

ECCO PERCHE' L'INTELLIGENZA ARTIFICIALE E' UN "CASTELLO DI CARTE"

Questi pensieri sono tratti da una analisi di ZoomyZoomm (grazie Carlo per la segnalazione) che ritengo estremamente interessanti e assolutamente da condividere (se riesci leggi fino alla fine)

Perché l'intelligenza artificiale è un castello di carte:

1. Paghi 200 dollari all'anno per un'app di intelligenza artificiale (come Cursor).
2. Cursor paga a OpenAI 500 \$ per i token API (di cui 300 \$ sono finanziamenti VC).
3. OpenAI paga ad AWS 1000 dollari per l'elaborazione (di cui 500 dollari sono finanziamenti VC).
4. AWS paga 10.000 \$ per le GPU \$Nvda .

Vedi il problema?

A meno che tu, come utente, non sia miracolosamente disposto a pagare 1.000 \$ per l'app AI

L'unica cosa che sostiene l'intelligenza artificiale è il finanziamento del capitale di rischio

Nessun finanziamento VC (Venture Capital):

Il livello applicativo dell'intelligenza artificiale non è redditizio

Il livello LLM (Large Language Model) non è redditizio

Il livello di elaborazione non è redditizio

Lo strato GPU non è sostenibile

I tori dicono che il costo dell'inferenza sta diminuendo esponenzialmente (e lo è)

Ciò significherebbe che anche il costo di elaborazione per LLM come OpenAI (e per estensione per le app di intelligenza artificiale come Cursor) diminuirebbe.

Ma questo deve accadere PRIMA che finiscano i finanziamenti VC

E dipende da quanto pagheranno gli utenti

Se gli utenti finali non sono in grado di pagare il prezzo di pareggio per l'intero stack:

Per esempio:

costo della GPU Nvda ammortizzato nel corso del ciclo di vita

Margini del cloud iperscalabile

Margini di inferenza LLM (Large Language Model)

Margini wrapper AI

Che potrebbe letteralmente arrivare a 1.000 \$ per utente

Poi tutto questo gioco implode

Al momento, le persone pagano \$ 20 al mese per il livello medio e \$ 200 al mese per i piani di livello professionale, ad esempio Cursor

Sono \$ 240 all'anno // \$ 2400 all'anno

Immagino che l'utente che utilizza molto probabilmente brucia più di 10.000 \$ di elaborazione

Quindi, a meno che tu non possa aumentare di 10 volte il costo per queste persone

Questo non è sostenibile

Vedo solo 3 percorsi praticabili per queste aziende:

1. Gli LLM devono ridurre drasticamente i costi dei modelli/inferenza (limitando l'utilizzo ai wrapper di intelligenza artificiale)
2. Le aziende che applicano l'intelligenza artificiale devono iniziare a far pagare MOLTO di più in generale (difficile da fare ora perché è in gioco l'accaparramento di terreni)
3. I prezzi delle GPU diminuiscono sostanzialmente (ad esempio, emerge un concorrente di NVDA)

Il che allevierebbe MOLTI costi a valle, ad esempio:

1. Gli hyperscaler possono ridurre i prezzi
2. Gli LLM possono ridurre i prezzi delle API
3. Le aziende di applicazioni AI possono raggiungere il pareggio di prezzo con i prezzi attuali

In definitiva, il "cigno nero" che vedo è che la deflazione sta arrivando per ogni aspetto della tecnologia:

1. Costi delle GPU in calo (grazie all'emergere di concorrenti)
2. Costi del cloud in calo (grazie ai neocloud e alla concorrenza degli hyperscaler)
3. Costi di inferenza in calo (grazie a hardware migliore)

Tutto ciò significa che l'economia unitaria FINALE migliorerà

Ma QUANDO questo accade

E QUALE profilo di margine ha ciascuno di questi giocatori quando tutto è spazzato via

Deve ancora essere terminato

Se dovessi indovinare:

Vincitori: fornitori di GPU e Hyperscaler

Perdenti: LLM e aziende di app di intelligenza artificiale

Ecco perché vedi l'intera comunità HF andare long:

NVDA MSFT GOOG AMZN

Perché queste 4 aziende sono destinate a fare un sacco di soldi nel prossimo decennio

Indipendentemente da ciò che accade nell'IA

MA...

