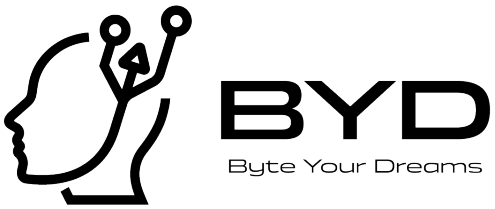

byteyourdreams.swe@gmail.com

Analisi dell'impatto green delle GenAI

Informazioni documento

Redattore	A.M. Margarit
Verificatore	A. Mio
Amministratore	O.F. Stiglet
Destinatari	Vimar S.p.A.

Registro delle modifiche

Versione	Data	Autore	Verificatore	Dettaglio
0.1.0	28/04/2025	A.M Margarit	A.Mio	Redazione documento

Indice

Byte Your Dreams

May 1, 2025

Contents

1 Confronto dell'Impatto Ambientale tra DeepSeek e altri Large Language Models

Deepseek emerge con la promessa di una maggiore sostenibilità energetica.
Tuttavia, il consumo energetico dei datacenter continua a crescere, sollevando dubbi su quanto sia realmente sostenibile rispetto ad altri modelli di AI generativa presenti sul mercato.

I modelli di AI generativa vivono due fasi distinte con impatto ambientale diverso: la **fase di training** e la **fase di inferenza**. La prima ha un significato ovvio: la costruzione di un modello richiede un costo computazionale e quindi ambientale importante. Su questa componente dell'impatto energetico, in effetti, Deepseek sembra molto efficiente. Si è parlato di 5,6 milioni di dollari, anche se più di qualche dubbio rimane. La seconda fase che ha un impatto energetico, quella di inferenza, è ancora più interessante. Infatti, mentre il training è un costo una tantum, l'inferenza è la fase di utilizzo del modello (quando diamo in pasto all'AI un "prompt"), quindi un costo ricorrente.

Negli ultimi anni, l'attenzione verso l'impatto ambientale delle tecnologie avanzate è cresciuta notevolmente. Tra queste tecnologie, gli LLM come DeepSeek, GPT-4, Claude, Gemini, Llama e Mistral sono al centro del dibattito. Viene mostrato di seguito come si confrontano in termini di consumo energetico, emissioni di CO2 e uso dell'acqua.

1.1 Consumo Energetico

I dati che abbiamo sulle AI generative attuali sono un po' preoccupanti. Una query di ChatGPT consuma circa **10 volte più energia** di una ricerca su Google (2,9 wattora contro 0,3 wattora). Questo enorme consumo eneretico deriva dai server che eseguono calcoli complessi per processare enormi quantità di dati. Ogni volta che utilizziamo un'AI per generare un testo o un'immagine, i server eseguono operazioni matematiche che richiedono energia elettrica. L'**International Energy Agency** stima che i datacenter ad oggi esistenti (più di 8000) utilizzano dall'**1% al 2%** dell'energia elettrica globale.

I modelli DeepSeek sono noti per la loro efficienza energetica. I server che supportano questi modelli consumano dal **50% al 75%** meno energia rispetto alle ultime unità GPU di Nvidia. Questo rappresenta un notevole passo avanti verso la riduzione dell'impatto ambientale delle tecnologie AI.
D'altra parte l'addestramento di modelli di grandi dimensioni come GPT-4 richiede risorse enormi. Si stima che l'addestramento di GPT-4 possa richiedere **fino a 25.000 GPU in funzione** per mesi, consumando quantità immense di elettricità. Claude utilizza circa 20.000 GPU, Gemini 22.000, Llama 18.000 e infine Mistral circa 19.000 GPU.

1.2 Emissioni di CO2

La partita vera della sostenibilità si gioca non tanto sui costi per costruire la macchina, ma sulle emissioni di CO2 una volta che la macchina viene messa in funzione. Quindi, per fare una valutazione corretta, andrebbero considerate più variabili: il costo della fase di training,in costo della fase di inferenza, il numero di interrogazioni medie per avere una risposta soddisfacente e l'impatto energetico delle diverse modalità di utilizzo (in cloud, in locale...).
Sebbene i **dati esatti** sulle emissioni di CO2 per i modelli DeepSeek **non siano facilmente disponibili**, la loro efficienza energetica suggerisce una minore impronta di carbonio rispetto ad altri modelli.
In confronto, l'addestramento di un singolo grande LLM come GPT-4 può produrre un ammontare stimato di **300 tonnellate di CO2**. Claude ne produce 250, Gemini circa 270 e Llama 220 tonnellate.

1.3 Utilizzo dell'acqua

Oltre al consumo energetico, l'AI ha anche un impatto significativo sul consumo di acqua. I server che alimentano i modelli AI generano enormi quantità di calore e devono essere raffreddati continuamente per funzionare in modo efficiente. Molti datacenter utilizzano torri di raffreddamento che richiedono grandi quantità di acqua per dissipare calore. L'acqua utilizzata viene spesso **riciclata dalle tre alle 10 volte** prima di essere scaricata, ma una parte significativa si perde per evaporazione.
Secondo uno studio di Washington Post e dell'Università della California Riverside, l'uso di ChatGPT per inviare una e-mail di 100 parole potrebbe richiedere fino a mezzo litro di acqua. Se un americano su 10 inviasse una e-mail settimanale tramite ChatGPT, i server AI consumerebbero **435 milioni di litri di acqua all'anno**, pari al fabbisogno idrico giornaliero di circa un milione di persone. L'uso dell'acqua è un fattore spesso trascurato ma cruciale nella produzione di chip AI. La **produzione di chip** utilizzati dai modelli di DeepSeek richiede oltre **8.300 litri di acqua per chip**.

1.4 Conclusione

In conclusione, mentre i modelli DeepSeek offrono vantaggi significativi in termini di efficienza energetica e potenzialmente minori emissioni di CO2, è chiaro che l'intero settore deve affrontare sfide importanti per diventare più sostenibile.

2 Riferimenti

- https://www.dw.com/en/what-does-chinas-deepseek-mean-for-ais-energy-and-water-use/a-71459557
- https://www.cutter.com/article/environmental-impact-large-language-models
- https://deteapot.com/chatgpts-carbon-footprint-how-much-energy-does-your-ai-prompt-really-use
- https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117
- https://www.nature.com/articles/s41598-024-76682-6.pdf
- https://www.rinnovabili.net/business/markets/deepseeks-energy-consumption-ais-75-power-cut/