Gli assetti di mercato dei modelli di intelligenza artificiale generativa

di Rossana Arcano, Carlo Cambini, Paolo Lupi,
Antonio Manganelli, Antonio Perrucci, Giovanni Trotta

SOMMARIO: 1. Dinamiche competitive lungo la catena del valore. – 1.1. Il mercato dei chip e del cloud. – 1.2. Il mercato dei dati. – 1.3. Il mercato dei Foundation Models: numerosità dei modelli e concentrazione. – 2. Partnerships e dinamiche di integrazione verticale. – 3. Investimenti e sviluppo dei modelli USA, Europea e Cina.

1. *Dinamiche competitive lungo la catena del valore*

1.1. Il mercato dei chip e del cloud

Come anticipato, il principale hardware necessario per lo sviluppo di modelli di IAG è rappresentato dai *chip* acceleratori. Attualmente, il leader mondiale nello sviluppo di GPU per applicazioni di IA è NVIDIA, la cui posizione dominante è sostenuta dalla sua architettura chiusa (Box 4), dall'efficienza dei suoi acceleratori e dalla diffusione dell'architettura CUDA. Infatti, la Fig. 1 mostra che ad ottobre 2024, NVIDIA ha registrato un fatturato record di 35,08 miliardi di dollari, con una crescita significativa rispetto ai trimestri precedenti. Il segmento *data center*, che include le tecnologie per l'intelligenza artificiale e il calcolo accelerato, ha generato 30,77 miliardi di dollari, più del doppio rispetto allo stesso trimestre dell'anno precedente. Questa crescita esponenziale è attribuita principalmente alla crescita dell'IA e alla domanda di *chip* specializzati.

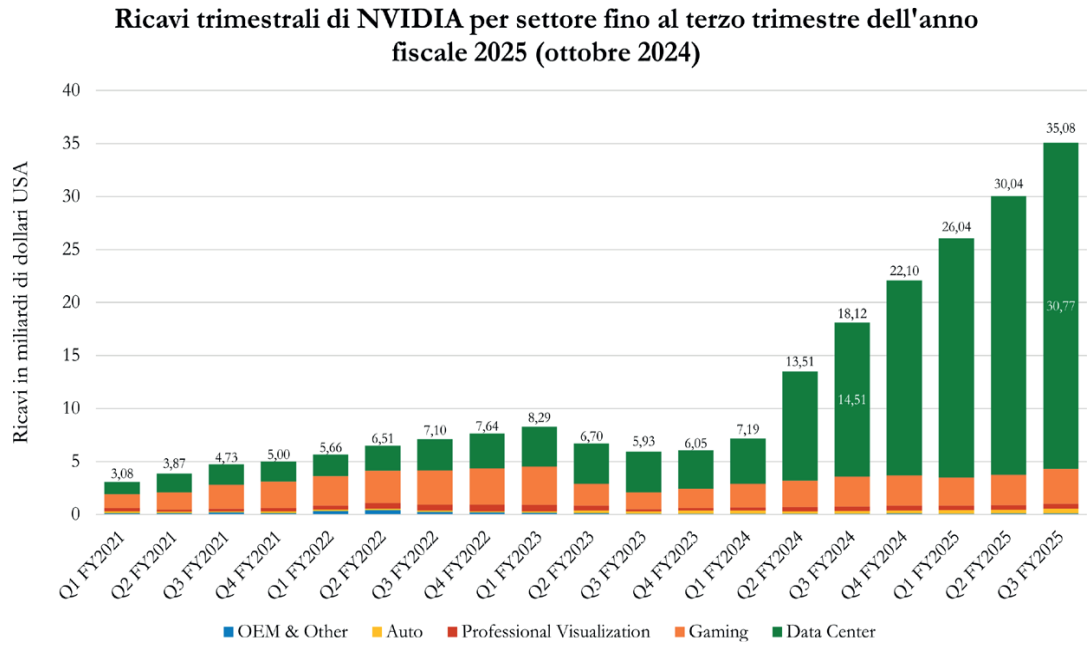

Fig. 1: Ricavi trimestrali di NVIDIA[1].

Box 4: NVLink e UALink

Un ulteriore elemento cruciale nella posizione dominante di NVIDIA è rappresentato dalla tecnologia *NVLink*, un'interconnessione proprietaria che consente alle GPU NVIDIA di comunicare tra loro con una velocità elevata. Tuttavia, questa architettura chiusa limita l'interoperabilità, in quanto NVLink è progettato esclusivamente per le GPU NVIDIA, creando un ecosistema hardware fortemente dipendente da un singolo produttore. Per rispondere a NVLink, è nato l'*Ultra Accelerator Link Promoter Group* (*UALink*)[a], un'alleanza tra le più grandi aziende tecnologiche come Intel, Google, Microsoft e Meta che mira a creare uno standard aperto per l'interconnessione tra acceleratori di diversi produttori. Questo standard permetterebbe di collegare acceleratori eterogenei, riducendo i costi e favorendo una maggiore concorrenza nel mercato degli acceleratori hardware per l'intelligenza artificiale. Attraverso un ecosistema interoperabile, le aziende avrebbero la possibilità di combinare hardware di diversi fornitori, eliminando il vincolo di dover dipendere esclusivamente dalle tecnologie NVIDIA. In questo modo, l'UALink non solo

[1] Nvidia (2025), *Nvidia specialized market revenue from fiscal year 2019 to 2025, by quarter (in million U.S. dollars)*, «Statista».

promuove un mercato più competitivo, ma incentiva anche l'innovazione tecnologica, riducendo le barriere all'ingresso per nuovi attori nel settore.

[a] F. Leone (2024, June 7), *Le più grandi aziende tecnologie fondano UALink Promoter Group*, «Italia A», https://ital-ia.it/ualink-promoter-group/.

Altri produttori, come Intel, stanno cercando di acquisire quote di mercato offrendo *chip* alternativi[2]; tuttavia, le loro soluzioni attuali non riescono a competere con le prestazioni dei *chip* acceleratori prodotti da NVIDIA. Per ridurre le dipendenze dall'azienda statunitense, i principali fornitori di infrastrutture hardware *cloud* per lo sviluppo di modelli, ossia Microsoft, Google e Amazon, stanno sviluppando internamente *chip* proprietari, trasformandosi così da clienti principali[3] a potenziali concorrenti. Un esempio sono gli acceleratori TPU di Google che rappresentano una soluzione hardware progettata specificamente per ottimizzare le operazioni di addestramento e inferenza nei modelli di intelligenza artificiale. Queste possono essere sfruttate tramite la piattaforma Cloud TPU di Google che permette di utilizzare i suoi acceleratori in base ad una tariffa a consumo misurata in ora *chip*. Anche Amazon Web Services ha rilasciato tramite la sua piattaforma *cloud* la possibilità di utilizzare i *chip* proprietari Trainium che permettono di ridurre i costi di addestramento dei modelli ottenendo allo stesso tempo elevate prestazioni. Tuttavia, questa trasformazione, sebbene in corso, non sembra rappresentare una minaccia immediata per NVIDIA. Infatti, oltre ad essere la principale fornitrice di GPU per l'addestramento dei modelli,

[2] M.R. Oregonian/OregonLive (2025, July 9), *Intel's CEO: We are not in the top 10' of leading chip companies. Oregonlive*, https://www.oregonlive.com/silicon-forest/2025/07/ intels-ceo-we-are-not-in-the-top-10-of-leading-chip-companies.html.

[3] Si stima che nel 2023 NVIDIA abbia spedito 150 mila GPU H100 a Microsoft e circa 50 mila unità a Google e Amazon. Statista (2024), *Estimated shipments of Nvidia H100 GPUs worldwide in 2023, by customer*, https://www.statista.com/statistics/1446564/ nvidia-h100-gpu-shipments-by-customer/.

è anche la principale cliente della *Taiwan Semiconductor Manufacturing Company* (*TSMC*), la più grande produttrice di semiconduttori al mondo. Il ruolo di TSMC è cruciale perché fornisce a NVIDIA le capacità produttive più avanzate, tra cui l'uso della tecnologia *Chip-on-Wafer-on-Substrate* (*CoWoS*), un sistema di *packaging* che migliora l'integrazione tra i componenti dei *chip*, aumentando l'efficienza e la velocità di elaborazione. La TSMC prevede di raddoppiare la sua attuale capacità produttiva a seguito della crescente domanda da parte di NVIDIA, Microsoft, Amazon e Alphabet, di cui più del 50%[4] verrebbe occupata esclusivamente da NVIDIA. È quindi probabile che questo controllo sulla catena di fornitura permetterà a NVIDIA di rimanere in una posizione dominante nel lungo periodo. A suggerirlo è anche l'adozione della nuova *GPU Blackwell* da parte di Microsoft, Google e Amazon, che hanno annunciato l'integrazione di questi *chip* nelle loro infrastrutture *cloud* per migliorare le capacità computazionali dedicate all'intelligenza artificiale. Questo ulteriore investimento da parte dei principali *hyperscaler* dimostra come, nonostante i tentativi di sviluppare soluzioni proprietarie, la dipendenza da NVIDIA rimanga ancora un elemento centrale nell'evoluzione del mercato dei *chip* per l'IA.

In un contesto in cui l'addestramento dei grandi modelli sembrerebbe dipendere da una sola azienda, per i modelli di piccole dimensioni numerose startup cercano di emergere con soluzioni più efficienti dal punto di vista computazionale ed energetico. Un esempio rilevante è la startup *deep tech* Neuronova di Milano che potrebbe aprire nuovi scenari rivoluzionari nell'implementazione dell'IA in piccoli dispositivi IoT. Neuronova sviluppa infatti *chip* per l'IA sfruttando le potenzialità delle *Spiking Neural Networks* (*SNN*). Le SNN sono delle reti neurali che si ispirano al funzionamento del cervello biologico che, a differenza delle reti tradizionali, comunicano attraverso degli impulsi elettrici chiamati "spike", in modo simile agli impulsi utilizzati dai neuroni biologici. Mentre le reti tradizionali hanno neuroni sempre attivi con alti costi computazionali ed energetici, nelle SNN vengono attivati solo i neuroni necessari in risposta a stimoli specifici portando enormi vantaggi in termini di efficienza energetica rendendo questi *chip* ottimali per l'elettronica di consumo.