Odiando il consenso, cerco modi per fare lo stesso scambio al 50% della valutazione

Per me, questa è ***China Tech***

NVDA = ***SMICY***

MSFT = ***TCEHY***

GOOG = ***BIDU***

AMZN = ***BABA***

INOLTRE, non ho menzionato altri enormi colli di bottiglia come:

- Terre rare
- Produzione di energia
- Scalabilità dell'IA incorporata

La ***Cina*** è al comando in tutti e 3 i settori e ha:

- Tutti i minerali delle terre rare
- Tutta la lavorazione delle terre rare
- Energia pulita illimitata (nucleare)
- Capacità produttiva robotica del 99%

Quello che voglio dire è che è inevitabile che la Cina recuperi il terreno perduto sul fronte LLM (sia attraverso modelli chiusi come Bytedance o open source come Qwen DeepSeek ecc.)

Quindi la domanda successiva diventa: cosa succede DOPO che i modelli raggiungono il plateau?

Tutto si riduce agli atomi.

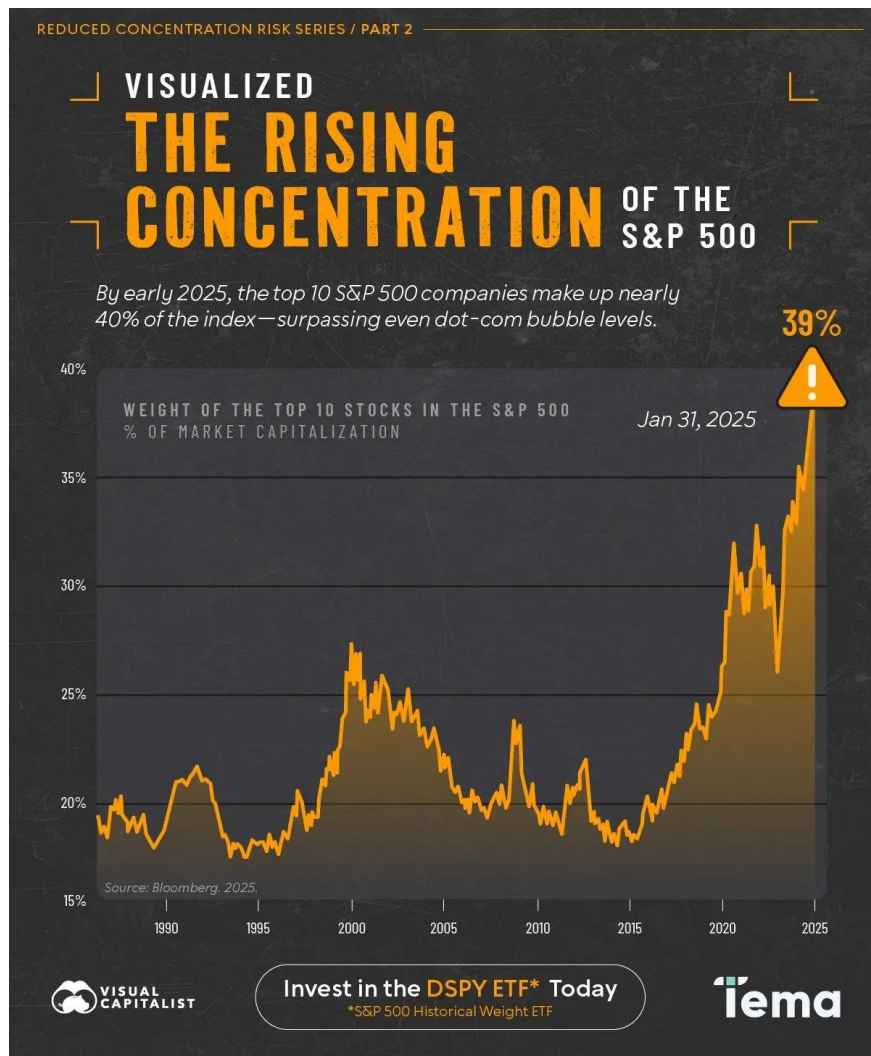
Tutto si riduce a una questione di scala.

QUI per voi su quello che potrebbe essere il primo tassello del domino a cadere in questa "casa di carte dell'IA".

