POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Gestionale

A.A. 2024/2025

Sessione di Laurea Marzo 2025

Concorrenza nel mercato dei Foundation Models

Dinamiche competitive & Policy recommendations

Relatore accademico:

Dottor Carlo Cambini, PhD

Candidato:

Marchetto Alessia

Abstract

Dal 2023 l'intelligenza artificiale è diventata una fedele compagna quotidiana nella vita di milioni di persone e aziende. Poche sono le *task* che non riesce a completare ed infinite le possibili applicazioni. L'entusiasmo di professionisti e *users* e le incredibili *performance* raggiunte hanno attirato l'attenzione di grandi investitori e colossi *tech* che oggi si sfidano nel mercato per assicurarsi più quote possibili di *share* e, di conseguenza, alti ritorni.

La filiera dell'Ai è articolata in più *layers*: risorse computazionali, dati e talenti come principali input a monte per il secondo livello formato dai *foundation models*, modelli generali più grandi e capaci che vengono, a valle, affinati nei singoli applicativi.

La presente tesi mira ad indagare le dinamiche della competizione nel mercato dei modelli fondativi non potendo però prescindere dall'analisi di tutte le interdipendenze tra i livelli della *value chain.* In particolare si concentra sull'analisi delle ripercussioni a mercato degli effetti di rete dei dati e dei *data feedback loop*, attraverso il commento di modelli economici realizzati per la descrizione di prodotti o sistemi *data enabled learning.* Contrariamente a quanto comunemente si pensa, in un modello di intelligenza artificiale generativa gli *user data* raccolti in fase di *deployment* non costituiscono l'elemento cardine del vantaggio competitivo a lungo termine di una azienda *incumbent.*

Nonostante le caratteristiche tecniche e strutturali del mercato, in particolare le forti economie di scala dell' infrastruttura *cloud* e *hardware,* indichino la presenza di una minaccia oligopolistica, l'attuale scenario è altamente dinamico e competitivo con le *big tech* americane che sfidano numerose giovani *startup*.

Da qui l'esigenza per i *policymakers* di tutto il mondo di rallentare la corsa alla regolamentazione e comprendere a fondo le dinamiche di mercato dell'intera filiera in modo da evitare una *over-regulation* che frenerebbe la crescita*,* e intervenire a garanzia dell'attuale dinamicità, dell'innovazione, e ovviamente, dei consumatori.

INDICE

INDICE DELLE FIGURE

Capitolo 1: Introduzione

Nei prossimi decenni l'intelligenza artificiale "sarà in grado di fare quasi tutto, incluse nuove scoperte scientifiche che potrebbero espandere la nostra idea di "tutto"". Così esordì Sam Altman, CEO di OpenAi in un post del 2021[1] ed oggi, qualche anno dopo la sua previsione non sembra discostarsi molto dalla realtà. Come vapore, motori, aerei, antibiotici, semiconduttori e internet nei secoli precedenti, l'intelligenza artificiale generativa ha il potenziale per essere la prossima grande tecnologia *disruptive* in grado di cambiare completamente le abitudini di vita di milioni, miliardi di persone. L'interesse in questa tecnologia è esploso a seguito del rilascio di OpenAI dei primi *large language model* per il testo (Chat GPT) e per la generazione di immagini (DALL-E) nella seconda metà del 2022. Da allora il mercato ha visto una proliferazione di applicazioni di ogni genere, da quelle più mondane per la quotidianità o di *business* per il lavoro, ad applicazioni artistiche, di *coding*, o per uso di ricerca o medico. Uno studio condotto da ricercatori di OpenAi ha affermato che in pochi anni circa il 49% delle persone affiderà all'intelligenza artificiale oltre la metà dei propri *task* quotidiani personali e professionali, diventando una fedele compagna di vita (Eloundou, Manning, Mishkin, & Rock, 2023).