[4] B. Neuro, (2024, 4 Novembre) *Nvidia e le altre spingono TSMC a raddoppiare il packaging*, «Yahoo Finance», https://it.finance.yahoo.com/notizie/ nvidia-e-le-altre-spingono-113049499.html.

1.2. Il mercato dei dati

Le stime riportate da recenti studi[5] evidenziano un problema strutturale che potrebbe limitare il futuro sviluppo dei modelli di IAG. Secondo queste analisi, la disponibilità di dati generati dall'uomo è destinata ad esaurirsi entro pochi anni. Tuttavia, il momento esatto in cui questi dati verranno utilizzati appieno dipende in larga misura dalla politica di scalabilità adottata durante l'addestramento dei modelli. Se i modelli venissero addestrati in modo ottimale dal punto di vista computazionale, ci sarebbero dati sufficienti per addestrare modelli fino al 2028. Tuttavia, le pratiche di addestramento recenti, come nel caso di Llama 3-70B, mostrano una tendenza al "sovra addestramento", in cui i modelli utilizzano meno parametri e più dati per ottimizzare l'efficienza computazionale durante l'inferenza. In scenari di sovra addestramento più modesti, ad esempio di 5 volte, si stima che lo *stock* di dati generati dall'uomo sarà completamente utilizzato entro il 2027. Invece, se si adotta una politica più aggressiva, come il sovra addestramento di 100 volte, lo *stock* di dati potrebbe esaurirsi già entro il 2025.

Questa dinamica porta molte aziende a orientarsi verso dati proprietari o acquisirli tramite licenza per sopperire alla crescente scarsità di dati generati dall'uomo. Tuttavia, l'adozione di tali dati introduce significative barriere competitive. I costi elevati associati alla loro acquisizione, manutenzione e preparazione rappresentano un vincolo accessibile principalmente alle grandi aziende con ampie risorse economiche. Questo controllo esclusivo sui dati consoliderebbe ulteriormente il potere di mercato degli *incumbent*, ampliando il divario competitivo e limitando le possibilità per le startup e le organizzazioni più piccole di entrare nel mercato.

A favorire ulteriormente l'utilizzo di dati proprietari è l'incertezza giuridica sull'utilizzo dei dati pubblici sia in termini legali sullo sfruttamento dei dati nell'addestramento sia per la protezione dei risultati generati dal modello. Come anticipato, i dati sintetici potrebbero rappresentare una soluzione alternativa ai dati pubblici; tuttavia, l'utilizzo di questi comporta dei costi che non tutte le aziende riuscirebbero a sostenere, inoltre la loro affidabilità è ancora argomento di discussione, in quanto attualmente hanno dimostrato le loro capacità solo in ambiti ristretti come la matematica e la codifica.

[5] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim e M. Hobbhahn (2024), *Will we run out of data? Limits of LLM scaling based on humangenerated data*, https://arxiv.org/abs/ 2211.04325.

1.3. Il mercato dei Foundation Models: numerosità dei modelli e concentrazione

Per analizzare il mercato dei FMs e offrire una panoramica completa, in questa sezione vengono presentati diversi grafici costruiti a partire dal database *Ecosystem Graphs del CRFM* (*Center for Research on Foundation Models*) di Stanford che, introdotto nel 2023, traccia l'ecosistema dei FMs, includendo modelli, applicazioni e dataset. La Fig. 2 mostra il numero di modelli rilasciati globalmente dal 2019 al 2024 suddivisi in base alle modalità di accesso (*closed*, *limited* e *open*) anticipate nel secondo capitolo di questo elaborato. L'analisi dell'evoluzione del numero di modelli evidenzia in primo luogo una crescita esponenziale nella loro diffusione. In particolare, il numero totale è aumentato da pochi esemplari negli anni 2019 e 2020 fino a un massimo di 184 modelli totali nel 2023. La riduzione a 133 modelli nel 2024 potrebbe suggerire che, dopo una rapida espansione iniziale, il mercato stia entrando in una fase di consolidamento. Dalla suddivisione tra le modalità di accesso, emerge invece che, nonostante i modelli *open source* offrano generalmente prestazioni inferiori rispetto ai modelli *closed*[6], il loro sviluppo ha registrato una crescita significativamente maggiore, raggiungendo nel 2023 un numero quasi quattro volte superiore a quello dell'anno precedente. Anche nel 2024, il numero dei modelli *open source* rappresenta la percentuale maggiore, pari a circa il 72% del totale. Questo mette in luce la tendenza nel mercato all'innovazione e alla ricerca, in quanto anche se i modelli *open* non offrono un elevato controllo e protezione, consentono ad un numero ampio di attori di testare, personalizzare e innovare, contribuendo alla crescita complessiva del settore.

[6] Vedere sezione 4.1 del capitolo precedente.

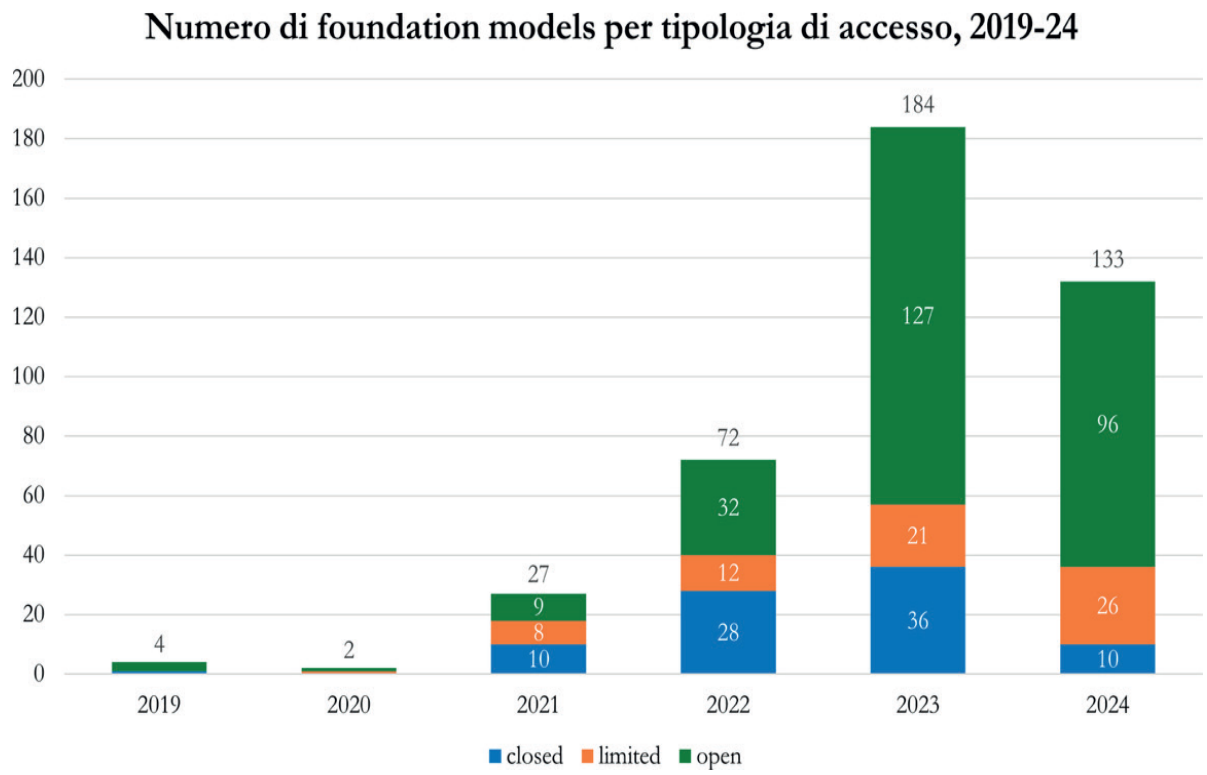

Fig. 2: Numero di FMs per tipologia di accesso, 2019-24[7].

Oltre alla suddivisione per tipologia di accesso, un ulteriore aspetto da analizzare è il numero di modelli rilasciati dai diversi *players* nel mercato. I successivi due grafici (Fig. 3 e Fig. 4) mostrano, rispettivamente, il numero di modelli rilasciati dai primi 15 *player* più attivi nel periodo 2019-2024 e il numero totale rilasciato, sempre dai 15 *player* più attivi, solo nel 2024. Quest'analisi permette di evidenziare, oltre che i principali attori, anche eventuali tendenze di concentrazione del mercato e l'evoluzione dei diversi sviluppatori nel tempo.

[7] *Ecosystem graphs for foundation models*, CRFM Stanford.

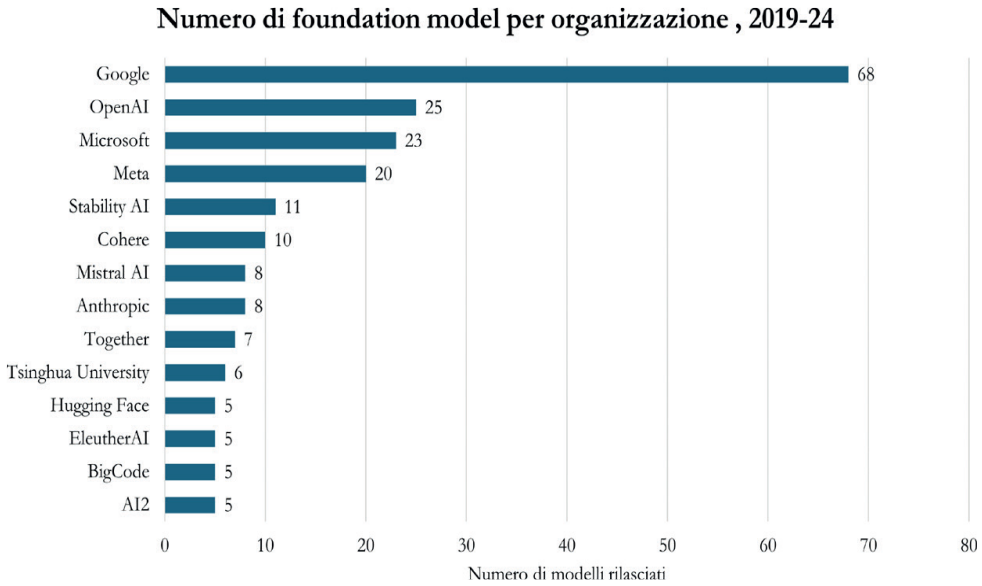

Fig. 3: Numero di FMs dei primi 15 *player* più attivi rilasciati nel periodo 2019-2024[8].

