

# USA vs Cina

## Quale modello economico sta emergendo nel software AI?

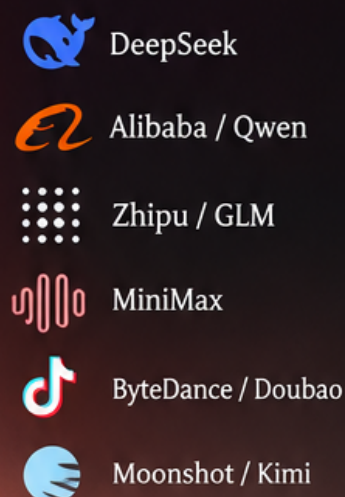
Innovazione, costi, investimenti, valutazioni e sostenibilità economica dei Foundation Models

UN CONFRONTO TRA LE PRINCIPALI REALTÀ USA E CINESI

### STATI UNITI



### CINA

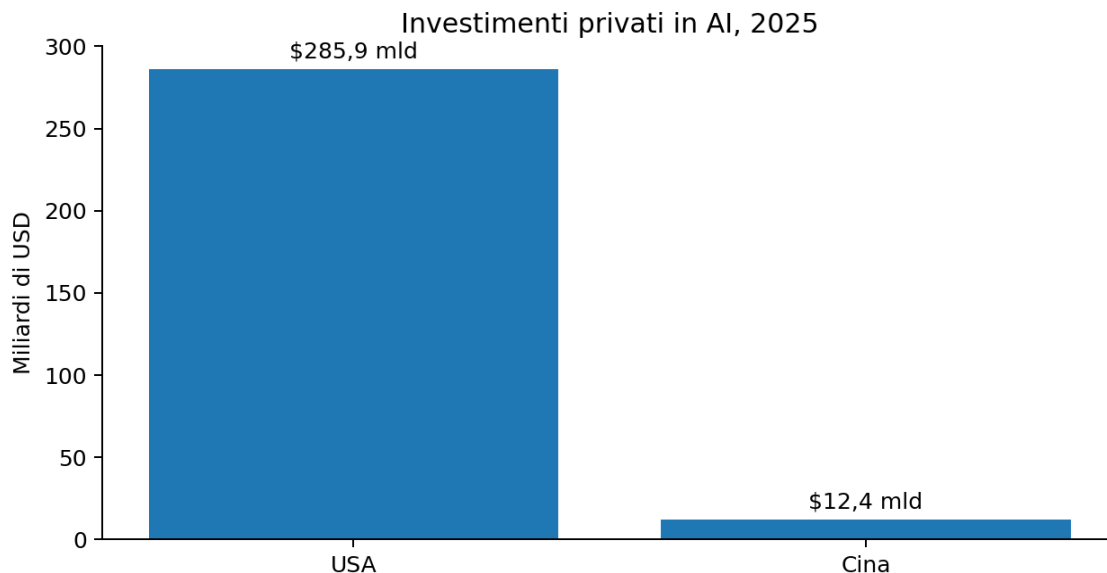


di  
**ROBERTO LO JACONO**

# 1. Executive summary

Domanda centrale. Quale modello economico sta emergendo nel software AI quando capacità sempre più simili vengono prodotte con combinazioni molto diverse di capitale, compute, architettura, distribuzione e prezzo?

Il quadro 2026 mostra una forte asimmetria finanziaria: Stanford HAI stima \$285,9 miliardi di investimento privato AI negli Stati Uniti nel 2025 contro \$12,4 miliardi in Cina, oltre 23 volte tanto. Il confronto va però interpretato con cautela perché il dato privato non cattura pienamente i government guidance funds e altri strumenti pubblici cinesi.



Fonte: Stanford HAI, AI Index Report 2026. Il dato privato può sottostimare il sostegno pubblico cinese.

Figura 1 - Investimento privato AI. Il grafico misura capitale privato a livello nazionale, non soltanto foundation-model software.

Sul piano industriale, gli USA combinano frontier labs pure-play e piattaforme integrate; la Cina combina grandi piattaforme digitali con laboratori che hanno reso l'efficienza architetturale e la pressione sui prezzi elementi centrali della competizione. La tesi del paper è che il confronto più utile non sia il solo benchmark tecnico, ma il costo per task riuscito e il margine catturato.