Se i modelli AI svolgeranno davvero un ruolo così importante nella nostra economia, allora la struttura del mercato in cui vengono offerti avrà implicazioni di primo ordine per il benessere sociale. I *policymakers*, comprese le autorità antitrust, devono quindi prestare particolare attenzione a questo mercato, al fine di evitare una prematura *o*ver regolamentazione che potrebbe limitarne l'innovazione, e, al contempo, garantire un ambiente competitivo e sicuro per aziende e cittadini.

Rispetto ai precedenti algoritmi di *machine learning,* l'AI generativa ha introdotto i *foundaton model* (*FM*), ossia modelli *general purpose* pre-addestrati su larga scala, progettati per essere successivamente affinati con i dati e le informazioni necessarie all'implementazione su una specifica applicazione. I *foundation models*, per i loro alti costi di produzione e il legame con le *big tech* americane, per la fornitura *cloud* e

[1] https://moores.samaltman.com/

hardware, hanno da subito attirato l'attenzione di *policymaker* e autorità, preoccupate per potenziali rischi di concentrazione.

Questa tesi mira ad essere un supporto alla comprensione del mercato e delle dinamiche competitive dei *FM* e dell'intera *value chain* in cui sono inseriti per rispondere alla "*trillion-dollar question"* che persegue *policymakers* e investitori: "I *Foundation Model* diventeranno una *commodity* standardizzata o saranno dominati da uno o due attori?" (Hagiu & Wright, 2024)

Nella prima parte il testo descrive la struttura tecnica del mercato introducendo le interdipendenze nella *value chain*, per poi concentrarsi sull'analisi delle dinamiche competitive nei mercati dei principali *input*, o *asset* complementari: talenti, dati e risorse computazionali. In riferimento alle teorie di Teece (1986) si cercherà di capire quale tra *appropriability* e *complementary assets* sia, quindi, la migliore fonte di vantaggio competitivo e potere di mercato e come le aziende ad oggi presenti stiano interagendo di conseguenza.

Il terzo capitolo approfondisce la questione dati ed effetti di rete, analizzando come e se i *data network effect* o i così detti *data feedback loop* costituiscano, in questo specifico settore, una fonte di vantaggio competitivo stabile che possa creare alte barriere all'ingresso e favorire la concentrazione. A tal proposito verrà esaminato un modello di competizione dinamica tra due aziende caratterizzate da un apprendimento *data-enabled,* i cui *insight* verranno interpretati alla luce delle dinamiche del mercato di riferimento.

L'ultimo capitolo propone, infine, un approfondimento in ambito *policy* con alcune proposte e considerazioni finali.

Capitolo 2: Struttura del mercato e concorrenza

2.1 Introduzione e *value chain*

L'intelligenza artificiale, un tempo relegata al mondo della fantascienza, è oggi una realtà che permea profondamente la nostra società e impatta significativamente molti aspetti della nostra quotidianità.

Nonostante i rapidi e recenti sviluppi degli ultimi tre anni, la storia di questa tecnologia ha inizio già negli anni '30 del ventesimo secolo grazie al contributo scientifico di Alan Turing. Egli non solo definì i primi modelli matematici alla base dei successivi computer ma anticipò anche molte delle sfide ed opportunità legate allo sviluppo dell'AI nel suo celebre testo "*Computing machinery and intelligence"*, in cui propose il test che porta il suo nome. Secondo Turing una macchina si poteva ritenere intelligente se fosse in grado di ingannare una persona, facendogli credere di essere un umano a sua volta. Questa i stimolò il dibattito scientifico e stabilì un obiettivo concreto per ricercatori e scienziati del settore, i più illustri dei quali si riunirono nel 1956 alla conferenza di Dartmouth, considerata il battesimo ufficiale dell'intelligenza artificiale come campo di studio. Negli anni successivi gli importanti risultati ottenuti nello sviluppo delle reti neurali (aventi radici nel 1943 nella proposta di W.S. McCulloch e W. Pitts di un "neurone artificiale" in grado di risolvere semplici funzioni booleane), e nella logica matematica per la dimostrazione di teoremi e l'inferenza di nuova conoscenza, dimostrarono che le speranze della comunità scientifica erano, di fatto, fondate.