L'analisi della distribuzione dei modelli tra il 2019 e il 2024 evidenzia che la maggior partecipazione all'interno del mercato è data da 4 principali società: Google (68), OpenAI (25), Microsoft (23) e Meta (20). Complessivamente, questi attori hanno rilasciato 136[9] modelli nell'arco di questi 6 anni. Questo dato assume particolare importanza se confrontato con il numero dei modelli totali a livello globale e il numero di sviluppatori a questi associati. Infatti, il numero totale di attori che hanno sviluppato almeno un FM nei 6 anni passati ammonta a circa 155 e comprende aziende private, istituzioni accademiche, organizzazioni no-profit, enti governativi e collaborazioni scientifiche internazionali. I quattro precedenti attori rappresentano quindi circa il 2,6% del panorama mondiale e solamente i loro modelli costituiscono quasi un terzo (32%) di tutti quelli sviluppati globalmente.

Questi dati suggeriscono che storicamente il settore è stato fortemente dominato da un ristretto gruppo di aziende tecnologiche, le quali, probabilmente grazie al loro accesso privilegiato a risorse computazionali, dati e competenze altamente specializzate, riescono a mantenere una posizione predominante nello sviluppo dell'IAG. Questa elevata concentrazione del mercato solleva interrogativi sul grado di concorrenzialità del settore, sulle barriere all'ingresso per nuovi attori e sul possibile impatto delle politiche regolatorie nel riequilibrare la distribuzione del potere tra i vari

[8] *Ibidem*.

[9] Da questo dato sono esclusi i modelli rilasciati in collaborazione con altre società o con istituzioni accademiche.

sviluppatori di modelli. Tuttavia, l'analisi della distribuzione dei FMs sviluppati nel 2024 rispetto all'intero periodo 2019-2024 suggerisce una diversificazione tra gli attori coinvolti. Se nel lungo periodo il mercato è stato dominato dalle *big tech*, con Google, OpenAI, Microsoft e Meta responsabili di oltre un terzo dei modelli sviluppati globalmente, il 2024 mostra un panorama più distribuito e competitivo.

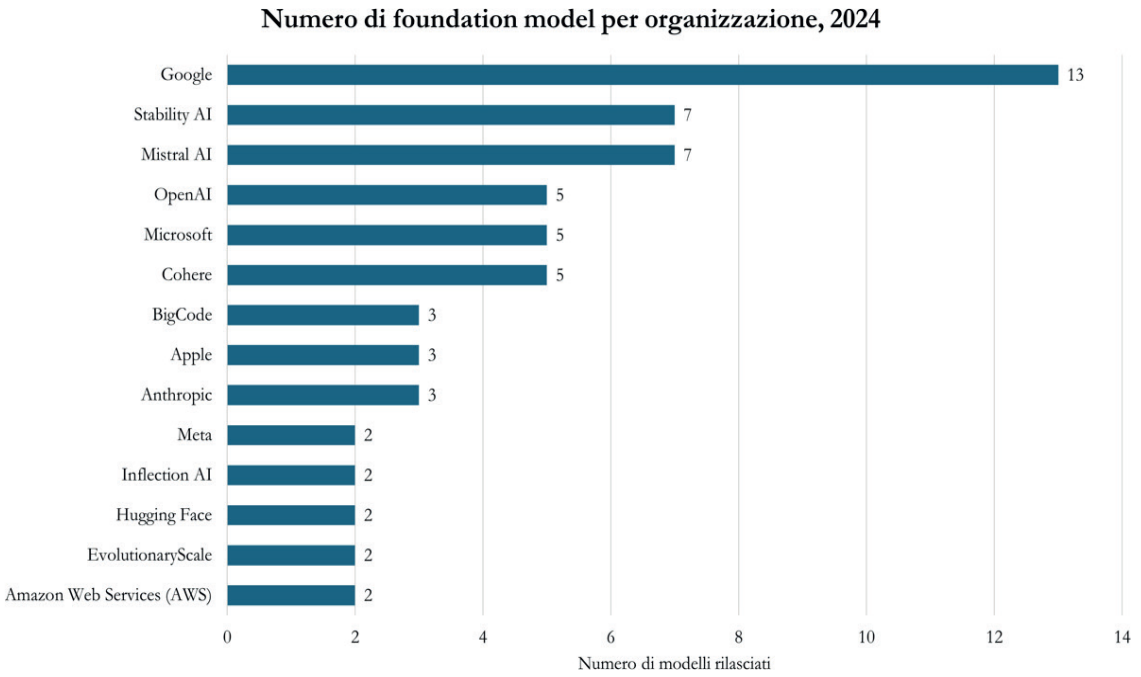

Fig. 4: Numero di FMs dei primi 15 *player* più attivi rilasciati nel 2024[10].

Google emerge nuovamente come il principale sviluppatore, con 13 modelli rilasciati. Tuttavia, Stability AI e Mistral AI si posizionano tra i leader del 2024 con 7 modelli ciascuna, evidenziando l'ingresso di nuovi attori che potenzialmente potrebbero competere con le grandi aziende tecnologiche. Inoltre, è importante evidenziare come questi due attori, rispettivamente con sede nel Regno Unito e in Francia, abbiano sviluppato un totale di 8 e 11 modelli nell'arco dei 6 anni di cui la maggior parte solamente del 2024. Questo potrebbe suggerire una partecipazione più attiva da parte di nazioni differenti in un mercato fino a questo momento dominato esclusivamente da aziende statunitensi e cinesi.

Tuttavia, il numero di modelli sviluppati e rilasciati dalle diverse organizzazioni fornisce sì un'indicazione della competitività del settore, ma non è sufficiente di per sé a determinare quali tecnologie stiano effettivamente dominando l'attenzione degli

[10] *Ecosystem graphs for foundation models,* CRFM Stanford.

utenti. Per comprendere l'effettivo impatto dei FMs sul mercato e il loro livello di adozione da parte degli utenti, è utile analizzare il traffico web generato dalle piattaforme che li ospitano. Alcuni modelli, pur essendo altamente innovativi, potrebbero avere un impatto limitato in termini di utilizzo, mentre altri potrebbero ottenere una diffusione ben superiore grazie a strategie di distribuzione efficaci o a un'integrazione diffusa in applicazioni di largo consumo. L'analisi del traffico web dei principali modelli consente quindi di valutare il successo commerciale e la rilevanza di ciascuna tecnologia, evidenziando quali attori stiano consolidando la propria leadership e quali, invece, fatichino a emergere nonostante gli investimenti nello sviluppo.

Per la seguente analisi vengono confrontate le piattaforme di 8 *chatbot* di IAG: ChatGPT (OpenAI), Le Chat (MistralAI), Gemini (Google), Meta AI (Meta), Claude (Anthropic), Copilot (Microsoft), Qwen2.5-Max (Alibaba) e DeepSeek (DeepSeek AI). La Fig. 5 mostra la percentuale di traffico a livello globale delle piattaforme[11] sul totale del traffico generato da queste ultime nei mesi di dicembre 2024, gennaio 2025 e la prima settimana di febbraio 2025.

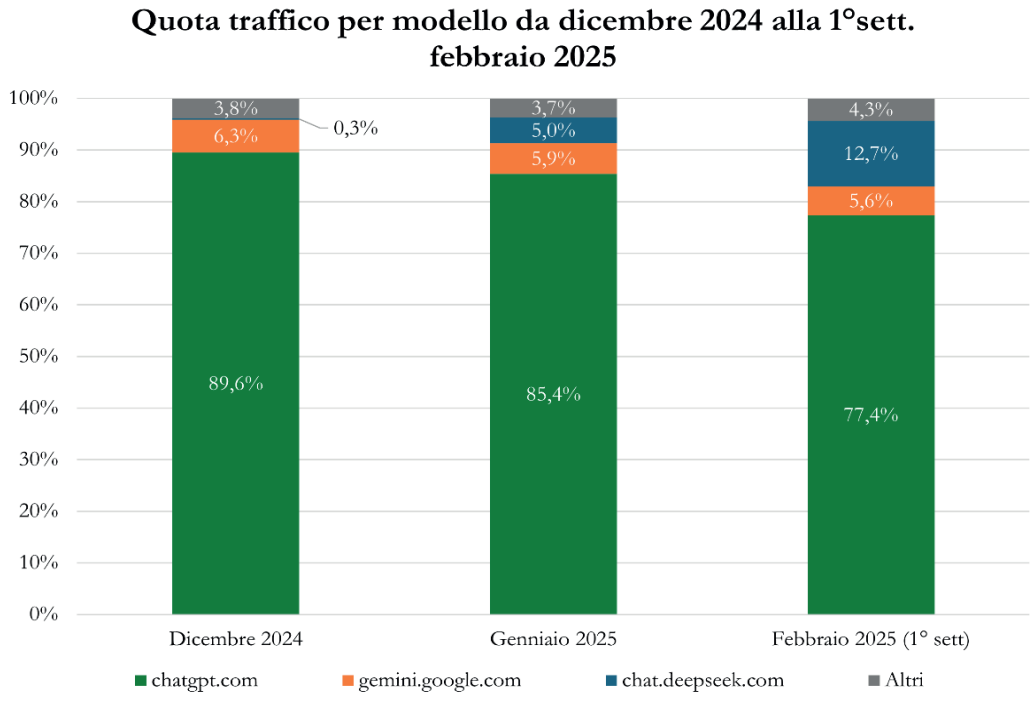

Fig. 5: Quota traffico generato dalle piattaforme di alcuni FM[12].

