La smaterializzazione del denaro è un processo che è progressivamente avanzato dalla nascita dell'era mercantile sino alla sua definitiva affermazione nell'era digitale, in tutte le sue tre determinazioni di unità di misura, mezzo di scambio e di tesaurizzazione. Il dispiegamento globale dei sistemi digitali in grado di innervare pervasivamente le relazioni sociali e produttive porta ad individuare nel flusso di informazioni il medium dei processi di creazione di valore, un nesso tra unità di informazione (bit, byte) l'unità di valore, ad esprimere il concetto di produttività dell'informazione, della creazione di valore nelle diverse fasi di produzione/circolazione dell'informazione.
L'innovazione fondamentale, il salto di qualità dall'inizio di questo secolo è stato l'affermarsi dei social networks come mediazione necessaria e capillare delle relazioni interpersonali e sociali, come mediazione dell'attività conversazionale a tutti i livelli, da cui in modo capillare e continuo si estrae valore non solo e non tanto per l'inserimento delle pubblicità quanto per la possibilità di profilare soggetti individuali e collettivi arrivando a creare mappe dinamiche in costante aggiornamento delle loro attitudini, propensioni, peraltro non in modo neutro in quanto vengono costantemente ricreati i contesti, aggregati, ecc.
La crescita esponenziale delle informazioni estratte dal veicolare il tessuto delle relazioni ai più diversi livelli ha spinto allo sviluppo dei dispositivi di memorizzazione ed elaborazione di questa molte di dati, definite col neologismo di 'big data'.
L'ulteriore salto di qualità è quello che stiamo vivendo e sperimentando in questi anni con una accelerazione che si misura nell'arco dei mesi, vale a dire le tecnologia dei modelli transformer, che conosciamo come Large Language Modules, che si sono affermati a partire dall'introduzione del meccanismo dell'attenzione entro i processi di elaborazione statistica dei testi e delle conversazioni che la rete ha messo e mette a disposizione ogni secondo che passa in un flusso che conosce una espansione più che esponenziale. Ricordiamo che il meccanismo dell'attenzione, introdotto in un paper dei ricercatori di Google dal titolo '*Attention is all you need*'[1]. In parole povere non solo si calcola la probabilità negli enunciati analizzati una parola ne segua una altra o meglio una sequenza di caratteri ne preceda un'altra, ma si analizza anche la rilevanza entro una frase dei termini presenti, si associa ad ogni termine un insieme di altri termini, che ne definiscano il contesto più probabile. Le unità di testo di cui si analizzano le ricorrenze e le modalità di aggregazione nella produzione linguistica sono denominati 'token'. Da quando è esploso l'utilizzo dei LLM con ChatGPT-3 con l'espandersi della platea dei modelli che si fanno concorrenza e la penetrazione del loro utilizzo in tutte le attività, il flusso dei token, il loro consumo sta diventando un parametro essenziale a definire e quantificare ogni sorta di attività. Il token entro i processi dell'A.I. man mano che questa va a intermediare ogni relazione, produzione e scambio, diventa l'unità di misura atomica non solo del flusso delle informazioni, sempre più veicolato attraverso dispositivi di A.I. ma anche della produzione di valore. Si annuncia la sua trasformazione in unità di valore anche con la creazione di collaterali sul mercato finanziario.

La Cina ha ufficialmente designato ciyuan come traduzione di “token” – le unità computazionali che alimentano i Large Language Modules (LLM) come Claude, ChatGPT e Gemini – in una mossa ampiamente vista come la creazione di una nuova forma di valuta globale per l’era dell’intelligenza artificiale. In cinese, ‘ci’ si traduce come “parola”, mentre yuan è comunemente usato come sinonimo di “valuta”. Ad esempio, l’unità base del renminbi cinese è lo yuan, e la maggior parte delle valute estere sono indicate come yuan in cinese, precedute dai rispettivi nomi dei loro paesi

Sebbene il dollaro statunitense sia da tempo la base del sistema finanziario globale, la Cina cerca di ottenere un tipo diverso di vantaggio in una “token economy” – dove le metriche chiave sono la produzione di []token per watt e il costo per milione di token. Uno di questi vantaggi deriva dall’elettricità, la materia prima essenziale necessaria per produrre i token. La Cina è il più grande produttore di elettricità al mondo e la sua capacità energetica si è espansa a un ritmo che ha ripetutamente superato gli obiettivi ufficiali[2].