🔗 Perché i mag 7 stanno per implodere:

1/ L'ondata di spese in conto capitale per l'intelligenza artificiale \$AMZN \$MSFT \$GOOG \$META dovuta all'aumento dei carichi di lavoro di inferenza dell'intelligenza artificiale di nuova generazione ha una conseguenza indesiderata significativa per la quale il mercato non è pronto.

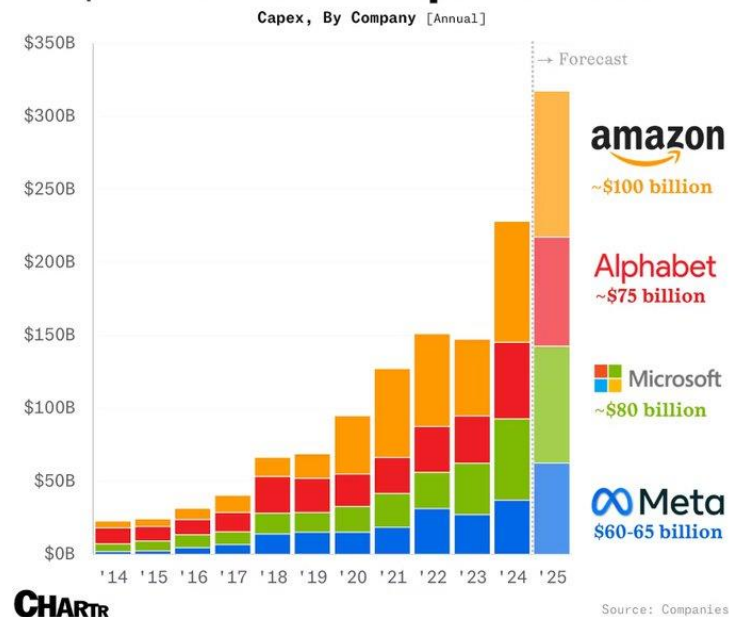
Il killer nascosto dell'IA è la PERDITA DI VALORE (causata dall'eccessivo utilizzo e dall'obsolescenza tecnologica).



2/ Nel 2025, la spesa in conto capitale per l'intelligenza artificiale raggiungerà i 300 miliardi di dollari > , convogliati verso data center e GPU per supportare i carichi di lavoro dell'intelligenza artificiale generativa.

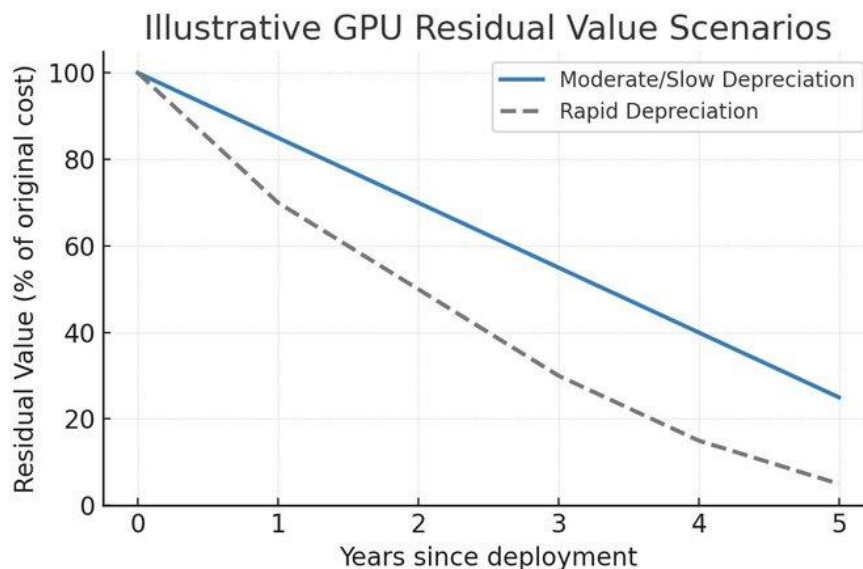
Questa è una decisione razionale data la richiesta di formazione e inferenza, ma trascura un problema critico: l'hardware viene utilizzato a velocità di gran lunga superiori alle specifiche di progettazione originali.

4 Big Tech Companies Plan On Spending \$315+ Billion On Capex This Year



3/ I tradizionali programmi di ammortamento per le GPU sono quinquennali, ma le intense richieste dell'IA di nuova generazione (operazioni continue ad alto carico) causano un'usura fisica molto più rapida (entro 2-3Y) .