[11] Nel grafico i competitor evidenziati in grigio (Altri) fanno riferimento alle seguenti piattaforme: claude.ai, copilot.microsoft.com, qwenlm.ai, meta.ai, chat.mistral.ai.

[12] Fonte: Similarweb.

L'analisi combinata dei dati sul traffico web delle principali piattaforme che permettono lo sfruttamento dei FMs e del numero di modelli sviluppati dalle aziende tra il 2019 e il 2024 evidenzia dinamiche di mercato particolarmente rilevanti. In primo luogo, ChatGPT di OpenAI continua a dominare il settore, con una quota di traffico che, nonostante un lieve calo dal 89,56% di dicembre 2024 al 77,36% nella prima settimana di febbraio 2025, rimane nettamente superiore a quella di qualsiasi altro competitor. Questo dato è particolarmente significativo se confrontato con il numero di modelli sviluppati dalle diverse aziende: Google, leader per numero di FMs rilasciati negli ultimi sei anni (68), non riesce a tradurre questa superiorità in un dominio effettivo sul mercato, con il suo modello di punta Gemini che raccoglie appena il 5,62% del traffico nella prima settimana di febbraio 2025, confermando una posizione marginale rispetto a ChatGPT.

Un altro elemento di particolare rilievo è rappresentato dalla rapida ascesa di DeepSeek AI, la cui piattaforma chat.deepseek.com, che consente l'utilizzo dei suoi modelli *open source*, è passata dallo 0,26% di dicembre 2024 al 12,7% nella prima settimana di febbraio 2025, registrando una crescita esponenziale in un arco temporale estremamente ridotto. Questo incremento repentino ha inizialmente rappresentato un'anomalia nel contesto competitivo, suggerendo l'adozione di una strategia di penetrazione particolarmente efficace. L'analisi approfondita di tale strategia e delle motivazioni per cui il rilascio dei modelli *open source* di DeepSeek abbia inciso in modo così marcato sulle dinamiche di mercato è oggetto del Box 5.

Un'analisi ulteriore[13] dei dati relativi al traffico web nel periodo marzo-maggio 2025 conferma, tuttavia, che l'ascesa di DeepSeek non è stata un fenomeno transitorio legato unicamente all'effetto novità: la sua quota di traffico si è infatti stabilizzata su un valore pari al 7,38% (Fig. 6), risultando comparabile a quella della piattaforma che ospita Gemini (7,08%). Tale andamento indica la capacità dell'azienda di consolidare una base utenti stabile, a conferma della solidità della propria offerta tecnologica. La stabilizzazione su livelli rilevanti suggerisce, in ultima analisi, che DeepSeek abbia raggiunto una fase di maturazione della propria presenza nel mercato, superando la fase iniziale di rapida espansione.

[13] In questo caso sono stati comparati cinque dei principali modelli del mercato.

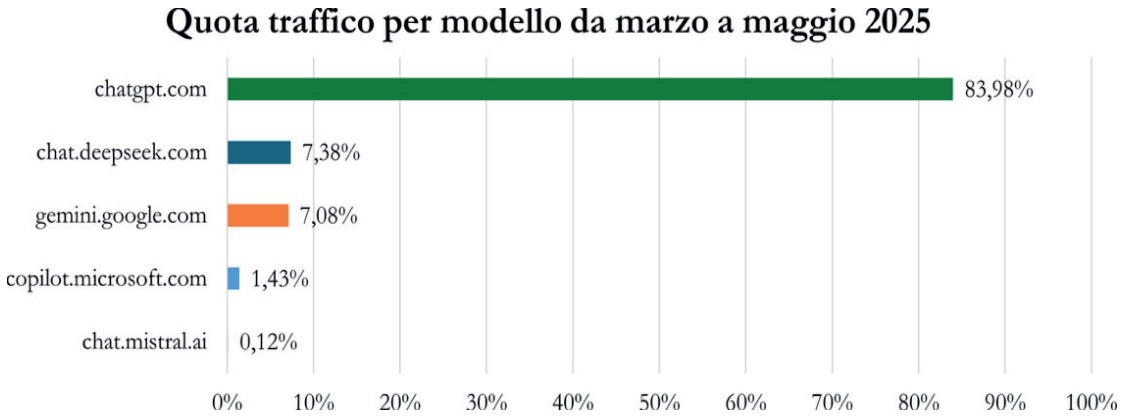

Fig. 6: Quota traffico generato dalle piattaforme di alcuni FM, marzo-maggio 2025[14].

Box 5: DeepSeek AI

DeepSeek è una startup cinese fondata nel maggio 2023 e costituita principalmente da neolaureati e giovani ricercatori che, a gennaio 2025, ha scosso i mercati globali rilasciando i suoi modelli di IAG in modalità *open source*. Ciò che ha reso i modelli di DeepSeek (DeepSeek-V3 e DeepSeek-R1) un fenomeno globale sono le loro prestazioni in relazione al costo di sviluppo. Fino a prima dell'avvento di questi modelli, i costi di sviluppo per l'addestramento dei FMs, in particolar modo quelli legati alle risorse computazionali necessarie all'addestramento, come anche anticipato nel capitolo precedente, rappresentavano un ingente investimento e quindi una barriera all'ingresso per numerosi potenziali concorrenti. DeepSeek, invece, dichiara che non solo i suoi modelli superano in termini prestazionali i modelli di OpenAI (e di altri competitors) su diversi *benchmark*, ma il costo di addestramento (pari a circa 6 milioni di dollari) è inferiore del 90-95% rispetto a quello sostenuto per addestrare i modelli di punta di OpenAI. Questa riduzione dei costi è attribuita all'utilizzo di solo 2048 GPU NVIDIA H800[a], acceleratori che hanno sia costi che prestazioni inferiori rispetto alle GPU di punta H100. DeepSeek si sarebbe quindi trovata quasi obbligata a utilizzare GPU meno performanti a causa delle rigide restrizioni sulle esportazioni tecnologiche dagli

[14] Fonte: Similarweb.

Stati Uniti, puntando quindi sullo sviluppo di tecniche che aumentassero l'efficienza del modello.

Tuttavia, il costo così ridotto per lo sviluppo dei suoi modelli ha sollevato alcuni dubbi sulla veridicità delle dichiarazioni della startup: ad esempio il CEO di Scale AI, Alexandr Wang, durante un'intervista ha sostenuto che in realtà DeepSeek possieda 50.000 *chip* GPU H100 che non possono essere dichiarate proprio a causa delle restrizioni sulle esportazioni. Inoltre, diverse analisi indipendenti (come quella di SemiAnalysis[b]) hanno osservato che la cifra di circa 6 milioni di dollari resa pubblica da DeepSeek rappresenta soltanto il costo di una specifica fase di addestramento su GPU H800, mentre il costo complessivo dell'infrastruttura e delle operazioni necessarie sarebbe nell'ordine di miliardi di dollari, se si considerano CAPEX e OPEX. Non solo, la stessa OpenAI ha accusato la startup cinese su una possibile violazione della proprietà intellettuale, in quanto secondo una speculazione DeepSeek avrebbe usato i modelli di OpenAI per addestrare i suoi modelli tramite tecniche di distillazione, il che violerebbe i termini d'uso di OpenAI.

Nonostante questo, DeepSeek ha completamente sconvolto i mercati, causando la più grande perdita azionaria giornaliera nella storia: NVIDIA a seguito del rilascio dei modelli di DeepSeek, ha perso 590 miliardi di dollari di capitalizzazione (calo del 17%)[c]. Questo calo potrebbe rispecchiare la sfiducia degli investitori in un mercato che fino a quel momento si pensava potesse essere dominato esclusivamente da chi possedeva un numero elevato di modelli di GPU NVIDIA di punta.

L'ingresso di DeepSeek ha quindi segnato il panorama globale, in quanto la strategia open source basata sull'efficienza, in netto contrasto con quella adottata dalle aziende statunitensi, ha sollevato sia interrogativi sulle modalità di sviluppo sia sulle attuali regolamentazioni, dando il via a una corsa allo sviluppo a livello globale che verrà analizzata nel seguito di questo elaborato.

[a] DeepSeek-Ai (2024, December 27), *DeepSeek-V3 Technical Report*, https://arxiv.org/abs/2412.19437.

[b] Feed (2025, January 31), *Is DeepSeek lying? CEO of Scale AI Alexandr Wang says the Chinese startup is using 50,000 Nvidia H100 chip*, «The Economic Times», https://economictimes.indiatimes.com/news/international/us/is-deepseek-lying-ceo-of-scale-ai-alexandr-wang-says-the-chinese-startup-is-using-50000-nvidia-h100-chips-butworkers-cant-talk-about-it/articleshow/117652897.cms.

[c] L. Bratton (2025, January 27), *Nvidia stock plummets, loses record $589 billion as DeepSeek prompts questions over AI spending*, «Yahoo Finance», https://finance.yahoo.com/news/nvidia-stock-plummets-loses-record-589-billion-asdeepseek-prompts-questions-over-ai-spending-135105824.html.