A sostegno dell’orientamento cinese sta la dichiarazione in proposito del CEO di Nvidia Jensen Huang.

I token sono la nuova merce,” ha dichiarato il CEO di Nvidia Jensen Huang, vestito con la sua iconica giacca di pelle, alla conferenza annuale per sviluppatori di punta dell’azienda, GTC, la scorsa settimana a San Jose, California. Il timone del progettista del chip vuole riproporre la sua azienda non come un fornitore di silicio, ma come l’architetto di quelle che chiama “fabbriche di intelligenza artificiale”, il cui prodotto standard è il “token”. Mentre Nvidia è impegnata a scrivere le regole di una nuova economia dei token, in Cina sta emergendo un dibattito parallelo sull’idea delle “esportazioni dei token*“. L’intelligenza generata dall’IA è essenzialmente un bene commerciabile misurato da token, e la Cina si sta posizionando lungo tutta la catena del valore – dall’energia e potenza di calcolo ai modelli e alla produzione*[3].

“Mentre le aziende cinesi di intelligenza artificiale continuano a spingere per l’adozione dei loro modelli di punta all’estero, hanno assunto un nuovo ruolo come fonte chiave delle cosiddette “esportazioni di token” verso il mercato globale. I modelli di IA cinesi hanno rappresentato quattro dei primi 10 modelli in termini di consumo di token sul popolare marketplace di modelli di IA OpenRouter[4] dal 18 marzo al 18 aprile, evidenziando la loro crescente visibilità nell’uso globale degli sviluppatori di sistemi di IA. La domanda interna di token è cresciuta a un ritmo esponenziale. I dati della National Data Administration cinese hanno mostrato che il consumo giornaliero di token aveva superato i 140 trilioni entro marzo 2026, un aumento di oltre 1.000 volte rispetto ai 100 miliardi all’inizio del 2024[5].”

Teniamo conto del fatto che “La Cina detiene un vantaggio sugli Stati Uniti quando si tratta di portare applicazioni di intelligenza artificiale nelle mani degli utenti comuni, secondo

dirigenti delle imprese tecnologiche e investitori, anche se avvertono che le aziende cinesi di IA appaiono sempre più sopravvalutate. La Cina era ancora indietro nella potenza di calcolo ma era solo “100 giorni indietro” rispetto agli Stati Uniti nelle capacità dei modelli di IA di frontiera, secondo Chi Zhang, direttore generale dell’industria finanziaria presso Alibaba Cloud Intelligence Group, parlando giovedì al HKEX Future Tech Summit 2026 a Shenzhen, ospitato da Hong Kong Exchanges and Clearing. Ma i maggiori vantaggi e potenziali del paese risiedono nelle applicazioni dell’IA grazie al suo vasto numero di imprenditori e ingegneri, così come nella fase attuale di sviluppo economico, ha dichiarato Zhang durante una tavola rotonda.”[6].
Significativa è l’analisi che viene fatta della nascita della Token economy nell’articolo Token Economics for LLM Agents: A Dual-View Study from Computing and Economics[7] che ha uno sviluppo analitico che non è riportabile nel livello di analisi di queste note. Il contributo consiste anche nella possibilità di consultare riferimenti bibliografici in aggiornamento.
Riferimento bibliografico https://github.com/SuDIS-ZJU/Token-Economics. “Questo repository organizza la letteratura dietro la nostra indagine per struttura di ricerca a livello di capitolo, coprendo l’economia dei token dalle fondazioni, l’ottimizzazione a singolo agente, il coordinamento multi-agente, la dinamica degli ecosistemi, l’economia della sicurezza e le opportunità future. È progettato come un archivio vivente di letteratura: invece di una bibliografia statica, funge da indice continuamente mantenibile degli articoli, raggruppati secondo il quadro analitico del sondaggio.”
*Il salto di qualità nell’affermarsi di una vera e propria Token Economy sta nello sviluppo della A.I. agentica, nell’affermarsi degli agenti, nella creazione di reti collaborative di agenti, vale a dire dispositivi di A.I. prima singoli poi sempre di più in rete, in grado di svolgere autonomamente compiti specifici essendo capaci di apprendere e autoregolarsi nel proprio contesto operativo. Il flusso di token non è più semplice come nell’interrogazione di un LLM, ma è sempre più stratificato e complesso nei processi di elaborazione interna al singolo agente, nelle reti collaborative e lo svolgimento dei propri compiti nel contesto assegnato. Entro un quadro così complesso si pongono problemi rilevanti per definire e quantificare il consumo e la produttività dei toke*n, *rispetto ai quali l’analisi è a tutt’oggi inadeguata, ponendo rilevanti problemi di gestione, progettazione e pianificazione.*