## AVVERTENZA

Documento di ricerca e analisi a finalità informativa. Non costituisce consulenza né raccomandazione di investimento. Si veda l'Avvertenza e limitazione d'uso a fine documento.

## FONTI E RIFERIMENTI - pagina 2

Stanford HAI, AI Index Report 2026, Top Takeaways / Economy: [hai.stanford.edu/ai-index/2026-ai-index-report](https://hai.stanford.edu/ai-index/2026-ai-index-report)  
Materiali forniti dall'autore: "AI - investimenti US vs CHINA" e "L'economia dell'IA cinese", agosto 2026.

## 2. Perimetro e metodo: Software AI, non infrastruttura

Il paper confronta i produttori di foundation models e i servizi software costruiti sopra di essi. Chip, data center ed energia sono inclusi soltanto come input economici del software AI, non come settori autonomi.

| Dimensione  | Che cosa confrontiamo                                 | Cautela metodologica                                       |
|-------------|-------------------------------------------------------|------------------------------------------------------------|
| Innovazione | capacità, architetture, multimodalità, agenti, coding | un benchmark non misura la qualità assoluta                |
| Economics   | training, inference, prezzo API, costo/task, margini  | prezzo API non equivale al costo industriale               |
| Capitale    | funding, valuation, impegni di compute                | valuation privata non equivale a market cap                |
| Adozione    | consumer, API, enterprise, distribuzione              | utenti non equivalgono a ricavi                            |
| Apertura    | proprietario, open-weight, licenze                    | open-weight non significa sempre open source               |
| Energia     | compute ed elettricità come costo del servizio        | dato nazionale non equivale al consumo del singolo modello |

### La catena economica corretta del software AI

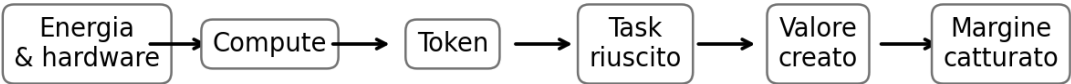


Figura 2 - La catena di unit economics. La metrica economicamente più utile è il costo per task riuscito, non il solo costo per token. Regola centrale:  $\text{prezzo/token} \neq \text{costo/token} \neq \text{costo/task} \neq \text{redditività dell'impresa}$ . I dati aziendali e i benchmark dichiarati dai produttori sono indicati come tali e non trattati come audit indipendenti.

### 3. Mappa competitiva: chi confrontiamo

| USA           | Modello economico / distribuzione         | Cina             | Modello economico / distribuzione      |
|---------------|-------------------------------------------|------------------|----------------------------------------|
| OpenAI        | consumer + enterprise + API; frontier lab | DeepSeek         | API + open-weight; efficienza          |
| Anthropic     | enterprise/API; coding e agenti           | Alibaba/Qwen     | cloud + open-weight + API              |
| Google Gemini | Search, Workspace, Cloud, Android         | Zhipu/GLM        | API + open-weight; pure-play           |
| xAI / Grok    | modello + piattaforma X + infrastruttura  | MiniMax          | consumer + API; internazionalizzazione |
| Meta / Llama  | open-weight + distribuzione Meta          | ByteDance/Doubao | consumer + cloud/API + distribuzione   |
|               |                                           | Moonshot/Kimi    | consumer + agenti + API                |

La distinzione più importante è tra pure-play frontier labs e piattaforme integrate. OpenAI, Anthropic, DeepSeek, Zhipu e MiniMax dipendono più direttamente dall'economia del modello. Google, Meta, Alibaba e ByteDance possono invece monetizzare l'AI anche attraverso cloud, advertising, productivity, social, e-commerce o distribuzione.

Per questo «USA vs Cina» è una scorciatoia geografica: la vera unità di analisi è il business model del provider e la sua capacità di trasformare capacità tecnica in ricavi e margini.

## 4. Innovazione: capacità e efficienza sono due frontiere

Nel 2026 la competizione tecnica non può essere descritta soltanto come una distanza lineare fra Stati Uniti e Cina. I laboratori americani continuano a competere sulla frontiera di reasoning, coding, multimodalità, tool use e agenti; i laboratori cinesi hanno reso molto visibile un secondo asse: quanta capacità utile si ottiene per unità di compute e di capitale.