Tuttavia, questo entusiasmo iniziale si dovette scontrare con importanti limiti tecnologici, primo fra tutti la scarsa potenza di calcolo allora disponibile, portando ad un periodo di stagnazione e riduzione dei finanziamenti.

Negli anni '80 l'avvento dei personal computer e l'aumento della capacità computazionale diedero nuovo slancio alla ricerca portando alla nascita del *machine learning,* branca dell'intelligenza artificiale che si concentra sulla progettazione di algoritmi e modelli statistici capaci di apprendere dai dati e migliorare in autonomia le proprie prestazioni nel tempo. L'entusiasmo dovuto ai sorprendenti risultati ottenuti nel riconoscimento vocale e di immagini fu il *driver* principale degli ingenti investimenti in ricerca che si concretizzarono nella rivoluzione del *deep learning*. Le reti neurali profonde di cui fanno uso gli algoritmi di *deep learning,* sono composte da più strati di

neuroni artificiali e sono in grado di apprendere da enormi quantità di dati non strutturati, come immagini, testi e audio. Questa capacità ha permesso, negli anni, di sviluppare applicazioni incredibilmente sofisticate, come i veicoli autonomi, i sistemi di traduzione automatica, rilevazione delle frodi finanziarie e i chatbot conversazionali. L'introduzione di algoritmi di *backpropagation* più efficienti e il perfezionamento delle unità di elaborazione grafica (GPU) per accelerare i calcoli hanno costituito un ulteriore punto di svolta nello sviluppo delle reti neurali portando alla nascita delle *convolutional neural network* (CNN) per il riconoscimento e classificazione di immagini e le *recurrent neural network* (RNN) per la comprensione del linguaggio naturale, reti che hanno dato avvio allo sviluppo dell'attuale AI generativa.

Se il *deep learning* si specializza nella comprensione e modellizzazione algoritmica di enormi dataset, la nuova frontiera dell'AI ha la funzione di generare un output (testo, immagini, video, musica, parole, codice…) completamente nuovo partendo da un *prompt* specifico e un preallenamento già fatto su enormi *corpus* di dati di ogni tipologia.

Questi modelli, chiamati *foundation models,* generano nuovi contenuti imparando dai *pattern* presenti nel *training dataset.* Durante il processo di *training,* infatti, il modello "digerisce" una enorme quantità di dati di diversa natura spacchettandoli in "*token"* più piccoli (una parola, una lettera, un pixel, una nota...) e imparando così a predire il prossimo token data una certa sequenza (Pierre Azoulay, 2024). Iterando il processo previsionale il modello impara a modellizzare il dataset e così si conclude la fase di "addestramento". Infine, il modello può essere sottoposto ad uno o più cicli di *fine-tuning* dove viene allenato ulteriormente su set di dati specifici per un dato contesto.

Ciò che rende i modelli di GenAI diversi dai precedenti algoritmi di *machine learning* e *deep learning* è il fatto che siano *pre-trained* e *context-independent* (riescono infatti a svolgere un gran numero di task differenti senza ulteriore allenamento specifico) (Pierre Azoulay, 2024).

La fase di *deployment* consiste successivamente nell'integrazione di questi modelli in applicazioni (app, siti web, software ecc…) specializzate in un settore specifico o nello svolgimento di determinate *task* attraverso operazioni di *fine-tuning* e/o di "*plug-in*".

I modelli di fondazione più conosciuti ad oggi sono i modelli *Generative Pre-Trained Transformer* (GPT) facenti parte della più ampia classe di *Large Language Models* (LLMs), appositamente realizzati per interagire con il "linguaggio", sia esso umano o

qualsiasi altra forma di rappresentazione concettuale come linguaggio informatico, matematico o linee di codice. Sono di fatto complessi algoritmi che costruiscono la sintassi della frase (o altro output) selezionando la successiva parola (o *token* generico) che, in quel dato contesto, è stata maggiormente utilizzata in quella posizione. Si tratta quindi di calcoli di tipo statistico che avvengono attraverso la rete neurale.

