970	0.918

Posizione	Modello	MMLU′	Sostenibilità	IC (50/50)
4	GPT-4.1	0.979	0.780	0.880
5	GPT-5	0.865	0.791	0.828
6	Claude-3.7 Sonnet (2025)	0.717	0.977	0.847
7	GPT-4.1 nano	0.621	1.000	0.811
8	O3	0.816	0.823	0.820
9	GPT-4o	0.819	0.522	0.671
10	LLaMA-3.2-90B Vision Turbo	0.830	0.940	0.885
11	Claude-Opus-4.1 (2025)	0.894	0.363	0.629
12	Claude-Opus-4 (2025)	0.928	0.323	0.626
13	Mixtral-8×7B	0.309	0.961	0.635
14	GPT-3.5-turbo	0.262	0.992	0.627
15	DeepSeek-Reasoner (R1)	1.000	0.000	0.500
16	Mistral-7B	0.000	0.996	0.498

I risultati evidenziano il ruolo trainante dei modelli compatti, in particolare le varianti “mini” e “sonnet”. GPT-4.1 mini si conferma come il modello più competitivo in entrambi gli scenari: unisce accuratezza molto elevata a una sostenibilità prossima all’unità, collocandosi stabilmente in testa alla graduatoria. Un andamento analogo si osserva per Claude-Sonnet-4 e GPT-5 mini, che restano costantemente nella fascia alta. Ciò mostra come le strategie di ottimizzazione architetturale abbiano permesso di bilanciare prestazioni cognitive e impatto ambientale.

All’interno della stessa famiglia emergono però differenze rilevanti. GPT-4.1 nano, pur eccellendo in sostenibilità, presenta un’accuratezza più contenuta e si posiziona quindi in area intermedia. Lo stesso accade a Mistral-7B, che figura tra i sistemi meno impattanti ma con livelli di accuratezza troppo bassi per primeggiare. Questi casi dimostrano come la sostenibilità, da sola, non sia sufficiente a garantire competitività: la qualità delle risposte resta un requisito imprescindibile.

Al contrario, i modelli di maggiori dimensioni, e in particolare le varianti Claude-Opus, mostrano lo squilibrio opposto: accuratezza molto alta, in alcuni casi oltre 0,92, ma sostenibilità penalizzante che li relega in fondo alla classifica. Anche DeepSeek-Reasoner rappresenta un esempio emblematico: nonostante il punteggio massimo in accuratezza, la sostenibilità nulla lo colloca tra gli ultimi. In entrambi i casi, l’assenza di equilibrio compromette il valore complessivo.

Un discorso a parte meritano i modelli multimodali e di fascia intermedia. GPT-4.1 mantiene un profilo equilibrato grazie a un’accuratezza prossima al massimo e a una sostenibilità discreta. GPT-4o e O3, pur

mostrando valori accettabili in entrambe le dimensioni, non eccellono in nessuna delle due, posizionandosi nella fascia mediana come esempi di solidità senza leadership.

Il confronto tra scenario europeo e globale conferma la stabilità delle posizioni di vertice e di coda. Modelli come GPT-4.1 mini, Claude-Sonnet-4 e GPT-5 mini restano leader in entrambi i contesti, mentre DeepSeek-Reasoner e le varianti Claude-Opus continuano a occupare le ultime posizioni. Alcuni miglioramenti si osservano invece nella fascia centrale: LLaMA-3.2-90B Vision e Mixtral-8×7B risultano leggermente più competitivi nello scenario globale, probabilmente grazie a mix energetici più favorevoli, senza tuttavia modificare gli equilibri generali.

In sintesi, i modelli più competitivi sono quelli che riescono a bilanciare le due dimensioni, evitando squilibri estremi. Le varianti ottimizzate come GPT-4.1 mini, Claude-Sonnet-4 e GPT-5 mini dimostrano che è possibile conciliare accuratezza e sostenibilità, mentre i casi opposti – come DeepSeek-Reasoner o Mistral-7B – mostrano che eccellere in una sola dimensione non basta a garantire un buon posizionamento.