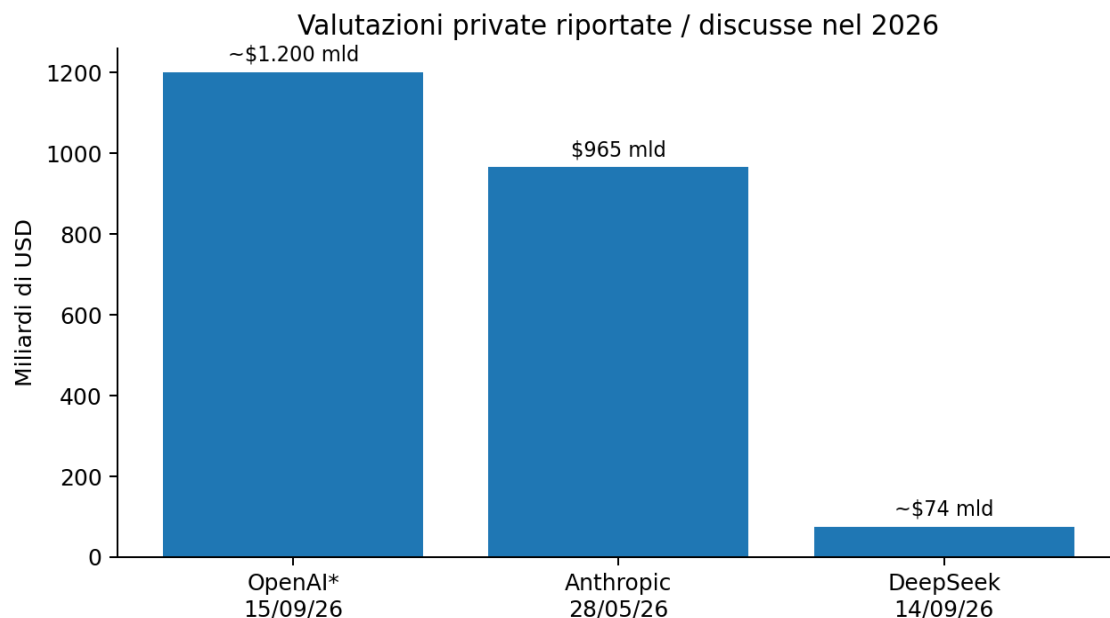
Il caso DeepSeek ha reso popolare questa lettura. Nel technical report di V3 il training run ufficiale è quantificato in 2,788 milioni di GPU-ore H800; la stessa fonte precisa che la cifra non include l'intero costo di ricerca e sperimentazione. Qwen e GLM hanno inoltre adottato estesamente Mixture-of-Experts e altre tecniche volte a ridurre i parametri attivi per token.

La conseguenza economica è importante: un modello che ottiene qualità comparabile con meno memoria, meno token o meno compute può comprimere il prezzo di mercato senza necessariamente comprimere nella stessa misura il proprio margine. L'efficienza è quindi una forma di innovazione, non soltanto una strategia low-cost.

Il paper evita volutamente di decretare un «vincitore» tecnico: benchmark, versioni e modalità di inference cambiano rapidamente e spesso non sono direttamente comparabili.

## 5. Capitale e valutazioni: una differenza di scala

La disponibilità di capitale determina quante iterazioni di training, quanta capacità di inference e quanto tempo un'impresa può sostenere prezzi aggressivi. Nel 2026 le valutazioni private dei principali laboratori sono diventate esse stesse una parte della tesi economica.



\*OpenAI: valutazione discussa in un possibile round, non round concluso. Fonti: Reuters/FT; Reuters.

Figura 3 - Valutazioni private riportate. Non sono direttamente comparabili con capitalizzazioni di Borsa; la cifra OpenAI è una valutazione discussa in un possibile round, non un round chiuso.

Reuters ha riportato il 15 settembre che OpenAI stava valutando un round a circa \$1,2 trilioni. Anthropic ha annunciato a maggio un round da \$65 miliardi a una valutazione di \$965 miliardi. Il 14 settembre Reuters ha riportato per DeepSeek una valutazione di circa RMB 500 miliardi (\$74 miliardi) nell'ambito di un round in corso.

Queste cifre misurano aspettative sul valore futuro, non la redditività corrente. Per Google/Alphabet, Meta, Alibaba e ByteDance il confronto è ancora più complesso perché il valore del modello AI è incorporato in gruppi con molte altre attività.