Con "modello" ci si riferisce alla derivazione della legge matematica che meglio descrive i dati di input e che opera attraverso una rete neurale *multi-layer.* Una rete neurale è formata da nodi organizzati in piu *layers* adiacenti e interconnessi. Ogni nodo è sede di un calcolo di regressione lineare che, come tale, è caratterizzato da input, pesi, distorsioni (o *bias*) e un output. L'output di un nodo, superata una determinata soglia, viene trasferito ad un nodo del *layer* successivo di cui ne diviene l'input (processo *forward*), attraverso una funzione di attivazione che introduce non-linearità nel modello (IBM, s.d.). I valori numerici dei parametri (pesi e *bias*) vengono determinati iterativamente minimizzando, con un processo *backward*, le funzioni di perdita. Viene quindi minimizzata la differenza tra l'output generato dal modello e gli esempi di training. In questo modo il modello impara a ricreare output simili al training set (Autoridade da concorrencia, 2023).

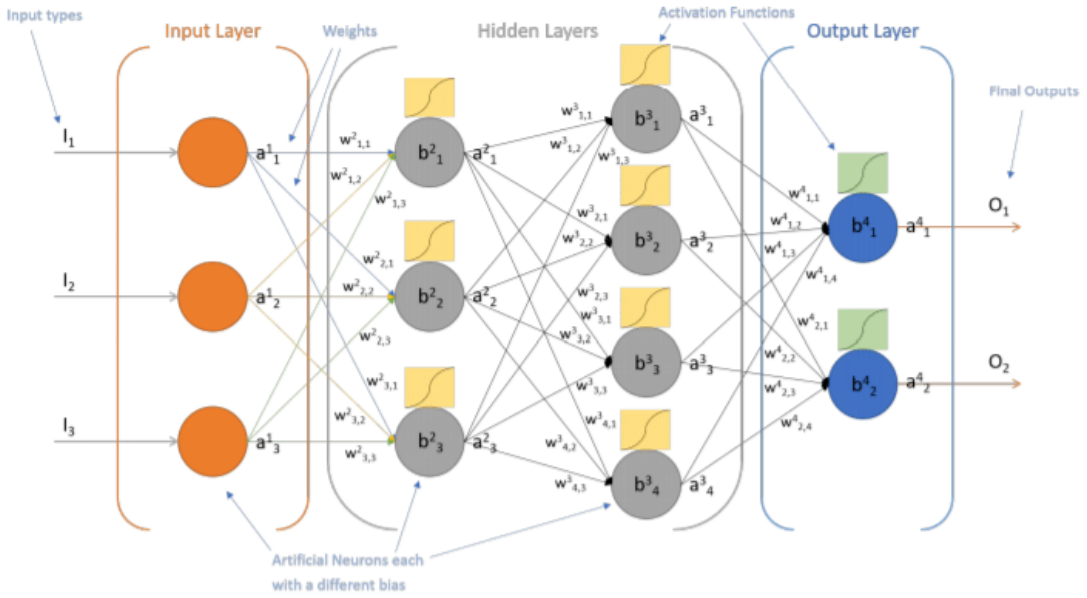


Figura 1: Descrizione grafica di una rete neurale profonda

Il numero di parametri di un modello è diventato un metodo approssimativo per valutarne la complessità in quanto modelli allenati su dataset più ampi tendono ad avere più

parametri e migliori performance (The AI Index 2024 Annual Report, 2024). L'introduzione dei modelli GPT di OpenAi è considerato, infatti, un momento *pivot* nella storia di questa tecnologia (Yiheng Liu, 2025) in quanto, rispetto alla media di allora, di circa 6-10B di parametri, GPT-3 ne possiede 175B e oggi, GPT-4, ne ha oltre un trilione (Pierre Azoulay, 2024).

