RATIO QUOTIDIANO
Il commento e la critica delle più importanti notizie del momento

CONSULENZA AZIENDALE, COMMERCIALE E MARKETING

di CRISTIAN ZULIANI

Meno costa il token, più sale la bolletta. Paradosso di Jevons nell'AI

Il prezzo unitario dei token crolla, ma i workflow agentici moltiplicano le chiamate: come l'efficienza dell'AI rischia di far esplodere la spesa aziendale se gestita solo a consumo.

Un dirigente Nvidia lo ha detto ad Axios lo scorso aprile, senza troppi giri di parole: *"per il mio team, il costo del calcolo supera di gran lunga il costo dei dipendenti".* La frase di Bryan Catanzaro, vicepresidente per il deep learning applicato, pesa doppio se si considera che arriva da chi vende l'hardware su cui gira tutta l'intelligenza artificiale del pianeta. Nello stesso articolo si racconta che il direttore tecnico di Uber ha esaurito l'intero budget AI stanziato per il 2026 per via dei costi di inferenza legati ai token, non un progetto pilota andato storto, ma il consumo ordinario di uno strumento già in produzione.

Il paradosso è che tutto questo accade mentre il prezzo di un token continua a scendere. Secondo lo Stanford AI Index Report 2025, tra novembre 2022 e ottobre 2024 il costo di elaborazione di un milione di token, per un modello con prestazioni equivalenti a GPT-3.5, è passato da 20 dollari a 7 centesimi: oltre 280 volte più economico in meno di 2 anni. Eppure Gartner prevede che la spesa mondiale in intelligenza artificiale salga a 2.590 miliardi di dollari nel 2026, il 47% in più rispetto al 2025. E secondo una rilevazione di Futurum Group su oltre 1.600 responsabili tecnologici, quasi metà delle aziende, il 46,9%, sta spendendo in AI più di quanto avesse preventivato.

Il meccanismo ha un nome che risale all'Ottocento: paradosso di Jevons, dall'economista che per primo osservò come un carbone reso più efficiente dai motori a vapore non riducesse i consumi ma li facesse esplodere, perché l'efficienza lo rende conveniente anche dove prima non lo era. Satya Nadella lo richiamò esplicitamente nel gennaio 2025, commentando l'arrivo di DeepSeek: *"il paradosso di Jevons colpisce ancora, l'AI più efficiente ed economica ne farà esplodere l'uso".* Nei token il fenomeno si traduce così: un compito che un anno fa risolveva un modello con una risposta diretta oggi passa attraverso un agente che ragiona per passaggi successivi, si autoverifica e richiama altri strumenti prima di restituire un output. **Il prezzo unitario scende, ma il numero di token necessari per lo stesso lavoro a volte sale più in fretta.**

Per uno studio o un'azienda che valuta l'automazione documentale o un assistente AI per i clienti, la cosa da tenere a mente è che il prezzo di listino di oggi dice poco sulla spesa reale una volta che lo strumento entra nei processi. Il rischio riguarda soprattutto chi paga a consumo - API come Azure OpenAI o AWS Bedrock, agenti costruiti su misura - mentre le licenze a canone fisso scaricano quel rischio sul fornitore, che infatti impone limiti di utilizzo o soglie di fair use per proteggersi. Un flusso che legge un documento, lo confronta con altri e propone una bozza può consumare in un solo passaggio quanto bastava prima per decine di richieste separate. È la differenza tra guardare il prezzo al litro e guardare la bolletta a fine mese.