4.5 Indicatore Accuratezza – Rapidità – Sostenibilità

Dopo avere costruito un indicatore composito basato su accuratezza e sostenibilità, si è ritenuto opportuno ampliare l'analisi includendo anche la variabile della rapidità di inferenza. Questa dimensione risulta particolarmente interessante perché consente di valutare i modelli non solo sul piano della qualità delle risposte e dell'impatto ambientale, ma anche in termini di efficienza operativa. La velocità con cui un sistema elabora un prompt e genera l'output è infatti un fattore determinante nella percezione di usabilità e nella possibilità di impiego in scenari applicativi reali, come l'assistenza interattiva o la produzione di testi estesi.

Integrare la rapidità nell'analisi permette quindi di ottenere un indicatore più completo, capace di rappresentare la tensione tra tre obiettivi: garantire risposte accurate, ridurre al minimo l'impronta ecologica e fornire risultati in tempi compatibili con le esigenze degli utenti. L'indicatore esteso è stato concepito proprio per sintetizzare queste tre dimensioni, fornendo una misura unica che renda immediatamente comparabili i modelli sulla base del loro equilibrio complessivo.

4.5.1 Variabile considerate

La costruzione dell'indicatore esteso richiede di integrare tre dimensioni principali, già definite nel percorso analitico ma qui ricondotte a un quadro unitario.

La prima variabile è l'accuratezza cognitiva, misurata attraverso i punteggi ottenuti dai modelli sul benchmark MMLU. Questa scelta consente di disporre di un indicatore sintetico e comparabile, in grado di rappresentare la capacità dei modelli di fornire risposte corrette in un ampio ventaglio di discipline.

La seconda variabile è la sostenibilità ambientale, già calcolata in precedenza come indice composito basato su consumo energetico, emissioni di CO_2 e utilizzo idrico. Tale indicatore, riportato su scala normalizzata e trasformato in valori di sostenibilità, esprime in modo unitario l'impatto ecologico dell'inferenza.

La terza variabile è la rapidità di inferenza, che rappresenta l'efficienza temporale con cui un modello produce i risultati a partire da un prompt. Essa è stata misurata attraverso la **latenza media per token**, ossia il tempo medio necessario per generare un singolo token di output. Valori più bassi corrispondono a una maggiore reattività del modello e, quindi, a migliori prestazioni dal punto di vista operativo.

$$LatencyPerToken_i = \frac{Latency\_ms}{Output\_tokens}$$

Sebbene la rapidità possa essere espressa anche in termini di throughput (numero di token generati al secondo), questa grandezza risulta matematicamente inversa alla latenza.

$$Throughput_i = \frac{1}{LatencyPerToken_i}$$

Utilizzare entrambe avrebbe dunque introdotto una ridondanza informativa, con il rischio di sovrappesare la dimensione temporale all'interno dell'indice composito. Per tale motivo, si è deciso di adottare come unica variabile rappresentativa la latenza per token, opportunamente trasformata mediante normalizzazione min–max.

La formula utilizzata è la seguente:

$$LatencyNorm_i = \frac{\max(LatencyPerToken) - LatencyPerToken_i}{\max(LatencyPerToken) - \min(LatencyPerToken)}$$

In questo modo, il modello con la latenza più bassa assume valore 1 (massima rapidità), mentre quello con la latenza più alta assume valore 0 (minima rapidità). Gli altri modelli si distribuiscono proporzionalmente all'interno di questo intervallo. Il risultato finale è un indicatore di rapidità espresso su scala 0–1, facilmente confrontabile con le altre variabili già normalizzate (accuratezza e sostenibilità), e pronto per essere integrato nell'indicatore composito esteso.