Le prime versioni di questi modelli non erano in grado di processare una intera frase in un unico "*run*", non riuscendo infatti a "ricordare" i primi elementi della sequenza e rendendo l'elaborazione di lunghi testi molto imprecisa. Queste difficoltà sono state superate grazie all'introduzione dell'architettura *transformer*, creata da Google nel 2017 (Ashish Vaswani, 2017). Ciò ha introdotto un meccanismo cosiddetto "di attenzione" che permette al modello di individuare il contesto del testo focalizzandosi sulle parole chiave a cui affida un peso maggiore e migliorando così la qualità delle previsioni del token successivo nella sequenza in output. Ad oggi la maggior parte dei modelli di intelligenza artificiale utilizzati quotidianamente possiede questa struttura, di fatto diventata la "*dominant design*" (Pierre Azoulay, 2024).

Per quanto riguarda i domini non testuali come immagini, audio e video, oltre alla *transformer architecture* si sono affermati, negli ultimissimi anni, i *diffusion models*, che dopo aver modificato e "sporcato" i dati con del rumore, si allenano a riconoscerlo per ricostruire l'input originale.

Gli LLMs o *foundation models*, o alternativamente chiamati *general-purpose models* per i loro ampi domini di applicazione, per quanto riescano a svolgere un gran numero di task, sono di per sé incompleti o imprecisi e fungono da sottostante per modelli *task-specific* costruiti attraverso processi di *adaptation* e *fine tuning* progressivi. Questi processi consistono nell'ulteriore allenamento del modello con dati di settore, regole, segnali e vincoli precisi, che permettono il *deployment* del modello integrandolo in software/chatbot/siti, o altro, che lo rendono utilizzabile dal grande pubblico (CRFM, 2024) (Venkatesh Balavadhani Parthasarathy, 2024).

In particolare, per essere utilizzato da una azienda con un fine specifico, un LLM non solo deve essere sottoposto ad un processo di *fine-tuning* su dati di settore, ma deve essere ulteriormente affinato con i dati propri aziendali. Questo viene svolto attraverso i così detti *plugins* che permettono la piena personalizzazione del modello. Infine, troviamo

l'applicazione vera e propria con una determinata *user interface* (Autoridade da concorrencia, 2023).

Aspect	Pre-training	Fine-tuning
Definition	Training on a vast amount of unlabelled text data	Adapting a pre-trained model to specific tasks
Data Requirement	Extensive and diverse unlabelled text data	Smaller, task-specific labelled data
Objective	Build general linguistic knowledge	Specialise model for specific tasks
Process	Data collection, training on large dataset, predict next word/sequence	Task-specific data collection, modify last layer for task, train on new dataset, generate output based on tasks
Model Modification	Entire model trained	Last layers adapted for new task
Computational Cost	High (large dataset, complex model)	Lower (smaller dataset, fine-tuning layers)
Training Duration	Weeks to months	Days to weeks
Purpose	General language understanding	Task-specific performance improvement
Examples	GPT, LLaMA 3	Fine-tuning LLaMA 3 for summarisation

Tabella 1: Overview comparativa tra il Pre-Training e il Fine-Tuning nei LLMs. Source: Ireland's Centre for AI

Lo sviluppo dei *foundation models* è stato possibile grazie al contemporaneo sviluppo delle tecnologie di calcolo, tra cui, prime per importanza, le *graphic processing units* (GPU) che permettono di gestire migliaia di operazioni in parallelo al secondo. In generale la complessità del modello e la grandezza del dataset di allenamento influenzano direttamente la potenza computazionale richiesta. In particolare, negli ultimi cinque anni il bisogno di capacità computazionale è esponenzialmente aumentato divenendo sempre più un fattore critico dal punto di vista economico e di *carbon footprint*.