### FONTI E RIFERIMENTI - pagina 6

Reuters, 15 Sep 2026, OpenAI possible funding round / ~\$1.2tn valuation.

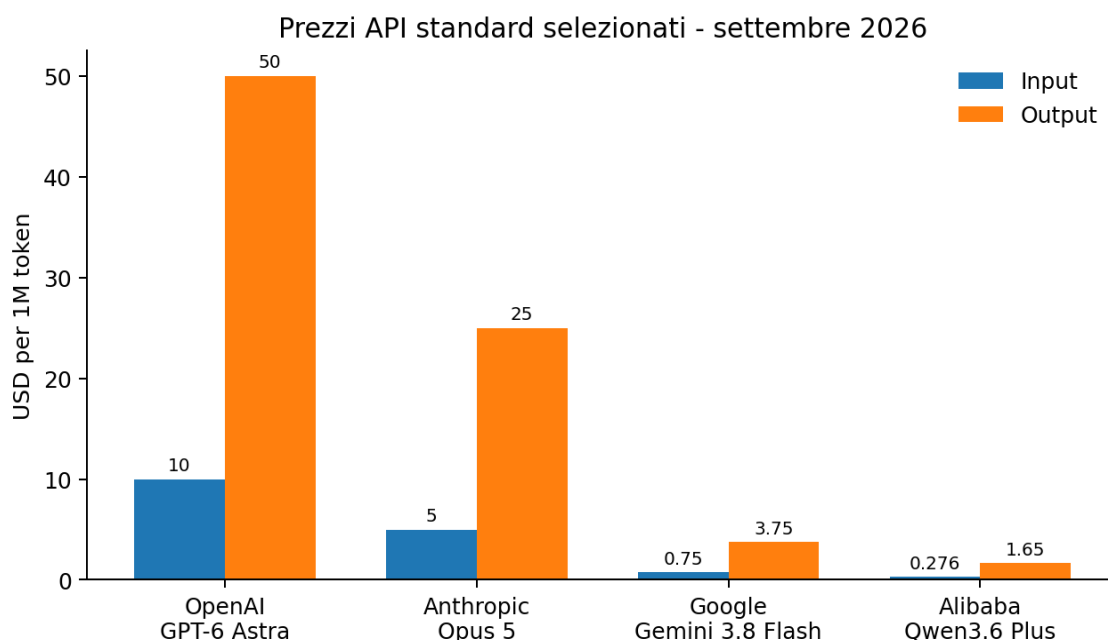
Reuters, 28 May 2026, Anthropic raises \$65bn / \$965bn valuation.

Reuters, 14 Sep 2026, DeepSeek CFO, funding and possible IPO / ~RMB500bn valuation.

Stanford HAI, AI Index Report 2026, Economy.

## 6. Prezzi API: la pressione competitiva è misurabile

I listini sono una delle poche grandezze direttamente osservabili e confrontabili, ma non vanno scambiati per costo industriale. A settembre 2026 la dispersione dei prezzi fra modelli è molto ampia.



Listini non equivalgono a costo industriale; modelli e qualità non sono direttamente equivalenti.

Figura 4 - Prezzi API standard selezionati. Google Gemini 3.8 Flash usa il prezzo promozionale valido fino al 31 dicembre 2026; Qwen3.6 Plus si riferisce alla fascia Global 0-256K.

OpenAI pubblica per GPT-6 Astra \$10/M token input e \$50/M output. Anthropic indica per Claude Opus 5 \$5/M e \$25/M. Google indica per Gemini 3.8 Flash \$0,75/M e \$3,75/M fino a fine 2026. Alibaba Cloud indica per Qwen3.6 Plus Global, entro 256K token, \$0,276/M input e \$1,651/M output.

La differenza di listino non implica equivalenza di qualità, latency, tool use o affidabilità. Per un cliente enterprise la metrica corretta resta \$/task riuscito a qualità costante.