Modello	Latenza media per token (ms)	Latenza normalizzata
gpt-5	17.40474	0.8477980750267755
gpt-5-mini	33.53287040714841	0.4813468362039831
gpt-4.1	18.300604444444442	0.827442917473086
gpt-4.1-mini	29.017542000000002	0.5839407294030297
gpt-4.1-nano	14.036710444444443	0.9243239073376349
gpt-4o	39.17087741972151	0.3532442848175388
o3	14.23326049648712	0.9198580450062019
gpt-3.5-turbo	15.602745649024653	0.8887416342364597
claude-3-7-sonnet-20250219	30.407853818181817	0.5523511108135818
claude-sonnet-4-20250514	35.64122289922481	0.43344243792021936
claude-opus-4-20250514	51.41112462626263	0.07513060151385316
claude-opus-4-1-20250805	48.405690222222226	0.14341781973184947
deepseek-reasoner	54.7177477466415	0.0
meta-llama/Llama-3.2-90B-Vision-Instruct-Turbo	42.6593376137779	0.2739821181916189
open-mistral-7b	10.706079348523152	1.0
open-mixtral-8x7b	27.27035924433402	0.6236389008029732

Prompt Brevi

Modello	Latenza/token (ms)	LatenzaNorm
gpt-5	13,81	1,000
open-mistral-7b	14,30	0,994
o3	15,71	0,979
gpt-3.5-turbo	28,55	0,835
gpt-4.1-nano	29,72	0,822
open-mixtral-8x7b	29,90	0,820
gpt-4.1	34,24	0,771
gpt-4.1-mini	36,24	0,748
gpt-4o	39,56	0,711

Modello	Latenza/token (ms)	LatenzaNorm
claude-3-7-sonnet (2025)	57,01	0,516
LLaMA-3.2-90B Vision Instruct Turbo	58,11	0,503
claude-sonnet-4 (2025)	58,16	0,503
deepseek-reasoner	62,18	0,458
gpt-5-mini	70,51	0,364
claude-opus-4.1 (2025)	75,02	0,314
claude-opus-4 (2025)	102,99	0,000

Prompt Medi

Modello	Latenza/token (ms)	LatenzaNorm
gpt-4.1-nano	5,87	1,000
gpt-3.5-turbo	7,66	0,967
gpt-4.1	8,58	0,950
open-mistral-7b	8,63	0,949
o3	12,67	0,875
gpt-5-mini	15,05	0,832
gpt-5	16,20	0,811
claude-3-7-sonnet (2025)	16,37	0,808
claude-sonnet-4 (2025)	22,14	0,702
open-mixtral-8x7b	24,22	0,664
claude-opus-4 (2025)	25,41	0,642
gpt-4.1-mini	29,38	0,569
LLaMA-3.2-90B Vision Instruct Turbo	34,30	0,479
claude-opus-4.1 (2025)	34,66	0,473
deepseek-reasoner	52,39	0,148
gpt-4o	60,47	0,000

Prompt Lunghi

Modello	Latenza/token (ms)	LatenzaNorm
gpt-4.1-nano	6,52	1,000
open-mistral-7b	9,18	0,938
gpt-3.5-turbo	10,59	0,905
gpt-4.1	12,08	0,871

Modello	Latenza/token (ms)	LatenzaNorm
o3	14,31	0,819
gpt-5-mini	15,04	0,802
gpt-4o	17,48	0,745
claude-3-7-sonnet (2025)	17,84	0,737
gpt-4.1-mini	21,44	0,654
gpt-5	22,21	0,636
claude-opus-4 (2025)	25,84	0,551
claude-sonnet-4 (2025)	26,62	0,533
open-mixtral-8x7b	27,69	0,508
claude-opus-4.1 (2025)	35,54	0,326
LLaMA-3.2-90B Vision Instruct Turbo	35,57	0,325
deepseek-reasoner	49,58	0,000

Per rappresentare in maniera più realistica le prestazioni complessive dei modelli linguistici, è stato sviluppato un **indicatore composito esteso** che integra tre dimensioni fondamentali:

- **Accuratezza cognitiva (*MMLU*)** misura la capacità del modello di fornire risposte corrette su un ampio spettro di discipline, normalizzata su scala 0–1.
- **Sostenibilità (*S*)** rappresenta l'efficienza ambientale del modello, ottenuta come trasformazione inversa dell'indice ambientale sintetico (1 – indice), in modo che valori elevati indichino maggiore sostenibilità.
- **Rapidità (**R**):** esprime la velocità di generazione dei token, derivata dalla normalizzazione min–max della latenza media per token, invertita affinché valori alti corrispondano a modelli più rapidi.