Ad esempio NVDA il passaggio agli aggiornamenti annuali dell'architettura GPU rende gli investimenti esistenti obsoleti più rapidamente



4/ Le prove di questa discrepanza stanno già emergendo.

\$AMZN ha ridotto la durata dei server AI da 6Y a > 5Y, comportando costi per circa 920 milioni di dollari.

\$META la riduzione della durata di vita delle attività potrebbe aumentare le spese di ammortamento di oltre 5 miliardi di dollari nel 2026

Tutto ciò ha un impatto diretto sui profitti.

Company	Ticker	Nature of the change	Effect
Alphabet Inc.	GOOGL	Change in estimated useful life of servers from 4 to 6 years, and of networking equipment from 5 to 6 years	Decrease in depreciation expense of \$3.9 billion or increase in EPS of \$0.24 for the year ending December 31, 2023
AMAZON COM INC	AMZN	Change in estimated useful life of servers from 5 to 6 years, due to "continuous improvements in...hardware, software, and data center designs". Note that for Amazon this is the second change in 3 years - the company changed the depreciable life of servers from 4 to 5 years and of networking equipment from 5 to 6 years effective January 1, 2022	Decrease in depreciation expense of \$695 million or increase in EPS of \$0.07 for the three months ending on March 31, 2024. The 2022 change in estimate decreased depreciation expenses by \$3.6 billion or \$0.27 per share for the year ending December 31, 2022.
MICROSOFT CORP	MSFT	Due to increased efficiencies and advances in technology, the estimated useful lives of both server and network equipment increased from 4 to 6 years	An increase in operating income of \$3.7 billion and net income of \$3.0 billion, or \$0.40 per share for the year ending June 30, 2023
INTERNATIONAL BUSINESS MACHINES CORP	IBM	Increased the estimated useful life of servers and networking equipment from 5 to 6 years for new assets and from 3 to 4 years for used assets	Increase in income from continuing operations before income taxes of \$208 million or \$0.18 per share for the year ending December 31, 2023
Meta Platforms, Inc.	META	Increase in estimated useful lives of a majority of the servers and network assets from four years to 4.5 years, effective the second quarter of 2022, and further extension of the useful lives to 5 years effective the fourth quarter of 2022	Decrease in depreciation expenses of \$860 million or increase in EPS of \$0.26 for the year ending on December 31, 2022
ORACLE CORP	ORCL	Increased the estimated useful life of servers from 4 to 5 years	Decreased operating expenses by \$434 million for the year ending May 31, 2023

Source: Data sources - SEC filings & Audit Analytics CAE database. Analysis by Deep Quarry

5/ \$MSFT e \$META hanno esteso i periodi di ammortamento per aumentare artificialmente i guadagni a breve termine.

Ma prima o poi i nodi vengono al pettine.

Non puoi tenere 100 miliardi di dollari di "attività GPU" nel tuo bilancio che si sono sciolte e sono inutili. Prima o poi dovrai contabilizzare

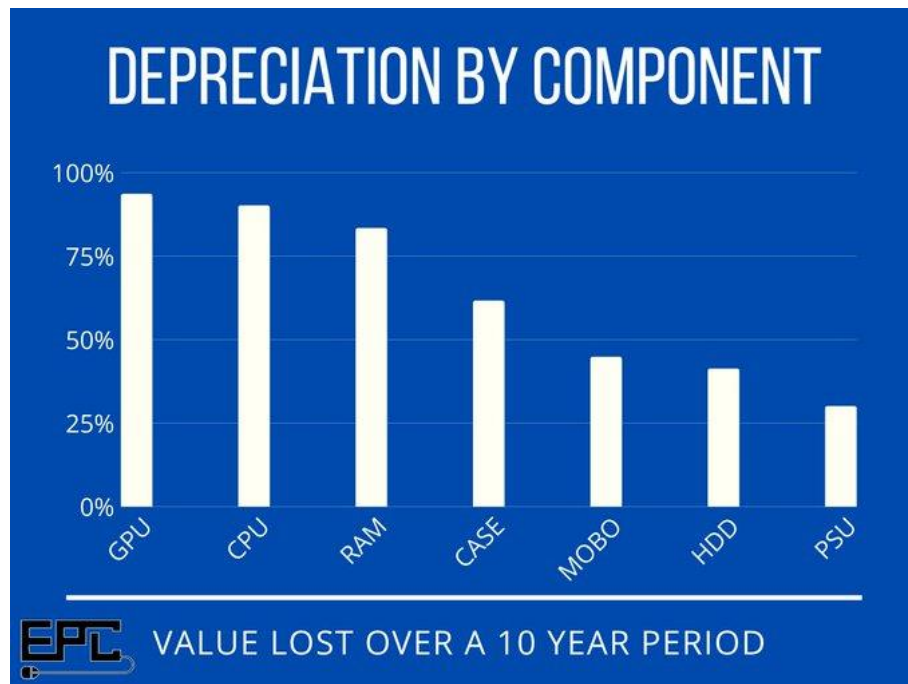
META PLATFORMS, INC.
CONDENSED CONSOLIDATED BALANCE SHEETS
(In millions)
(Unaudited)

