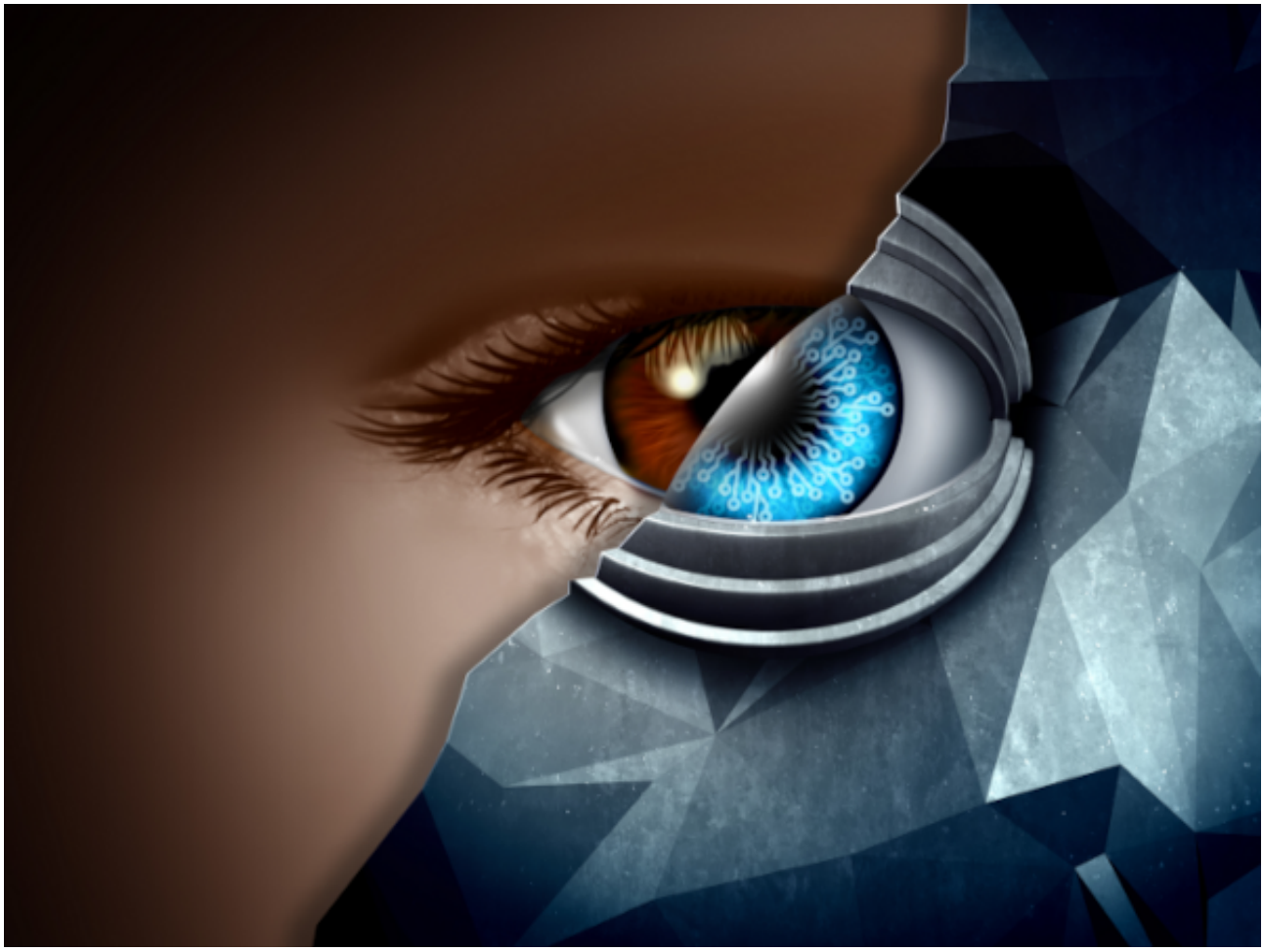
Cassandra Crossing 665- Token, è finita la pacchia!



image

(665) - Le aziende che vendono LLM stanno trattando i loro utenti come avviene sempre quando qualcuno raggiunge un monopolio di fatto. Nel modo peggiore possibile. E quel che è peggio è che ci sono costrette.

11 maggio 2026 - Nell'ultima puntata di "Archivismi", Cassandra aveva inserito alcune note riguardo la performance dell'account Claude che stava utilizzando.

E, finito quel lavoro, dato che aveva sempre "praticato" molto poco il settore degli LLM, ha fatto qualche altra indagine, volta a comprendere perché l'argomento del "cosa compro" e del "cosa uso" fosse sempre così bistrattato e vago nell'intero settore degli LLM.

Non si tratta di questioni così vitali come quella finanziaria o quella dell'inquinamento dell'Infosfera, ma è comunque un argomento interessante.

Non c'è voluto poi molto, è bastato mettere insieme i pezzi per avere il panorama completo; Cassandra ne ha quindi distillato un "Bignami", che si permette di offrirvi in queste poche righe.

Quando usate un LLM, lo potete fare in due modi, eseguendolo sul vostro computer oppure nel cloud, quindi sul computer di qualcun altro.

Della prima modalità forse non sapevate nulla, e poi certamente state usando la seconda, perciò, riservando all'uso degli LLM "in locale" una prossima esternazione di Cassandra, continuiamo.