[d] SemiAnalysis (2025, January 31), *DeepSeek debates: Chinese leadership on cost, true training cost, closed model margin impacts*, SemiAnalysis, https://semianalysis.com/2025/01/31/deepseek-debates/

L'analisi congiunta di questi dati suggerisce dunque che il numero di modelli sviluppati non rappresenta, di per sé, un indicatore sufficiente per determinare il successo di mercato o la concentrazione dello stesso. Infatti, tale numero può essere fuorviante anche per ulteriori ragioni: spesso, le organizzazioni sviluppano un singolo modello base accompagnato da numerosi altri modelli derivati, progettati per essere più economici o più veloci rispetto al modello principale[15]. Questi modelli rispondono prevalentemente a logiche di differenziazione del prodotto piuttosto che riflettere una reale frammentazione competitiva del mercato. Tuttavia, se da un lato OpenAI, con un numero inferiore di modelli rispetto a Google, ha consolidato la propria leadership attraverso un singolo prodotto dominante, dall'altro la rapida crescita di DeepSeek dimostra che nuovi entranti possono scalare rapidamente se dotati di un prodotto competitivo e di una strategia efficace di distribuzione. Questo scenario potrebbe prefigurare una maggiore frammentazione del mercato nel prossimo futuro, con il potenziale emergere di nuovi concorrenti in grado di sfidare il predominio consolidato di OpenAI, derivato in parte dalle sue strategie di integrazione verticale.

2. *Partnerships e dinamiche di integrazione verticale*

L'integrazione verticale rappresenta una delle strategie più rilevanti nel mercato dell'IAG, permettendo alle aziende di acquisire un maggiore controllo sulla catena del valore e di ottenere vantaggi competitivi significativi. In questo contesto, l'integrazione può avvenire sia a monte che a valle. L'integrazione a monte fa riferimento al controllo delle infrastrutture computazionali nonché sui dati necessari all'addestramento del modello. In quella a valle, invece, l'integrazione avviene con le piattaforme di

[15] Come anticipato nella sezione 3.2 del capitolo precedente.

distribuzione dei servizi basati sui FMs. Queste tipologie di integrazione, che possono avvenire singolarmente o in modo congiunto, forniscono vantaggi competitivi significativi per le aziende integrate ma creano potenziali barriere di ingresso per i nuovi partecipanti. Partendo dalle integrazioni a monte, di seguito vengono esaminate due casistiche:

1. *controllo diretto sulle risorse computazionali*: alcune aziende sviluppatrici di modelli posseggono risorse computazionali proprietarie utilizzate per addestrare i propri modelli. Questo garantisce un accesso stabile e prioritario alla potenza di calcolo necessaria per l'addestramento e diminuisce la dipendenza dai fornitori esterni. Un esempio è rappresentato dai già citati *chip* acceleratori TPU di Google, utilizzati per addestrare i suoi modelli come PaLM e Gemini, o dalle GPU H100 utilizzate da NVIDIA per addestrare la famiglia di modelli open source NVLM;

2. *accordi esclusivi con fornitori di cloud computing*: un ulteriore metodo per ottenere risorse computazionali senza possedere acceleratori interni è stipulare accordi esclusivi con i *provider* di *cloud computing*. Ad esempio, nel box 6 viene approfondito l'accordo tra OpenAI e Microsoft, in cui quest'ultima fornisce tutta l'infrastruttura *cloud* di Azure in modo esclusivo a OpenAI sia per l'addestramento che per la distribuzione. A sua volta, Microsoft ha stipulato un accordo con Coreware, una società sostenuta da NVIDIA che permette di affittare le sue GPU. Un altro esempio rilevante è rappresentato da Google Cloud, il quale è il *provider* preferenziale per la società sviluppatrice Anthropic.

Box 6: Accordo tra OpenAI e Microsoft

La collaborazione tra Microsoft e OpenAI ha avuto inizio nel 2019, quando Microsoft ha effettuato un investimento iniziale di un miliardo di dollari in OpenAI[a], con l'obiettivo di accelerare le innovazioni nel campo dell'intelligenza artificiale e di democratizzare l'accesso a queste tecnologie avanzate. Questa *partnership* è stata ulteriormente rafforzata nel gennaio 2023[b], con un investimento pluriennale e multimiliardario da parte di Microsoft, volto a estendere la collaborazione nella ricerca sull'IA e nello sviluppo di supercomputer specializzati.

I termini dell'accordo prevedevano che Microsoft diventasse il fornitore *cloud* esclusivo per OpenAI, con Azure che alimentava tutti i carichi di lavoro. In cambio, Microsoft ha ottenuto diritti sulla proprietà intellettuale sui modelli di OpenAI, consentendone l'integrazione nei propri

prodotti, come Copilot, e l'offerta di servizi basati su questi modelli attraverso l'Azure OpenAI Service. L'accordo inoltre prevedeva una condivisione dei profitti bilaterale, in modo che entrambe le aziende potessero trarre vantaggio dall'utilizzo più diffuso dei modelli di OpenAI. Questa collaborazione ha portato vantaggi significativi a entrambe le aziende: per OpenAI, l'accesso alle vaste risorse computazionali di Microsoft ha facilitato l'addestramento di modelli sempre più avanzati, per Microsoft, invece, l'integrazione delle tecnologie di OpenAI ha arricchito la propria offerta di prodotti e servizi con funzionalità di IA all'avanguardia, rafforzando la sua posizione nel mercato del *cloud computing* e dell'intelligenza artificiale.

Tuttavia, a gennaio 2025 i due colossi hanno annunciato una revisione dell'accordo[c] che ne lascia invariata la struttura portante, ma che offre ad OpenAI più flessibilità. In particolare, a seguito del progetto Stargate, OpenAI potrà accedere a risorse di calcolo anche da fornitori di *cloud* diversi da Microsoft, tra questi figura Oracle, una delle società coinvolte nel progetto Stargate.

[a] OpenAI, *Microsoft invests in and partners with Ope nAI to support us building beneficial AGI*, OpenAI, 22 July 2019, https://openai.com/index/microsoft-invests-in-and-partners-with-openai/

[b] OpenAI, *OpenAI and Microsoft extend partnership*, OpenAI, 23 January 2023, https://openai.com/index/openai-and-microsoft-extend-partnership/.

[c] M.C. Blogs, *Microsoft and OpenAI evolve partnership to drive the next phase of AI - The Official Microsoft Blog*, «The Official Microsoft Blog», 21 January 2025, https://blogs.microsoft.com/blog/2025/01/21/microsoft-and-openai-evolve-partnership-to-drive-the-next-phase-of-ai/.

Questa tipologia di accordi potrebbe portare le aziende a adottare comportamenti anticoncorrenziali. Questi accordi, infatti, hanno sollevato preoccupazioni di diverse autorità antitrust, tra cui la *FTC* (*Federal Trade Commision*). Nel report *Partnerships Between Cloud Service Providers and AI Developers* di gennaio 2025, la FTC identifica diverse aree di attenzione che potrebbero avere implicazioni sul mercato:

1. *Controllo sulle risorse critiche*: i *Content Service Provider* (*CSP*) investono massicciamente in aziende di IA, ma una parte significativa di tali investimenti viene vincolata all'utilizzo esclusivo dei propri servizi *cloud*, creando un sistema di reinvestimento chiuso. Questo significa che aziende come OpenAI, Anthropic e altre

devono utilizzare le infrastrutture di Microsoft Azure, Amazon Web Services o Google Cloud per l'addestramento e la distribuzione dei propri modelli, riducendo la possibilità di diversificazione e consolidando il dominio dei CSP. Inoltre, gli sviluppatori beneficiano di tariffe agevolate per l'accesso alle risorse computazionali, mentre i CSP ottengono un accesso privilegiato alla proprietà intellettuale e ai dati sulle prestazioni dei modelli, consentendo loro di migliorare le proprie soluzioni e consolidare la loro posizione di mercato.

2. *Aumento dei costi di switching*: le *partnership* includono vincoli contrattuali ed esclusività, rendendo oneroso per le aziende di IA cambiare fornitore di *cloud computing*. Oltre agli ostacoli legali e finanziari, esistono anche barriere tecniche, poiché diversi CSP potrebbero implementare soluzioni proprietarie che complicano la migrazione dei modelli da una piattaforma all'altra. Questa situazione rafforza la dipendenza degli sviluppatori dai loro investitori *cloud*, creando un effetto di *lock-in*.

3. *Vantaggio informativo dei CSP*: le aziende di *cloud* ottengono accesso privilegiato a dati sensibili, tra cui metodi di sviluppo dei modelli, prestazioni finanziarie e requisiti infrastrutturali dei partner. Questo determina asimmetrie informative, permettendo ai CSP di sviluppare modelli proprietari con un vantaggio competitivo sleale rispetto ad altri concorrenti nel settore.

4. *Rischio di integrazione verticale*: le *partnership* permettono ai CSP di integrare i modelli di IA nei propri prodotti e servizi, come nel caso di Microsoft che utilizza OpenAI per potenziare le funzionalità di Copilot in Office e Teams, o Google che integra l'IAG in Gmail e in Google Docs. Questa tendenza solleva preoccupazioni per il mercato, poiché i CSP possono favorire i propri modelli a discapito di altri fornitori, riducendo la pluralità e l'innovazione.

L'impatto delle *partnership* tra i principali fornitori di servizi *cloud* e gli sviluppatori di intelligenza artificiale potrebbe determinare un progressivo consolidamento del mercato, con il rischio di una tendenza alla monopolizzazione. Queste dinamiche potrebbero ostacolare l'ingresso di nuovi attori, rafforzando il dominio di un numero ristretto di aziende tecnologiche. Inoltre, la stretta collaborazione tra CSP e sviluppatori potrebbe generare un significativo squilibrio competitivo a livello globale, limitando le opportunità per ecosistemi alternativi, come quello europeo, che già oggi incontrano difficoltà nel competere con i colossi americani. Se da un lato tali accordi favoriscono lo sviluppo tecnologico e l'efficienza operativa, dall'altro lato, potrebbero ridurre la diversità dell'offerta e rallentare

l'emergere di modelli innovativi indipendenti, con implicazioni rilevanti per la competizione e la regolamentazione del settore.