### FONTI E RIFERIMENTI - pagina 7

OpenAI API pricing, consultato 19 Sep 2026: [openai.com/api/](https://openai.com/api/) ; [developers.openai.com/api/docs/models/gpt-6-astra](https://developers.openai.com/api/docs/models/gpt-6-astra)

Anthropic Claude pricing: [platform.claude.com/docs/en/about-claude/pricing](https://platform.claude.com/docs/en/about-claude/pricing)

Google Gemini API pricing: [ai.google.dev/gemini-api/docs/pricing](https://ai.google.dev/gemini-api/docs/pricing)

Alibaba Cloud Model Studio pricing: [alibabacloud.com/help/en/model-studio/model-pricing](https://alibabacloud.com/help/en/model-studio/model-pricing)

## 7. Redditività: crescita, compute e cash burn

Il software AI può avere crescita eccezionale e contemporaneamente richiedere finanziamenti eccezionali. Reuters, citando il Financial Times il 18 settembre 2026, ha riportato una previsione di circa \$278 miliardi di cash burn cumulato per OpenAI tra 2026 e 2030, principalmente per compute e infrastruttura. È una previsione riportata dalla stampa, non un bilancio auditato.

Anthropic ha raccolto \$65 miliardi nel maggio 2026 anche per espandere capacità di calcolo e prodotto. Sul lato cinese, i materiali di base riportano per MiniMax nel primo semestre 2026 \$116,6 milioni di ricavi, gross margin 17,9% e adjusted net loss di \$293 milioni; Zhipu riportava RMB 953,9 milioni di ricavi e circa RMB 2 miliardi di perdita netta.

Il punto è simmetrico: efficienza tecnica non equivale automaticamente a profitto; allo stesso modo, una forte crescita dei ricavi non garantisce un ritorno adeguato sul capitale quando il costo di compute e R&D cresce molto rapidamente.

Ne deriva una domanda chiave per i prossimi bilanci: il gross margin del serving migliora più rapidamente dell'aumento di spesa per modelli, agenti e capacità?

---

### **FONTI E RIFERIMENTI - pagina 8**

Reuters, 18 Sep 2026, OpenAI expects cash burn near \$278bn through 2030 (reporting on FT).

Reuters, 28 May 2026, Anthropic funding and valuation.

MiniMax 1H 2026 results: [minimax.io/news/minimax-announces-first-half-2026-financial-results-1787744160](https://minimax.io/news/minimax-announces-first-half-2026-financial-results-1787744160)

Reuters, 31 Aug 2026, Zhipu first-half revenue, losses and R&D riportato nel materiale dell'autore.

# 8. Proprietario vs open-weight: il problema della cattura del valore

| Strategia              | Vantaggio economico                                   | Rischio                                  |
|------------------------|-------------------------------------------------------|------------------------------------------|
| Proprietario frontier  | controllo di prodotto, pricing, sicurezza, dati d'uso | compute elevato e pressione competitiva  |
| Open-weight            | adozione, fine-tuning, ecosistema, standardizzazione  | commoditizzazione del modello            |
| Conglomerato integrato | distribuzione, cloud, advertising, productivity       | ROI del singolo modello poco trasparente |
| Pure-play lab          | focus, velocità, brand                                | dipendenza da funding e unit economics   |

Meta/Llama e Alibaba/Qwen mostrano come l'open-weight possa essere razionale quando la monetizzazione avviene altrove. DeepSeek rende ancora più evidente la possibilità che capacità avanzate e facilmente distribuibili comprimano la rendita del foundation model.

In questo scenario il valore può migrare verso cloud, dati proprietari, integrazione, distribuzione, agenti e applicazioni verticali. La domanda decisiva non è quindi soltanto «quanto valore crea l'AI?», ma chi riesce a trattenerlo?

## 9. Energia e compute: il costo fisico del software

Energia e hardware restano essenziali, ma in questo paper sono trattati come input del software. Il problema metodologico è che i provider non pubblicano una contabilità completa e comparabile del consumo elettrico e del costo marginale per singolo modello.

L'economia va letta lungo la catena energia → compute → token → task → valore. MoE, sparse attention, caching, batch inference e modelli specializzati possono ridurre il compute necessario per una certa quantità di lavoro. Al contrario, agenti e reasoning più lunghi possono aumentare i token e le chiamate per task.

Il paradosso di Jevons è quindi centrale: se il costo per task crolla, il numero di task può crescere più rapidamente del risparmio unitario. L'AI può diventare molto più efficiente e, nello stesso tempo, aumentare il consumo aggregato di elettricità.

Per confrontare USA e Cina in modo serio, la domanda non è «chi consuma meno elettricità in assoluto?», ma «quanta energia e quanto capitale servono per produrre un'unità comparabile di lavoro utile?»