“Con l’evoluzione degli agenti LLM, i token sono emersi come i fondamentali primitivi economici dell’IA Agentica. Tuttavia, il loro consumo esponenziale introduce gravi colli di bottiglia computazionali, collaborativi e di sicurezza. Le indagini attuali rimangono frammentate tra ottimizzazione del sistema, progettazione architettonica e fiducia, prive di un quadro unificato per valutare il compromesso fondamentale tra qualità di output e costo economico. (…)
Storicamente, le principali epoche tecnologiche sono state definite dai cambiamenti nei loro primitivi economici fondamentali. Il kilowattora (kWh) ha galvanizzato l’Era Industriale, e la larghezza di banda di rete (GB) ha sostenuto l’Era dell’Informazione. Oggi, il “token”

alimenta l'Era dell'Intelligenza, l'era dell'IA generativa e degli agenti dei grandi modelli linguistici (LLM), fungendo da substrato universale della creazione digitale. Ogni interazione multimodale, che sia testo, visione o suono, viene infine distillata in flussi di token; Attraverso questi flussi, la cognizione umana si traduce in esecuzione meccanica. In questo nuovo paradigma, il token non funziona più semplicemente come unità tecnica di calcolo. È diventata la primitiva economica dell'IA agentica: l'unità fondamentale con cui l'intelligenza viene prodotta e misurata, e la valuta pratica con cui viene scambiata. In questo ruolo, segue la logica ferrea di qualsiasi risorsa fondamentale: man mano che l'economia costruita su di essa si espande, aumenta anche la domanda stessa della risorsa. Il gettone era quindi destinato a essere consumato a una scala che sfida l'estrapolazione lineare.

Questa tendenza è già in via di sviluppo, e da nessuna parte è più visibile che nell'ascesa dell'IA agentica [zhong2024memorybank, yue2025masrouter, bian2026tokendance]. A differenza dell'inferenza LLM a singolo passaggio tradizionale, i flussi di lavoro degli agenti operano attraverso cicli iterativi di ragionamento, uso degli strumenti e autocorrezione, con ogni ciclo che consuma token come input diretto alla cognizione. Inoltre, poiché i flussi di lavoro degli agenti sono intrinsecamente molto più intensivi in termini di token rispetto alle call LLM convenzionali, la loro proliferazione ha generato un aumento esponenziale dei consumi.

Con l'emergere di sempre più piattaforme agente e applicazioni per utenti finali (figura 1), il volume settimanale di elaborazione dei token sulla piattaforma OpenRouter è schizzato alle stelle da 0,4 trilioni a dicembre 2024 a 27,0 trilioni entro marzo 2026, un aumento di quasi 68 volte in soli 15 mesi.

La proliferazione delle architetture agentiche amplifica la tensione tra calcolo ed economia, rimodellando la ricerca di frontiera. A differenza del consumo lineare di token negli LLM convenzionali, i flussi di lavoro degli agenti moderni sono altamente iterativi. (…) Questo passaggio dall'inferenza isolata al coordinamento organizzativo introduce costi interni di transazione sostanziali e costi generali ridondanti che non possono essere catturati da una singola dimensione tecnica. (…)

La comunità accademica ha già sviluppato traiettorie di ricerca multidimensionali attorno all'economia dei token. Questi includono meccanismi di accelerazione dell'inferenza, ottimizzazione dell'invocazione della catena di strumenti e sistemi di memoria per agenti. Tuttavia, la letteratura esistente dei sondaggi rimane fortemente compartimentata in compartimenti tecnici isolati.