Per quanto riguarda l'integrazione a valle, l'analisi condotta dalla Competition and Market Authority (CMA) nel report del 2023 dedicato ai *foundation models* e successivamente aggiornata nell'update paper del 2024 evidenzia come tale integrazione non si limiti alla semplice distribuzione dei modelli attraverso interfacce proprietarie, ma coinvolga la capacità delle grandi piattaforme digitali di sfruttare i propri ecosistemi esistenti per consolidare ulteriormente il proprio vantaggio competitivo. Secondo la CMA, «*the growing presence across the FM value chain of a small number of incumbent technology firms, which already hold positions of market power in many of today's most important digital markets, could profoundly shape FM-related markets to the detriment of fair, open and effective competition, ultimately harming businesses and consumers, for example by reducing choice and quality, and by raising prices*». In linea con questa prospettiva, anche l'OECD nel paper *Artificial Intelligence, Data and Competition* pubblicato nel 2024 sottolinea i rischi derivanti dall'asimmetria di potere economico e infrastrutturale, osservando che «*if a firm has substantial market power, they may be able to leverage this to foreclose rivals in downstream markets*».

Queste preoccupazioni nascono dalla capacità delle aziende con una base utenti consolidata in mercati digitali adiacenti, come sistemi operativi, suite di produttività o motori di ricerca, di integrare le proprie soluzioni di intelligenza artificiale direttamente nei loro servizi, rendendo più complessa la competizione per gli sviluppatori indipendenti. Esempi di questo fenomeno fanno riferimento alle aziende GAMMA (Google, Amazon, Microsoft, Meta e Apple), che, come evidenziato dalla Fig. 7 agiscono lungo tutta la catena del valore dell'IAG.

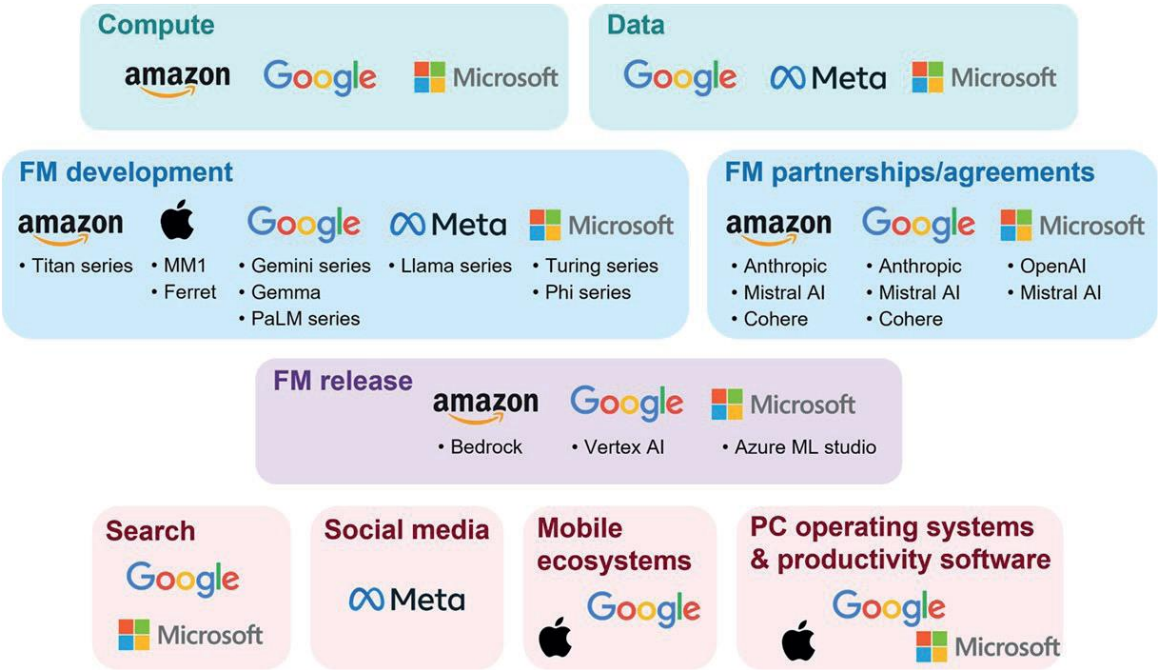

Fig. 7: Presenza delle aziende GAMMA lungo la catena del valore dell'IAG[16].

Tutte queste società hanno integrato i loro modelli all'interno dei propri servizi, ad esempio Microsoft, attraverso la *partnership* con OpenAI, ha incorporato modelli di intelligenza artificiale nel proprio software di produttività (Copilot in Office), nel sistema operativo Windows e nel motore di ricerca Bing. Google segue una strategia simile, implementando i propri FM nella *Search Generative Experience*. Queste aziende, oltre ad avere il vantaggio di poter utilizzare per l'addestramento i dati generati dai propri servizi, come Meta con i dati di Facebook e Instagram o Google con le trascrizioni dei video di YouTube, posseggono l'ulteriore vantaggio legato alla possibilità di utilizzare i dati generati dagli utenti durante l'utilizzo del modello per migliorare il modello stesso, innescando dei *data feedback loops*[17].

Box 7: *Data feedback loops*

[16] *AI Foundation Models Update paper*, CMA (2024).

[17] Hagiu e Wright (2025), "*Artificial intelligence and competition policy",* International Journal of Industrial Organization.

I dati generati durante l'utilizzo di un modello, noti come *real-time data*, rappresentano una risorsa cruciale per il miglioramento continuo dello stesso. Questi dati, raccolti da interazioni dirette con gli utenti, permettono di identificare debolezze, *bias* o comportamenti non ottimali, fornendo una base per migliorare l'accuratezza, la robustezza e la personalizzazione del modello. Questo processo può innescare dei *data feedback loops*, circuiti auto-rinforzanti in cui i dati raccolti migliorano il modello, attirando più utenti e generando ulteriori dati.

I *data feedback loops* hanno un impatto variabile sull'efficacia e sul miglioramento di un modello di IA, in funzione sia del *task* assegnato al modello che del dominio di applicazione. L'utilità di questi dati dipende dalla capacità del modello di raccogliere input significativi dall'utente e di integrarli nel proprio processo di apprendimento. Ad esempio, il dominio di ChatGPT nel mercato dei modelli conversazionali può essere attribuito soprattutto al processo di *feedback loop*. Ogni interazione dell'utente con il modello fornisce dati preziosi che possono essere utilizzati per ottimizzare la qualità delle risposte: quando un utente riformula una domanda o indica che una risposta non è corretta, questi dati possono essere integrati nei successivi cicli di apprendimento per migliorare l'accuratezza del modello. Differente è invece il caso di un modello che analizza immagini radiologiche per rilevare anomalie e che opera su dati preesistenti e fornisce un output diretto. In questo caso, l'utente raramente fornisce un *feedback* immediato o strutturato che possa alimentare un ciclo di miglioramento continuo.

Questi aspetti evidenziano che non esiste un approccio univoco per sfruttare i dati di *feedback*: la loro utilità è determinata dall'interazione tra la tipologia del dato, che deriva dal dominio applicativo (ad esempio medico o finanziario) e il *task* del modello. Nei contesti in cui il *feedback* è particolarmente utile, come nei modelli conversazionali, i *data feedback loops* possono creare significative barriere all'ingresso. Le aziende leader, grazie alla loro ampia base di utenti (come OpenAI), possono raccogliere volumi consistenti di dati in tempo reale, migliorando continuamente i propri modelli e consolidando la loro posizione di mercato, creando una fidelizzazione per il cliente e un effetto di *lock-in*. Questo ciclo rende difficile per nuovi concorrenti competere, poiché l'accesso a dati equivalenti o di qualità comparabile è spesso limitato, rafforzando ulteriormente il vantaggio competitivo degli attori già presenti.

Questa dinamica di integrazione dei modelli a valle, sebbene possa generare efficienze e miglioramenti per gli utenti, potrebbe quindi produrre effetti negativi sulla concorrenza. Questo perché, seppur vero che il mercato delle applicazioni a valle basate sui modelli di IAG sia dinamico e ricco di concorrenti, i vantaggi che i *player* integrati lungo tutta la catena posseggono rispetto a coloro che si limitano allo sviluppo di applicazioni basate su FM preesistenti potrebbe portarli ad attuare comportamenti anticoncorrenziali che soffocherebbero la concorrenza a valle. Ad esempio:

1. *bundling*: si verifica quando un'azienda vende due o più prodotti insieme come un pacchetto, spesso a un prezzo inferiore rispetto all'acquisto separato. Se il *bundling* avvantaggia in modo sleale un prodotto rispetto a quello di un concorrente (es. Microsoft che include gratuitamente Copilot in Office, rendendo più difficile per altri sviluppatori competere), può limitare la concorrenza;
2. *tying*: si ha quando un'azienda impone ai clienti l'acquisto di un prodotto per poter accedere a un altro. Ad esempio, se un'azienda richiedesse l'uso del proprio FM per accedere a determinati servizi *cloud*, potrebbe impedire ai concorrenti di entrare nel mercato;
3. *self-preferencing*: accade quando una piattaforma digitale o un'azienda favorisce i propri prodotti rispetto a quelli dei concorrenti. Un esempio potrebbe essere Google che mostra il proprio FM prima di quelli di altri sviluppatori nei risultati di ricerca, limitando la visibilità di alternative competitive.