Pesi attribuiti

L'attribuzione dei pesi alle tre dimensioni considerate riflette un approccio orientato all'uso pratico dei modelli linguistici. In particolare, all'accuratezza cognitiva è stato assegnato un peso pari a 0,5, poiché rappresenta la condizione necessaria per garantire la qualità e l'affidabilità delle risposte. La sostenibilità ambientale riceve un peso di 0,3, riconoscendone il ruolo crescente nei processi di valutazione delle tecnologie digitali, ma senza che questa dimensione prevalga sull'accuratezza. Infine, alla rapidità operativa è stato attribuito un peso pari a 0,2: si tratta infatti di un aspetto rilevante per l'esperienza d'uso e l'efficienza applicativa, pur avendo un impatto meno determinante sul valore complessivo del modello.

4.5.2 Formula dell'indicatore

L'indicatore composito esteso per il modello *i* è calcolato come:

$$IC_i^{esteso} = 0.5 \cdot MMLU_i' + 0.3 \cdot S_i + 0.2 \cdot R_i$$

dove MMLU indica l'accuratezza normalizzata del S_i rappresenta la sostenibilità ambientale derivata dall'inversione dell'indice d'impatto, e R_i corrisponde al punteggio di rapidità calcolato sulla base della latenza media per token.

Modello	MMLU′	Rapidità	Sostenibilità EU	Sostenibilità GL
GPT-5	0,865	0,848	0,819	0,819
GPT-5-mini	0,865	0,481	0,974	0,970
GPT-4.1	0,979	0,827	0,810	0,810
GPT-4.1 mini	0,883	0,584	0,983	0,983
GPT-4.1 nano	0,621	0,924	1,000	1,000
GPT-4o	0,819	0,353	0,586	0,522
O3	0,816	0,920	0,847	0,847
GPT-3.5-turbo	0,262	0,889	0,993	0,993
Claude-3.7 Sonnet (2025)	0,717	0,552	0,979	0,979
Claude-Sonnet-4 (2025)	0,848	0,433	0,975	0,975
Claude-Opus-4 (2025)	0,928	0,075	0,389	0,323
Claude-Opus-4.1 (2025)	0,894	0,143	0,425	0,363
DeepSeek-Reasoner (R1)	1,000	0,000	0,000	0,000
LLaMA-3.2-90B Vision Instruct Turbo	0,830	0,274	0,549	0,549
Mistral-7B	0,000	1,000	0,996	0,996
Mixtral-8×7B	0,309	0,624	0,813	0,813

Rank	Modello	Indicatore Esteso EU	Indicatore Esteso GL
1	GPT-4.1	0,898	0,898
2	GPT-4.1 mini	0,853	0,853
3	GPT-5	0,848	0,848
4	O3	0,846	0,846
5	GPT-5-mini	0,820	0,820
6	GPT-4.1 nano	0,815	0,815
7	Claude-Sonnet-4 (2025)	0,793	0,793
8	Claude-3.7 Sonnet (2025)	0,785	0,785
9	Mixtral-8×7B	0,648	0,648
10	LLaMA-3.2-90B Vision Instruct Turbo	0,640	0,640
11	GPT-4o	0,643	0,619
12	GPT-3.5-turbo	0,662	0,662
13	Claude-Opus-4.1 (2025)	0,615	0,591
14	Claude-Opus-4 (2025)	0,591	0,575
15	Mistral-7B	0,598	0,598
16	DeepSeek-Reasoner (R1)	0,500	0,500

L’indicatore esteso sviluppato in questo lavoro ha consentito di evidenziare con chiarezza quali modelli riescano a bilanciare al meglio accuratezza cognitiva, sostenibilità ambientale e rapidità operativa. In cima alla graduatoria si collocano **GPT-4.1 e GPT-4.1 mini**, che con valori superiori a 0,85 nell’indicatore esteso risultano i modelli più competitivi nel complesso, confermando la capacità delle varianti “main” e “mini” di coniugare prestazioni elevate con un impatto ambientale relativamente contenuto. **GPT-5** e **O3** seguono a brevissima distanza, mostrando anch’essi un equilibrio tra le tre dimensioni considerate.