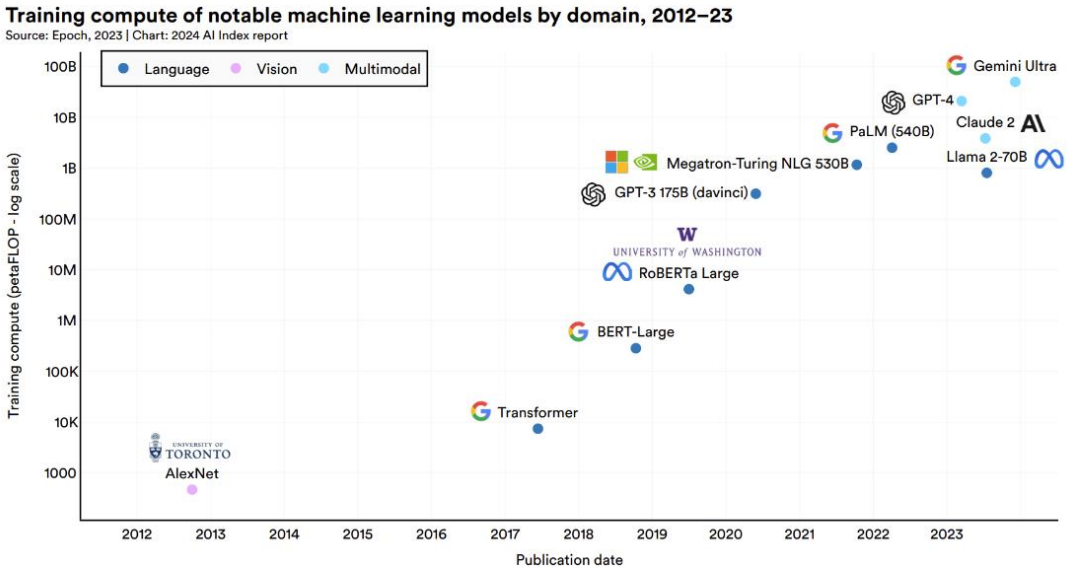


Figura 2: Training compute of notable machine learning models by sector, 2012–23. FLOP stands for "floating-point operation", a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division. The number of FLOPs a processor or computer can perform per second is an indicator of its computational power. The higher the FLOP rate, the more powerful the computer is. An AI model with a higher FLOP rate reflects its requirement for more computational resources during training

Alla luce di quanto appena descritto, questa tecnologia può essere suddivisa in quattro macroaree. Il primo *layer* è l'infrastruttura *hardware* composta da migliaia di GPUs sede del calcolo vero e proprio; successivamente si ha l'ecosistema dei dati con grandi *data storage* e tutti i meccanismi di raccolta e stoccaggio; vi è poi il *layer* di sviluppo dei *foundation models* e infine il livello degli applicativi.

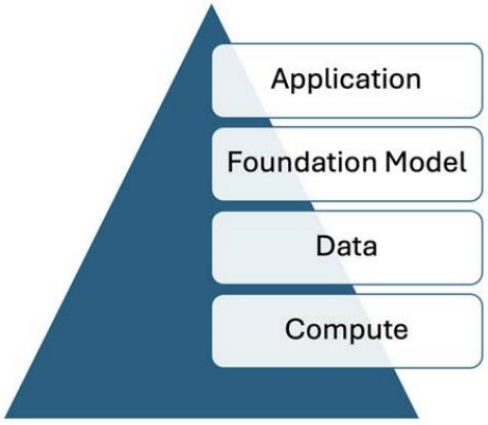


Figura 3: Il Technology Stack della GenAI, Source:

A titolo esemplificativo, GPT-4 è il modello di base mentre ChatGPT è la specifica applicazione che permette di interrogare GPT-4 attraverso una interfaccia *chatbot*.
Tuttavia, i differenti *layers* dell'AI *stack* sono strettamente interconnessi favorendo la penetrazione di una azienda da un dato *layer* ad un altro. Recenti esempi del fenomeno includono il tentativo di OpenAi di raccogliere circa $7 trilioni per sviluppare i propri *chip* (Hagey, 2024) o Nvidia che rilascia il proprio modello *open source* (Patel, 2024).