Queste possibili pratiche hanno attivato le autorità regolatorie ed antitrust mondiali, le quali hanno avviato indagini sulle grandi aziende integrate verticalmente, al fine di evitare una distorsione del mercato da parte delle imprese consolidate per permettere agli utenti di accedere liberamente ad una diversità di modelli senza il verificarsi dell'effetto di *lock-in*. In Fig. 8, secondo una recente analisi dell'Antitrust inglese, viene rappresentato il possibile *feedback loop* che le grandi aziende potrebbero innescare col fine di aumentare la loro posizione di mercato a discapito dei concorrenti:

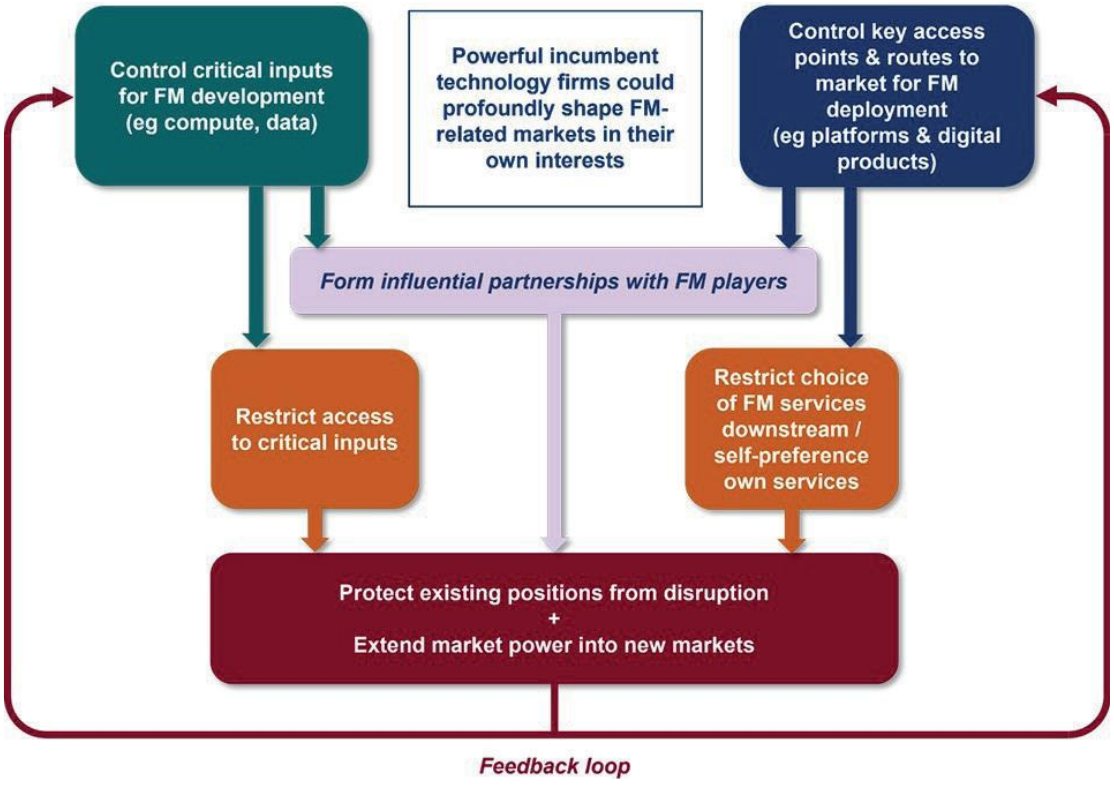

Fig. 8: Possibile *feedback loop* che le aziende leader potrebbero innescare per favorire i loro interessi[18].

3. *Investimenti e sviluppo dei modelli in USA, Europa e Cina*

Come è emerso dai dati presentati, i principali *players* all'interno di questo mercato provengono principalmente dagli Stati Uniti e dalla Cina, con concorrenti europei che faticano a penetrare nel mercato. Questa tendenza è visibile anche a livello del numero di modelli rilasciati dalle diverse nazioni[19]; infatti, la Fig. 9 mostra chiaramente una dominanza da parte degli USA sul numero di modelli rilasciati (109 nel 2023 e 65 nel 2024), rispecchiando il suo dominio a livello globale. La tendenza interessante è quella relativa al 2024, con una diminuzione del numero di modelli rilasciati dagli Stati Uniti e dalla Cina, il che suggerisce, come precedentemente accennato, a un consolidamento dei modelli già rilasciati negli anni passati. L'Europa e il Regno Unito hanno invece una tendenza opposta che vede il numero dei modelli

[18] *AI Foundation Models Update paper*, CMA (2024).

[19] In questa analisi, un modello viene associato ad una nazione se la sede legale dello sviluppatore ha sede in quella nazione. Inoltre, vengono esclusi i modelli rilasciati dalle collaborazioni internazionali.

rilasciati aumentare da 16 nel 2023 a 28 nel 2024, da cui si intuisce una partecipazione più attiva da parte di queste ultime.

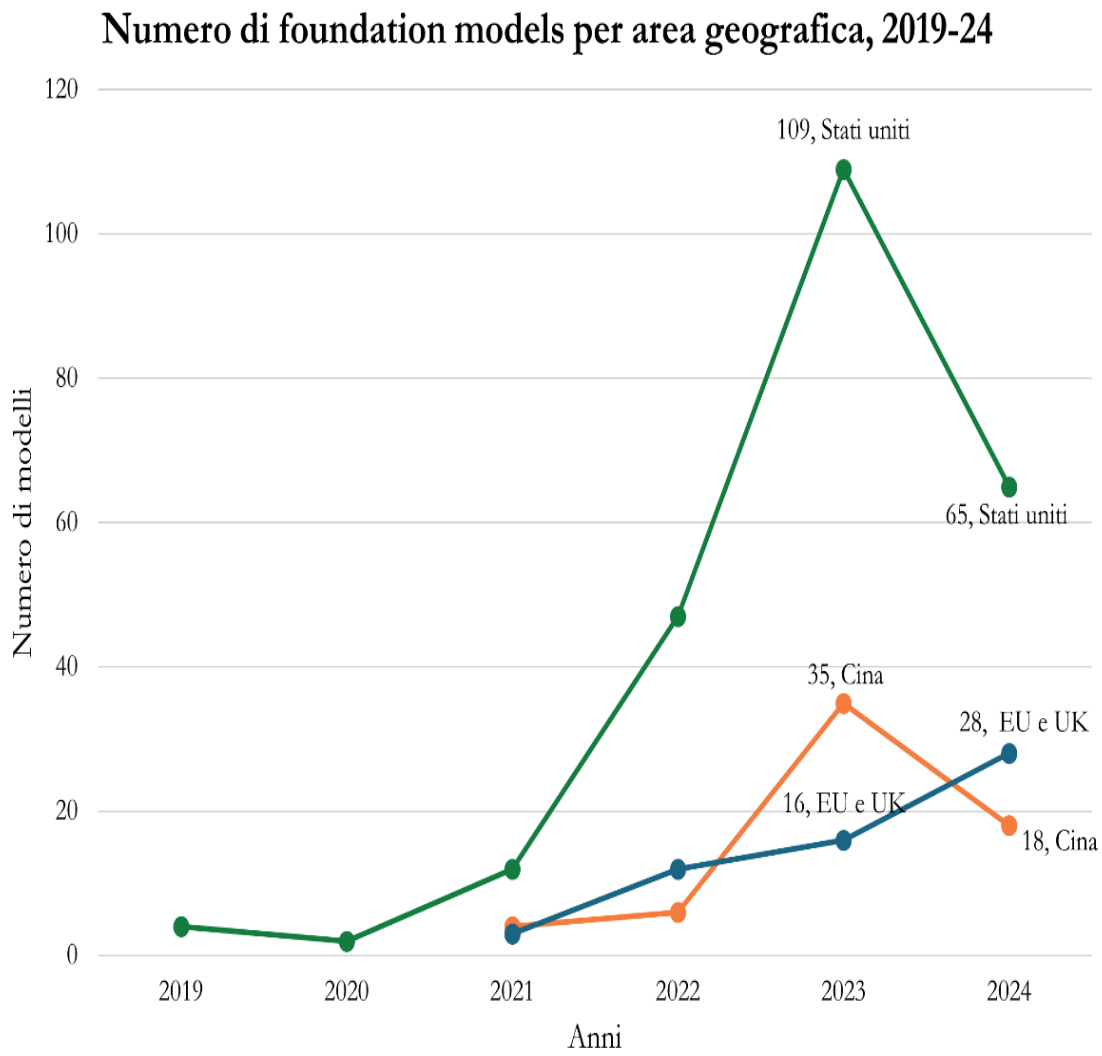

Fig. 9: FMs rilasciati da US, UE, UK e Cina dal 2019 al 2024[20].

Un elemento chiave di questa leadership è rappresentato dal volume degli investimenti privati (Fig. 10). Nel 2024, gli Stati Uniti hanno investito 29,04 miliardi di dollari nel settore dell'IAG, una cifra significativamente superiore rispetto a quella di Europa e Regno Unito (1,49 miliardi di dollari) e Cina (2,11 miliardi di dollari). Questa disparità suggerisce che la capacità statunitense di attrarre capitali privati sia un fattore determinante per il rapido sviluppo del settore. Gli elevati investimenti consentono infatti di finanziare infrastrutture computazionali avanzate, l'acquisizione

[20] *Ecosystem graphs for foundation models,* CRFM Stanford.

di talenti altamente specializzati e progetti di ricerca su larga scala, creando un vantaggio competitivo significativo.

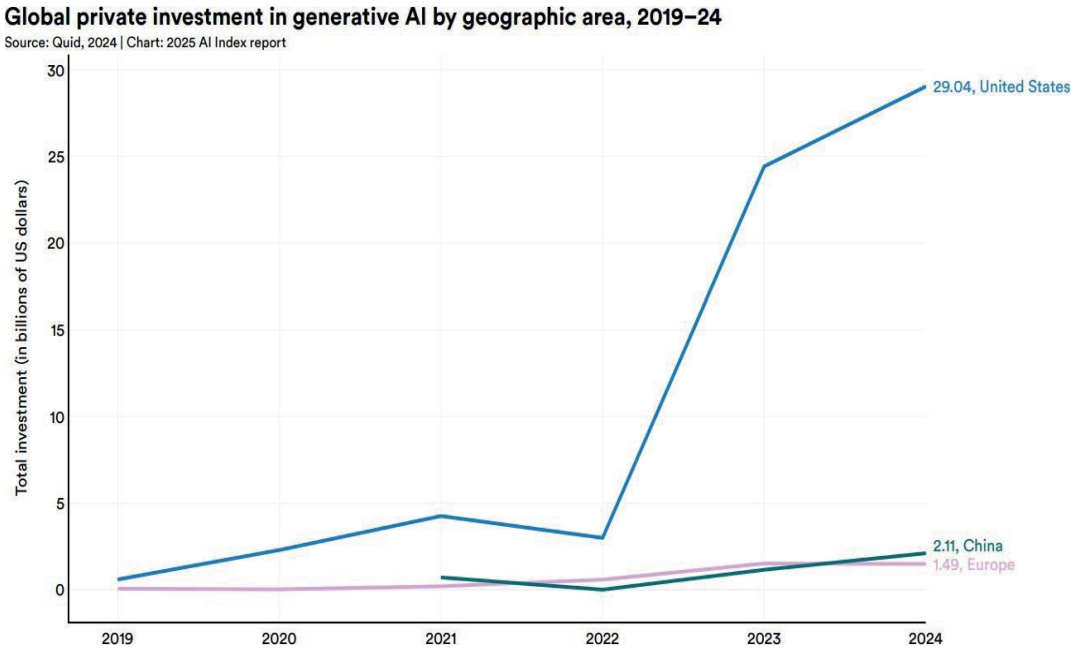

Fig. 10: Investimenti privati in IAG negli USA, UE, UK e Cina nel, 2019-24[21].