Nella fascia intermedia compaiono modelli come **Claude-Sonnet-4 e Claude-3.7 Sonnet**, che raggiungono punteggi prossimi a 0,79: si tratta di sistemi solidi sul piano cognitivo, ma meno performanti sotto il profilo ambientale rispetto ai GPT. Ancora più penalizzate risultano le varianti **Claude-Opus-4 e Claude-Opus-4.1**, che pur superando 0,92 nei valori di accuratezza normalizzata, faticano a superare 0,60 nell’indicatore composito a causa di consumi ambientali elevati.

Al polo opposto della classifica si colloca **DeepSeek-Reasoner,** che rappresenta un caso emblematico: con un’accuratezza normalizzata pari a 1,000 è il modello più potente dal punto di vista cognitivo, ma la sostenibilità nulla (0,000) lo relega in fondo alla graduatoria, con un indicatore complessivo fermo a 0,50.

Situazione inversa per modelli come **Mistral-7B e GPT-4.1 nano,** che mostrano altissimi valori di sostenibilità (0,996 e 1,000 rispettivamente) ma non riescono a scalare la classifica per via di un'accuratezza troppo bassa.

4.5.3 Conclusioni

In un contesto in cui la sostenibilità delle tecnologie digitali è sempre più al centro del dibattito, disporre di una classifica che integra l'impatto ambientale consente di effettuare scelte più consapevoli e responsabili.

L'indicatore sviluppato in questo lavoro non ha soltanto un valore tecnico, ma assume una rilevanza strategica per tutti gli stakeholder coinvolti nell'adozione e nell'uso dei modelli linguistici di grandi dimensioni. In un contesto economico caratterizzato dalla crescente diffusione dell'intelligenza artificiale nei processi organizzativi, la disponibilità di una misura sintetica che integri accuratezza cognitiva, rapidità e sostenibilità ambientale rappresenta uno strumento essenziale per il management. Le imprese, in particolare, sono chiamate a valutare non solo le performance operative dei modelli, ma anche il loro impatto in termini di reputazione e di allineamento con le politiche di sostenibilità: la scelta di un modello energivoro può generare implicazioni negative sul piano dell'immagine aziendale e sul rapporto con stakeholder sempre più attenti alle tematiche ESG.

In questo quadro, l'indicatore composito consente di orientare le decisioni in modo consapevole, permettendo a imprese, pubbliche amministrazioni e organizzazioni della società civile di selezionare modelli che massimizzino il valore economico senza trascurare la responsabilità sociale e ambientale. L'adozione di un LLM non costituisce infatti una decisione puramente tecnologica, ma implica scelte strategiche che toccano la competitività, il posizionamento sul mercato e la coerenza con le aspettative degli stakeholder interni ed esterni. L'utilizzo di una classifica integrata, come quella proposta, riduce l'asimmetria informativa e fornisce un benchmark utile per garantire che le decisioni di adozione siano coerenti con gli obiettivi di lungo periodo e con le pressioni normative emergenti.

4.6 Analisi di Pareto applicata agli LLM

L'analisi di Pareto nasce in economia come strumento per individuare situazioni di efficienza, ossia configurazioni in cui non è possibile migliorare una dimensione senza peggiorarne un'altra (Pareto, *Manuale di economia politica*, 1906). Nella microeconomia classica questo concetto si applica ai panieri di consumo o alle curve di produzione; in un contesto tecnologico, la stessa logica può essere tradotta per valutare i compromessi tra prestazioni e impatti ambientali.