# 10. La vera unit economics: dal token al task

| Metrica          | Che cosa misura                     | Limite                                         |
|------------------|-------------------------------------|------------------------------------------------|
| \$/1M token      | listino commerciale                 | non misura qualità né token necessari per task |
| token/task       | efficienza di ragionamento          | dipende da benchmark e prompting               |
| latency/task     | esperienza e costo infrastrutturale | non misura correttezza                         |
| success rate     | qualità operativa                   | serve un benchmark rappresentativo             |
| \$/task riuscito | unit economics del cliente          | richiede misurazione end-to-end                |
| gross margin     | sostenibilità del provider          | spesso non separa il singolo modello           |

Un modello cinque volte più caro per token può essere competitivo se completa il task al primo tentativo usando meno token, meno tool call e meno verifiche. Per gli agenti questa dinamica diventa ancora più importante.

Il benchmark economico ideale dovrebbe quindi misurare, sullo stesso paniere di task, qualità, token totali, tempo, retry e costo. Coding, analisi documentale, customer service e ricerca sono candidati naturali per un paniere standardizzato.

# 11. Scenari interpretativi 2027-2030

Scenario A - Commoditizzazione. Open-weight, distillazione, MoE e concorrenza riducono rapidamente il costo per task. Il valore migra verso applicazioni, dati proprietari e distribuzione; il pricing power del foundation model si riduce.

Scenario B - Frontier premium + mass market efficiente. Pochi modelli ad altissima capacità servono task ad alto valore, mentre modelli efficienti gestiscono la maggioranza dei volumi. Il mercato assume una struttura a classi di servizio.

Scenario C - Jevons dell'intelligenza. Il costo unitario scende e il numero di task cresce ancora più rapidamente. Ricavi, compute ed energia aggregati possono aumentare anche con prezzi per token in calo.

Gli scenari non sono mutuamente esclusivi. Potrebbero verificarsi contemporaneamente in segmenti diversi. La variabile che li distingue economicamente è dove rimane il margine dopo la riduzione del costo unitario.

## 12. Indicatori da monitorare nei prossimi 18 mesi

| Segnale                                   | Perché conta                                        |
|-------------------------------------------|-----------------------------------------------------|
| Prezzo per task a qualità costante        | misura la vera curva di costo del software AI       |
| Gross margin dei pure-play labs           | indica se inference e pricing diventano sostenibili |
| Ricavi / impegni di compute               | misura la leva finanziaria implicita della crescita |
| Quota open-weight nei workload enterprise | misura il rischio di commoditizzazione              |
| Retention e switching enterprise          | misura il pricing power reale                       |
| Token/task e success rate agentic         | misura l'efficienza oltre il listino                |
| Disclosure su energia/compute             | collega innovazione software a costo fisico         |
| IPO e disclosure dei lab cinesi           | può rendere confrontabili economics oggi opachi     |

DeepSeek ha nominato nel settembre 2026 il suo primo CFO mentre prepara una possibile IPO. Se il percorso proseguirà, una maggiore disclosure su ricavi, margini, compute e capitale potrebbe rendere il confronto fra pure-play USA e cinesi molto più rigoroso.

## 13. Conclusione

Il confronto fra software AI americano e cinese porta a una conclusione meno semplice di «USA più avanzati / Cina più economica». La competizione avviene contemporaneamente sulla capacità, sull'efficienza, sulla distribuzione e sulla disponibilità di capitale.

Gli Stati Uniti dispongono di una capacità eccezionale di finanziare frontier labs e di piattaforme globali in grado di distribuire rapidamente nuove capacità. La Cina combina minore capitale privato osservabile con forte concorrenza, grandi piattaforme domestiche e una pressione particolarmente visibile sull'efficienza e sui prezzi.

La variabile decisiva è la cattura del valore. Se l'intelligenza di base diventa abbondante e a basso costo, il successo tecnologico dell'AI non garantisce automaticamente rendimenti straordinari a chi produce il foundation model. Il valore può spostarsi verso chi possiede clienti, workflow, dati, distribuzione e applicazioni.

Il modello economico che sta emergendo appare quindi stratificato: frontier premium, modelli efficienti a basso costo, open-weight e piattaforme integrate. La metrica finale da osservare sarà il rapporto fra capitale investito, costo per task, qualità, adozione e margine catturato.