Per andare al sodo, quando usate un LLM nel cloud, state consumando tre tipi di risorse:

- Token
- Contesto
- Tool

che tradotte in termini reali, stiracchiando solo un po' le cose, sono rispettivamente:

- Tempo di calcolo (tempo macchina)
- Memoria
- Esecuzione di programmi esterni

Quindi, quando consumate token state occupando una o più unità di elaborazione, soprattutto **GPU**, cosa che implica consumo di tempo macchina ed energia elettrica, nonché oneri finanziari del datacenter del quale state temporaneamente "occupando" una piccola parte.

Quando state consumando token nella vostra chat, lo fate con un dato "**contesto**" fisso, anche questo espresso in token (attualmente da 100.000 ad 1.000.000), che rappresenta lo spazio di memoria di lavoro; questo spazio viene realizzato (in termini degni di un uomo delle caverne) **usando RAM e SSD**; state quindi occupando, oltre che processori, anche risorse di questo tipo.

Quando l'LLM che state usando utilizza un tool, crea un "**agente**" "*usa e getta*", scritto di solito in Bash o Python; Claude, ad esempio, ve lo mostra con utilissima chiarezza, permettendo di vedere l'agente addirittura mentre viene scritto. In pratica il vostro agente consumerà memoria e tempo macchina, ma di tipo "classico", su un'altra piccola frazione di datacenter.

Perché fa questo? Perché il "contesto" è una mercanzia molto costosa, e far fare il lavoro ad un algoritmo deterministico come un script aiuta a consumarne di meno, ovvero ad avere risultati migliori a parità di contesto. Serve anche a generare output più lunghi prima che l'LLM cominci a sparare "stronzate" (che è un **termine tecnico** e condiviso, come ben spiegato in questo paper).

Nel ***Paese dei Balocchi*** che è sempre stato, fino a pochi mesi fa, il mondo degli LLM, tutti consumavano tutto quello che veniva messo a disposizione gratuitamente, e nessuno, nemmeno Altman ed il duo Amodei, si preoccupava dei costi, che infatti venivano allegramente pagati dalla speculazione finanziaria.

Ma da qualche tempo chi ha tirato fuori centinaia di miliardi, tipo Blackrock, ha cominciato a voler vedere dei conti che avessero almeno una parvenza di correttezza. E siccome nulla, mai, sta finanziariamente in piedi se non produce delle entrate, questo ha portato alla necessità di grossi cambiamenti.

Il mondo degli LLM, infatti, è l'unico caso nella storia in cui acquistare nuovi utenti aumenta le perdite in maniera più che lineare. E siccome in un piano di business gli utenti devono, prima o poi, rappresentare fonti di reddito, si doveva cambiare registro.

A parte provare timidamente ad introdurre pubblicità e sesso, OpenAI, Anthropic e compagnia bella si sono trovati costretti a far generare entrate crescenti al proprio "parco buoi".

Ci sono solo due modi per farlo; far pagare quello che era aggratis, e far pagare di più quello che veniva venduto a poco. Questo è quanto basta a spiegare quello che sta succedendo.

Per far pagare di più, basta aumentare il prezzo degli abbonamenti (e dell'utilizzo via API); ma questo non è sufficiente.

Infatti perché un utente possa, in prospettiva, diventare una fonte di reddito crescente in maniera più che lineare, fino ad arrivare ad essere una fonte di guadagno, bisogna evidentemente che non rimanga una fonte di perdite; quindi non solo lo si deve far pagare di più, ma anche e contemporaneamente dargli di meno. Come fare?

Già, ecco perché i dati di utilizzo e di soglia degli LLM sono sempre nascosti; perché questo permette (tra l'altro) di diminuirli senza farsene accorgere, almeno dalla grandissima maggioranza degli utenti.

In questo modo quello che si ottiene, a parità delle altre cose percepite dall'utente, è "solo" un peggioramento delle "prestazioni" del modello, che si traduce molto spesso nella necessità di passare ad un modello più grande, e quindi ad un account migliore ed ovviamente più costoso.

E' quindi possibile aggiungere nuovi tipi di account che permettano di alzare la barriera del prezzo prima oltre i 20 euro, poi oltre i 100, poi oltre i 200 … e via così!

Per voi è un dejà vu? Una scomodità?

Pensate cosa rappresenta allora per quelle aziende che hanno già incorporato processi guidati dagli LLM, che improvvisamente non funzionano più, malgrado non sia cambiato nulla, perché il loro modello ci "indovina di meno". A quali possibilità di "ricatto" sono ora esposte.

Per oggi poco altro rimane da dire, se non ricordare che usare i modelli linguistici serve solo quando si deve elaborare del linguaggio. Non come oracoli. Non per sostituire il pensiero di chi lavora.

Chi lo usa per altri scopi lo fa a suo danno (e se lo merita tutto), ma talvolta (purtroppo) anche a danno di tanti altri.

Scrivere a Cassandra - Mastodon - Buttondown.com
Videorubrica "Quattro chiacchiere con Cassandra
Lo Slog (Static Blog) di Cassandra
L'archivio di Cassandra: scuola, formazione e pensiero

Cassandra per i posteri: l'archivio su Internet Archive