L'obiettivo non è identificare il "miglior modello assoluto", poiché tale opzione non esiste: un sistema può eccellere in una dimensione ma risultare carente in un'altra. La frontiera di Pareto consente dunque di individuare l'insieme di modelli che risultano efficienti, ossia quelli che non possono essere migliorati in accuratezza o in sostenibilità senza che vi sia un peggioramento nell'altra dimensione. Al tempo stesso permette di escludere i modelli dominati, per i quali esiste almeno un'alternativa che li supera in entrambe le metriche, e di restituire la logica dei compromessi, mostrando che ogni modello presente sulla frontiera rappresenta un equilibrio possibile, la cui scelta dipende dalle priorità strategiche delle imprese.

Alla base dell'analisi di Pareto vi è il concetto di **dominanza**. Si definisce che un modello A domina un modello B quando A presenta valori maggiori o uguali in tutte le dimensioni considerate e un valore strettamente superiore in almeno una. In questo caso, il modello B è razionalmente scartabile, poiché esiste un'alternativa (A) che lo supera in modo netto. Al contrario, i modelli che non risultano dominati da nessun altro compongono la cosiddetta **frontiera di Pareto**, ossia l'insieme delle scelte efficienti (Pareto, Vilfredo, *Manuale di economia politica*, 1906).

Nei modelli di Pareto, in generale, si parte da un insieme di alternative, ciascuna caratterizzata da più variabili. Attraverso la verifica delle relazioni di dominanza, si riduce lo spazio decisionale escludendo le opzioni inefficienti e mantenendo solo quelle efficienti. La frontiera così ottenuta non fornisce un "miglior modello assoluto", ma individua i compromessi non eliminabili, cioè quelli che rimangono validi a seconda delle preferenze e delle priorità di chi decide (Mas-Colell, Andreu, Whinston, Michael, Green, Jerry, *Microeconomic Theory*, 1995).

Nel nostro caso specifico, applicato ai modelli generativi, abbiamo deciso di concentrarci su due dimensioni: accuratezza e sostenibilità. La scelta è motivata dal fatto che queste due variabili rappresentano il trade-off più rilevante per imprese e stakeholder: da un lato la performance cognitiva del modello, dall'altro il suo impatto ambientale. Seguendo la regola della dominanza, sono stati esclusi tutti i modelli che risultavano superati contemporaneamente in entrambe le dimensioni, mantenendo sulla frontiera soltanto quelli non dominati.

Ogni modello viene rappresentato come un punto nello spazio bidimensionale con coordinate (Acc_i, Sus_i) (dove Acc_i indica il livello di accuratezza e Sus_i il livello di sostenibilità del modello i.

Un modello A domina un modello B se e solo se valgono le seguenti condizioni:

$$Acc_A \geq Acc_B \quad \wedge \quad Sus_A \geq Sus_B$$

e almeno una delle due disuguaglianze è stretta, ossia:

$$(Acc_A > Acc_B) \vee (Sus_A > Sus_B)$$

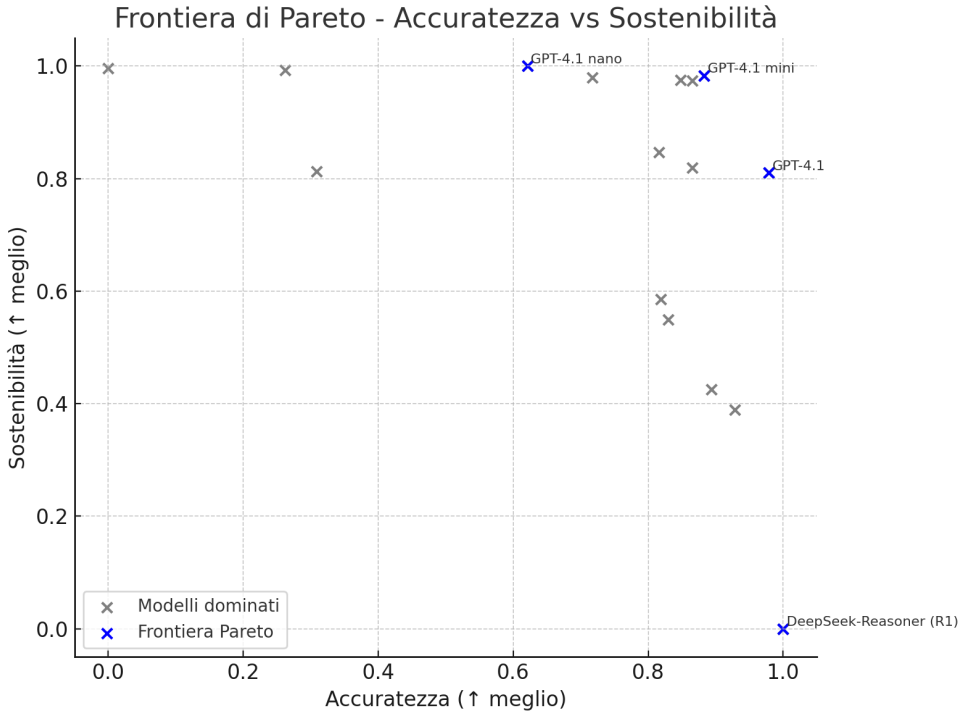
In tale circostanza, il modello BBB è considerato dominato e pertanto non può appartenere alla frontiera efficiente.

La f**rontiera di Pareto** è allora definita come l'insieme dei modelli non dominati da nessun altro elemento dell'universo considerato, cioè:

$$\mathcal{F} = \left\{ i \in M \;\middle|\; \nexists j \in M : \left(Acc_j \geq Acc_i \wedge Sus_j \geq Sus_i \wedge \backslash big\left(Acc_j > Acc_i \vee Sus_j > Sus_i \right) \right) \right\}$$

dove M rappresenta l'insieme di tutti i modelli in analisi, F la corrispondente frontiera efficiente e $i, j \in M$ sono due modelli generici, dove i è quello che stiamo valutando e j è un potenziale dominatore.

In altre parole, un modello i appartiene alla frontiera se non esiste nessun altro modello j che lo superi simultaneamente in accuratezza e in sostenibilità, con un miglioramento stretto almeno in una delle due dimensioni. Se tale modello j esiste, allora i è dominato e non può far parte della frontiera.



Il risultato dell'analisi ha individuato quattro modelli sulla frontiera: GPT-4.1, GPT-4.1 mini, GPT-4.1 nano e DeepSeek-Reasoner R1. Essi rappresentano i compromessi efficienti: da un lato DeepSeek-R1, che massimizza l'accuratezza sacrificando completamente la sostenibilità, dall'altro GPT-4.1 nano, che massimizza la sostenibilità a fronte di un livello di accuratezza più basso. In mezzo si collocano GPT-4.1 e GPT-4.1 mini, che offrono compromessi bilanciati. Tutti gli altri modelli sono stati esclusi dalla frontiera perché dominati, ossia superati in modo simultaneo da alternative più efficienti.

L'analisi ha mostrato che la frontiera è composta da un numero limitato di modelli, a conferma del fatto che non esiste una soluzione universalmente ottimale, bensì un insieme di compromessi che riflettono priorità differenti. Tale risultato offre alle imprese un quadro di riferimento chiaro all'interno del quale definire le proprie strategie: la scelta potrà orientarsi verso modelli che privilegiano la sostenibilità, quando la reputazione ambientale e gli obiettivi ESG assumono un ruolo competitivo, oppure verso modelli che massimizzano l'accuratezza, nei casi in cui la qualità cognitiva rappresenti un fattore critico di successo. In alternativa, la decisione potrà ricadere su compromessi intermedi, qualora l'obiettivo sia trovare un equilibrio tra prestazioni e responsabilità ambientale. In questa prospettiva, la frontiera di Pareto non costituisce soltanto uno strumento tecnico di analisi, ma si configura come un dispositivo interpretativo utile a guidare scelte manageriali coerenti con le attuali sfide di innovazione e sostenibilità.