## 14. Avvertenza e limitazione d'uso

Il presente documento ha esclusivamente finalità informative, divulgative e di analisi. È stato predisposto attraverso la raccolta, il confronto e l'elaborazione di informazioni disponibili pubblicamente e delle fonti indicate nel documento, ritenute attendibili alla data di pubblicazione. Non viene fornita alcuna garanzia circa la completezza, l'accuratezza o il successivo aggiornamento delle informazioni riportate.

Le valutazioni, gli scenari e le considerazioni contenute nel documento rappresentano elaborazioni dell'autore sulla base delle fonti citate e non costituiscono consulenza finanziaria, di investimento, legale o fiscale, né una raccomandazione, un'offerta o una sollecitazione all'acquisto o alla vendita di strumenti finanziari o alla realizzazione di qualsiasi investimento.

Il documento non deve essere utilizzato come unica base per decisioni di investimento. Qualsiasi decisione finanziaria o di investimento deve essere assunta autonomamente dal lettore sulla base delle proprie valutazioni e, ove opportuno, con il supporto di professionisti qualificati.

Valutazioni societarie, prezzi API, benchmark, dati economici e altre informazioni di mercato possono cambiare rapidamente. I dati riportati devono pertanto essere considerati riferiti alla data indicata nel documento.

L'autore non assume responsabilità per decisioni di investimento o altre decisioni economiche assunte sulla base del presente documento, nei limiti consentiti dalla normativa applicabile.

### CRITERIO DI LETTURA

Prezzi API, valutazioni private e benchmark sono snapshot soggetti a variazione. Dove una fonte riporta stime, discussioni di funding o dati aziendali non auditati, il paper li presenta come tali e non come valori certi o consolidati.

# 15. Riferimenti principali e nota metodologica

## Fonti primarie / ufficiali

Stanford HAI - AI Index Report 2026: [hai.stanford.edu/ai-index/2026-ai-index-report](https://hai.stanford.edu/ai-index/2026-ai-index-report)  
OpenAI - API e pricing: [openai.com/api/](https://openai.com/api/) ; [developers.openai.com/api/docs/models/gpt-6-astra](https://developers.openai.com/api/docs/models/gpt-6-astra)  
Anthropic - Claude pricing: [platform.claude.com/docs/en/about-claude/pricing](https://platform.claude.com/docs/en/about-claude/pricing)  
Google - Gemini API pricing: [ai.google.dev/gemini-api/docs/pricing](https://ai.google.dev/gemini-api/docs/pricing)  
Alibaba Cloud - Model Studio pricing: [alibabacloud.com/help/en/model-studio/model-pricing](https://alibabacloud.com/help/en/model-studio/model-pricing)  
DeepSeek - V3 technical report: [github.com/deepseek-ai/DeepSeek-V3](https://github.com/deepseek-ai/DeepSeek-V3)  
Qwen - technical blog: [qwenlm.github.io/blog/qwen3/](https://qwenlm.github.io/blog/qwen3/)  
MiniMax - company results / technical material: [minimax.io](https://minimax.io)

## Stampa finanziaria

Reuters, 15 Sep 2026 - OpenAI possible funding round / valuation discussion.  
Reuters, 18 Sep 2026 - OpenAI projected cash burn through 2030 (reporting on FT).  
Reuters, 28 May 2026 - Anthropic \$65bn round / \$965bn valuation.  
Reuters, 14 Sep 2026 - DeepSeek CFO, fundraising and possible IPO.  
Reuters, 31 Aug 2026 - Zhipu first-half revenue, losses and R&D.;

## Materiali di partenza

“AI - investimenti S vs CHINA.docx” e “Report\_Economia\_IA\_Cinese\_Confronto(1).pdf”, forniti dall'autore e utilizzati come base concettuale per le ipotesi da verificare.

Nota metodologica. Le valuation private sono legate a specifici round o discussioni di funding e non equivalgono alla capitalizzazione di Borsa. I prezzi API sono listini dinamici e non rappresentano il costo industriale. I benchmark dei produttori non sono trattati come audit indipendenti. Dove non esiste disclosure verificabile su energia, costo marginale o margine del singolo modello, il paper segnala il limite anziché stimare valori non osservabili.