Questa frammentazione crea una limitazione centrale: non esiste ancora un linguaggio unificato per misurare il compromesso sistemico tra capacità algoritmica e sovraccarico di coordinamento. Poiché le indagini esistenti non trattano il token come un primitivo economico fondamentale — e, più specificamente, come fattore di produzione, mezzo di scambio e unità di conto — non possono spiegare pienamente perché le scelte ingegneristiche localmente ottimali spesso innescano diseconomie di scala globali nei flussi di lavoro complessi degli agenti. All'interno di silos di ricerca isolati, migliorare una dimensione spesso impone costi nascosti a un'altra. Ad esempio, massimizzare

aggressivamente il throughput del sistema può compromettere la qualità del ragionamento, mentre difese di sicurezza rigide possono aggravare l'attrito del token-economic. Senza una lente economica unificata, la vera frontiera Prodotto-Costo di Pareto rimane difficile da caratterizzare.
Infine, come riporta l'agenzia Reuters. Exclusive: China works on AI token futures market, sources say, in race with US[8].

La Cina sta progettando un mercato dei futures per i token di IA, hanno detto fonti a conoscenza della questione, poiché il paese potrebbe adottare una strada diversa rispetto alle borse statunitensi che sviluppano futures sulla potenza di calcolo per sfruttare la crescente inclinazione a coprire i costi dell'IA. La Shanghai Futures Exchange è nelle prime fasi di progettazione di contratti futures per i cosiddetti token AI – la più piccola unità di informazione elaborata dai modelli di IA, ha detto una delle persone a conoscenza della questione.
La ricerca della borsa di Shanghai sul design dei prodotti per i futures sui token è preliminare e guidata in parte dalla rivalità tra l'IA e gli Stati Uniti, secondo la prima fonte. (…) negli Stati Uniti si stanno preparando a lanciare futures di calcolo su GPU, che sono legati al costo di noleggiare la potenza di calcolo per l'IA. Al contrario, il prodotto della borsa di Shanghai sarebbe legato ai token di IA, utilizzati per il prezzo dei servizi di IA. Tutti questi prodotti derivati sono progettati per aziende lungo tutta la catena di approvvigionamento dell'IA, per coprirsi contro il costo della potenza di calcolo. (…)
La Cina vede l'IA come un settore strategico e un motore di crescita, e sta accelerando lo sviluppo di un mercato spot per la potenza di calcolo, supportato da operatori di data center, modelli di IA e altri che utilizzano la potenza di calcolo. I token, che sono una misura del consumo di calcolo, funzionano come il "carburante digitale" o la "materia prima" che alimenta i modelli di intelligenza artificiale, ha detto Xiao Feng, Presidente e CEO di HashKey Group. L'uso giornaliero dei token in Cina è aumentato di 1.000 volte dall'inizio del 2024, superando i 140 trilioni entro la fine di marzo, secondo i dati ufficiali.
L'amministratore delegato di BlackRock, Larry Fink, ha dichiarato in una conferenza all'inizio di questo mese che la domanda crescente di token potrebbe generare una classe di asset completamente nuova nell'acquisto di futures di computo.
Nel frattempo, la carenza di potenza di calcolo ha costretto molti modelli di IA cinesi a razionare l'accesso degli utenti negli ultimi mesi. Zhang Yunquan, ricercatore in tecnologia informatica presso l'Accademia Cinese delle Scienze, ha anche proposto a marzo il lancio dei 'compute futures' al parlamento cinese, secondo i media ufficiali."

Concludiamo queste note riportando le preoccupazione nelle big tech sul costo del consumo dei token[9].
"I costi interni dell'IA in crescita di Meta e il calo collettivo dei giganti tecnologici stanno mettendo sotto i riflettori una contraddizione di settore da tempo trascurata: la logica commerciale dell'IA viene minata proprio dalla crisi di fatturazione che ha creato. Secondo