Tuttavia, tali valori non restituiscono un quadro completo della competizione globale.

In primo luogo, i dati sugli investimenti cinesi risultano spesso parziali e poco trasparenti, poiché una quota rilevante dei finanziamenti proviene da programmi statali o fondi a partecipazione pubblica, difficili da monitorare attraverso le metriche internazionali di *venture capital*. Parallelamente, i fondi di *venture capital* a partecipazione pubblica hanno mobilitato, nell'ultimo decennio, centinaia di miliardi di dollari a sostegno di settori strategici, fra cui l'IA. In secondo luogo, alcuni indicatori alternativi mostrano come la Cina abbia ormai raggiunto, e in alcuni ambiti superato, gli Stati Uniti sul piano scientifico e tecnologico. Alla conferenza AAAI 2026, una delle più prestigiose al mondo per l'intelligenza artificiale, quasi 20.000 dei circa 29.000 articoli sottomessi provenivano da ricercatori cinesi. Rispetto a dieci anni fa, quando la presenza cinese era marginale, si tratta di una crescita impressionante, che conferma il dinamismo della comunità scientifica nazionale. Inoltre, laboratori come

[21] *Artificial Intelligence Index Report*, Stanford University (2025).

DeepSeek, Alibaba e Baichuan rilasciano modelli *frontier* sempre più sofisticati, segnalando un'accelerazione nella capacità di ricerca e sviluppo.

Invece, un fattore determinante che potrebbe aver limitato gli investimenti europei nell'IAG è l'approccio regolatorio adottato dall'Unione Europea rispetto ad altre aree geografiche, in particolare agli Stati Uniti. L'UE ha posto un forte accento su sicurezza, trasparenza e protezione dei dati, imponendo vincoli più severi per lo sviluppo e l'implementazione di modelli di intelligenza artificiale. Un esempio chiave è l'*AI Act*, la prima normativa organica sull'IA a livello mondiale, che introduce obblighi specifici per i modelli, incluse restrizioni sull'uso dei dati, requisiti di interpretabilità e *audit* periodici. Sebbene queste misure abbiano l'obiettivo di mitigare i rischi etici e garantire un utilizzo responsabile dell'IA, esse comportano un incremento significativo dei costi di conformità per le aziende europee, e possono quindi rallentare l'adozione e lo sviluppo di nuove tecnologie rispetto alle controparti statunitensi.

L'effetto complessivo di questa regolamentazione più incisiva è il rischio di una minore attrattività del mercato europeo per gli investitori privati, che probabilmente preferiscono destinare i propri capitali a contesti più favorevoli alla crescita dell'IA. Se da un lato l'UE punta a costruire un ecosistema di intelligenza artificiale etico e sicuro, dall'altro lato, è possibile che questa complessa normativa finisca con il limitare l'innovazione e la competitività delle aziende europee nel panorama globale dell'IAG.

Tuttavia, l'*AI Action Summit*, svoltosi a Parigi il 10 e 11 febbraio 2025, potrebbe aver segnato una svolta significativa per lo sviluppo di modelli europei su scala globale. Infatti, l'Unione Europea ha annunciato due piani, per un totale di 200 miliardi di euro tra fondi pubblici e privati, per supportare la crescita dell'intelligenza artificiale in Europa:

1. *European AI Champion*[22]: una coalizione pubblico-privata guidata dal settore industriale che mira a rafforzare la competitività europea nel campo dell'intelligenza artificiale. Oltre sessanta imprese europee, tra cui grandi gruppi industriali e tecnologici come Airbus, Siemens, ASML, Volkswagen e startup emergenti come Mistral AI hanno aderito a una piattaforma di cooperazione volta a sviluppare e a diffondere soluzioni di IA nei settori strategici dell'economia europea. L'iniziativa punta a promuovere un ecosistema IA aperto, competitivo e affidabile, intervenendo in particolare su quattro aree chiave: il rafforzamento delle infrastrutture di calcolo, l'attrazione e formazione di talenti, l'adozione industriale dell'IA e la semplificazione

[22] *Launching the EU AI Champions Initiative' to unlock Europe's full potential in AI* (2025, February 10), https://www.generalcatalyst.com/stories/euaici.

del quadro regolatorio. I membri della coalizione hanno elaborato un'agenda strategica condivisa l'*Ambitious Agenda for European AI* che prevede azioni coordinate con le istituzioni europee per favorire la crescita del settore, anche attraverso forme di co-regolazione. Sebbene l'iniziativa non sia finanziata direttamente dalla Commissione, i partecipanti hanno annunciato un impegno complessivo di investimento privato di circa 150 miliardi di euro, da affiancare a fondi pubblici europei provenienti da altri strumenti, in particolare dal programma InvestAI.

2. *InvestAI*[23]: in aggiunta ai 150 miliardi di euro stanziati da azienda private, l'UE mira a mobilitare 50 miliardi di euro con l'obiettivo di colmare il divario tecnologico con le grandi potenze mondiali, dotando l'UE di infrastrutture computazionali avanzate. Infatti, dei 50 miliardi di euro, 20 miliardi saranno destinati alla costruzione di cinque Gigafactories[24], ossia strutture su larga scala dedicate allo sviluppo e all'addestramento di modelli di intelligenza artificiale di nuova generazione, con oltre 100.000 *chip* acceleratori.

Inoltre, la Francia, che si è dimostrata essere il principale *player* europeo nel mercato dell'IAG globale grazie alla startup Mistral, conferma la sua posizione annunciando un piano nazionale di 109 miliardi di euro finanziato sia da investitori francesi che esteri con il principale obiettivo di costruire *data center* finalizzati allo sviluppo di modelli. Nonostante il coinvolgimento limitato dell'Italia in queste iniziative, sono numerosi i progetti realizzati legati all'IAG (Tab. 1), tra i più rilevanti vi sono:

Tab. 1: Modelli di IAG italiani.

Sviluppatore	**Modello**	**Descrizione**
iGenius	Colosseum 355B	Un LLM progettato specificamente per operare i settori altamente regolamentati e nella pubblica amministrazione. Rispetta le rigorose normative sulla protezione dei

[23] *EU launches InvestAI initiative to mobilise 200 billion* (2025, February 11), European Commission, https://ec.europa.eu/commission/presscorner/detail/en/ip25467.

[24] AI factories (2025, June 19), *Shaping Europe's Digital Future*, https://digital-strategy.ec.europa.eu/en/policies/ai-factories.

		dati, offrendo alle aziende anche la possibilità di ospitare il modello all'interno delle proprie infrastrutture, garantendo il pieno controllo sui dati sensibili
Almawave	Velvet 14B e Velvet 2B	Modelli addestrati tramite il supercomputer Leonardo e rilasciati in modalità *open source* che rispettano tutto il quadro regolatorio europeo garantendo efficienza e bassi consumi energetici
ASC 27	Vitruvian-1	Un LLM addestrato con 8 GPU H100 di NVIDIA con un costo di addestramento dell'ordine delle decine di migliaia di euro. Progettato prevalentemente per compiti logici, ha ottenuto un punteggio di 94 sul *benchmark* MATH-500, dimostrandosi competitivo con i modelli di punta dell'aziende statunitensi e cinesi
Engineering	EngGPT	Un LLM proprietario sviluppato per le aziende e la pubblica amministrazione tramite tecniche che garantiscono la trasparenza e la sicurezza dei dati in conformità alle linee guida dell'*AI Act*

L'emergere di DeepSeek ha quindi influenzato i mercati globali, spingendo sia Europa che Stati Uniti a intensificare gli investimenti nel settore. Tuttavia, se da un lato l'Europa ha stanziato 200 miliardi di euro tramite i piani *European AI* e *Invest AI*, negli Stati Uniti, il 21 gennaio 2025, è stato annunciato il progetto *Stargate* che mira a consolidare la leadership americana nel settore: una *joint venture* tra OpenAI, SoftBank, Oracle e MGX che investiranno 500 miliardi di dollari entro il 2029 per la costruzione di infrastrutture di calcolo per l'intelligenza artificiale. Inoltre, il progetto vede la partecipazione anche di NVIDIA, Microsoft e Arm come partner tecnici. Questa disparità di investimenti, con un volume di investimento americano superiore

di circa 300 miliardi di dollari rispetto a quello europeo, potrebbe riflettere strategie diametralmente opposte. Mentre l'Europa si focalizza sulla regolamentazione e sulla creazione di un ecosistema di IA etico e conforme al quadro normativo, gli Stati uniti puntano sulla creazione di infrastrutture computazionali massicce. Un esempio concreto è rappresentato da Grok 3, la nuova famiglia di modelli sviluppata dalla startup xAI fondata da Elon Musk. Grok 3 è stato addestrato utilizzando una delle più grandi infrastrutture di calcolo mai realizzate sfruttando oltre 200.000 GPU.

La questione chiave per il futuro sarà comprendere se, attraverso questi piani di investimento e una regolamentazione armonizzata, l'Europa riuscirà a ridurre il gap tecnologico con gli Stati Uniti o se, al contrario, il divario continuerà ad ampliarsi.