	June 30, 2025	December 31, 2024
Assets		
Current assets:		
Cash and cash equivalents	\$ 12,005	\$ 43,889
Marketable securities	35,066	33,926
Accounts receivable, net	16,561	16,994
Prepaid expenses and other current assets	9,981	5,236
Total current assets	73,613	100,045
Non-marketable equity investments	21,988	6,070
Property and equipment, net	147,039	121,346
Operating lease right-of-use assets	15,662	14,922
Goodwill	20,654	20,654
Other assets	15,788	13,017
Total assets	\$ 294,744	\$ 276,054

6/ Non puoi avere entrambe le cose. Avere > un utilizzo del 100% sulle tue GPU non è privo di svantaggi.

L'intelligenza artificiale di generazione richiede molte risorse.

Si finisce per sottoporre l'hardware a sollecitazioni superiori ai limiti operativi standard, con conseguente esaurimento più rapido delle risorse, riduzione del valore di rivendita e maggiori svalutazioni.



7/ Gli oneri per svalutazione si presenterebbero come un impatto una tantum sulla redditività.

Non è la svalutazione in sé a preoccuparmi. È il segnale che invierà al mercato.

Ammortamento più rapido = minore utilizzo = minori ricavi = ROI più basso.

E se spendessero troppo per le GPU? Il mercato è preparato?

$$\text{Return on Invested Capital (ROIC)} = \frac{\text{Operating Income}_t (1 - \text{Tax Rate})}{\text{Book Value of Invested Capital}_{t-1}}$$

8/ Questo non significa che l'intelligenza artificiale sia una bolla enorme e che queste azioni crolleranno dell'80%.

Sono delle vacche da mungere con un debito netto negativo e hanno chiari casi d'uso nei settori verticali in cui operano.

L'intelligenza artificiale è reale.

MA il mercato si renderà presto conto della realtà. Niente sale per sempre.

9/ In breve:

- \$META \$AMZN \$MSFT \$GOOG ha puntato tutto sulla costruzione di investimenti in intelligenza artificiale

- Per ora, il mercato è soddisfatto di questo livello di spesa.

- Sebbene strategicamente abbia senso, l'ammortamento accelerato dell'hardware potrebbe forzare \$10b + svalutazioni

- Se ciò dovesse accadere, aspettatevi una svendita del -20%.

CONCLUSIONE

L'intelligenza Artificiale è presente e sarà sempre più presente nel panorama finanziario globale ma i nomi sui quali INVESTIRE non sono negli U.S.A. ma in CHINA

Ciò che stiamo osservando è il comportamento di trader che non sono abituati a periodi di rialzo sostenuti.

Ecco perché vendono non appena vedono un +5%.

Ma quando si renderanno conto che questo rally ha ancora gambe per andare +20/30/50/100% passeranno dalla ricerca del profitto all'AVIDITÀ e acquisteranno volumi crescenti e li manterranno per periodi più lunghi.

Il software è a rischio a causa dell'intelligenza artificiale.

L'e-commerce NON è a rischio a causa dell'intelligenza artificiale.

Il settore Fintech NON è a rischio a causa dell'intelligenza artificiale.

Gli Stati Uniti sono in una bolla di valutazione.

L'America Latina NON si trova in una bolla di valutazione.

La Cina NON si trova in una bolla di valutazione.

Il Sud-est asiatico NON si trova in una bolla di valutazione.

Investire è così facile, amico.

BUY CHINA TECH