The Information, Meta ha inviato questa settimana un promemoria interno a circa 6.000 dipendenti annunciando che avrebbe imposto limiti all'uso dei token e costruito una piattaforma di tracciamento in tempo reale per frenare la crescita esponenziale dei costi interni dell'IA. Il promemoria affermava esplicitamente che la spesa di Meta solo per l'uso interno dell'IA è prevista per raggiungere diversi miliardi di dollari nel 2026. Questa mossa arriva dopo che Meta aveva precedentemente fortemente incoraggiato i dipendenti a integrare strumenti di IA nei loro flussi di lavoro quotidiani, per poi essere ora costretta a un'improvvisa inversione di marcia."
Analogamente per Uber[10].
"Ogni cosa buona ha un costo e, per Uber, questo è il prezzo della sua massiccia adozione dell'IA. L'azienda tecnologica dei trasporti ha già esaurito il budget per l'IA per il 2026 a causa dell'aumento dell'uso di strumenti di programmazione, secondo il CTO dell'azienda Praveen Neppalli Naga. Le spese totali di ricerca e sviluppo (R&D) di Uber sono aumentate del 9% su base annua nel 2025, raggiungendo i 3,4 miliardi di dollari USA, con l'IA come fattore chiave dei costi. Come riportato da The Information, gli ingegneri stanno utilizzando ampiamente strumenti come Claude Code, con circa l'11% degli aggiornamenti live del backend di Uber completamente scritti da agenti AI."
La Token Economy sta superando soglie di complessità e di dimensioni tali da rendere problematico il suo sviluppo, mentre la sua pervasività è in grado dare una nuova forma al rapporto tra denaro e informazione. Con queste prime note abbiamo aperto il confronto sulle pagine della nostra rivista e non solo. E' solo un inizio nell'analisi e nello svolgimento concreto dei processi.

Roberto Rosso

1. https://arxiv.org/abs/1706.03762.[↩]
2. (https://www.scmp.com/news/china/science/article/3347887/china-names-trillions-ai-token-after-yuan-should-us-worry-dollar.[↩]
3. https://www.scmp.com/tech/big-tech/article/3347495/how-china-could-dominate-ai-eras-tokenomics-vast-power-grids-and-low-cost-models?module=inline&pgtype=article.[↩]
4. https://openrouter.ai/rankings.[↩]
5. https://www.scmp.com/tech/article/3351011/can-token-exports-give-china-edge-ai-era.[↩]
6. https://www.scmp.com/tech/big-tech/article/3356763/china-leads-us-everyday-ai-apps-firms-are-overvalued-experts-say".

Il termine token era diventato familiare con l'affermarsi delle cripto valute il cui sistema è basato sulla tecnologia BlockChain.

"Il termine "tokenomics" si riferisce attualmente all'economia dei token di criptovalute, ma potrebbe esserci una nuova definizione all'orizzonte man mano che l'adozione dell'IA cresce vertiginosamente in un contesto di grave carenza di offerta informatica: grandi

società (exchange) stanno progettando prodotti derivati attorno ai token di IA, che sono sempre più considerati meno un output computazionale e più un input di materie prime per infrastrutture, come elettricità o banda larga. La Shanghai Futures Exchange cinese sta attualmente progettando un mercato dei derivati per token AI, riporta Reuters. La notizia arriva mentre la principale borsa di derivati CME Group e l'Intercontinental Exchange (proprietaria della NYSE) hanno dichiarato separatamente di essere al lancio di contratti futures per il noleggio di GPU. (…)

Borse come ICE e CME affermano che il mercato del calcolo assomiglia sempre di più a un mercato globale delle materie prime rispetto al tradizionale mercato cloud. Ma mentre esistono mercati maturi per le GPU, c'è meno infrastruttura attorno ai token stessi — i mattoni fondamentali dei modelli di IA contemporanei. I piani enterprise per le grandi aziende di IA sono comunemente denominati in token: OpenAI, ad esempio, addebita 5 dollari per milione di token di input e 30 dollari per ogni milione di token di output se si vuole utilizzare l'API per il suo ultimo modello GPT-5.5. Anche i provider cloud offrono sempre più la possibilità di addebitare per token, come nel sistema Bedrock di Amazon" ((Just like gold and oil, we'll soon be able to trade AI token futures – Tcrunch https://techcrunch.com/2026/05/28/just-like-gold-and-oil-well-soon-be-able-to-trade-ai-token-futures/.[↩]

7. https://arxiv.org/html/2605.09104v1.[↩]
8. https://www.reuters.com/world/china/china-works-ai-token-futures-market-sources-say-race-with-us-2026-05-28/.[↩]
9. https://news.futunn.com/en/post/74550599/after-amazon-meta-has-also-imposed-limits-on-ai-usage?level=1&data_ticket=1781686986245915.[↩]
10. https://fortune.com/2026/05/26/uber-coo-ai-spending-tokens-claude-code/.[↩]