Oltre a OpenAi GPT, altri significativi esempi di *foundation models* multimodali (ossia capaci di elaborare diverse tipologie di dati come testo, audio o immagini, anche senza raggiungere prestazioni eccellenti su ogni dominio) includono Gemini di Google, Large Language Model Meta AI (Llama) di Meta e Claude di Anthropic (ampiamente partecipata da Amazon), con tutte le loro versioni. Altri, invece, specializzati in una data tipologia di dati sono ad esempio DALL-E di OpenAI, progettato per la generazione di immagini a partire da un input testuale, AlphaFold di DeepMind (Google) per la previsione della struttura tridimensionale delle proteine, i modelli BERT o T5 di Google per la classificazione di testi e traduzioni o AudioLM di Google per la produzione di musica.

La differenza tra un modello fondativo e un applicativo può, tuttavia, non essere così definita, ma sfumata introducendo un ulteriore complessità nelle definizioni per i *policymakers* (Hagiu & Wright, 2024). Un esempio di ciò è il modello AstroLLaMa, un *foundation model* specializzato in astronomia basato sul modello LLaMa2 che agisce, quindi, da versione intermedia tra i *general purpose foundation models* e le applicazioni di settore.

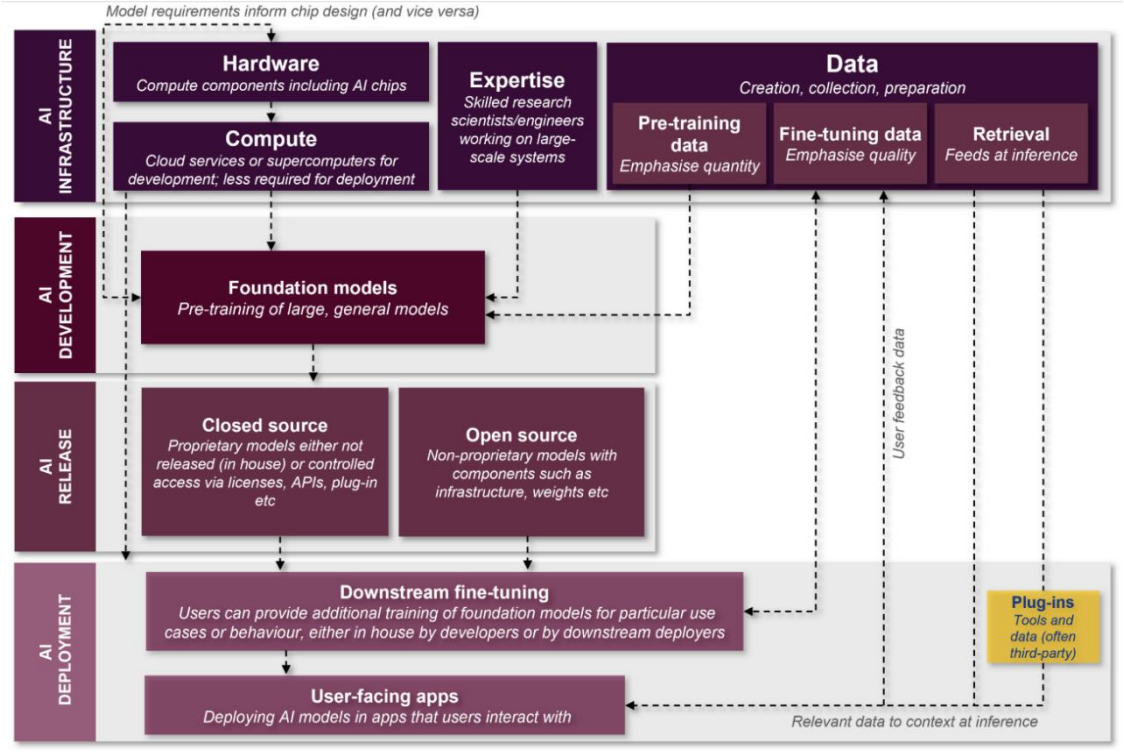


Figura 4: Overview della value chain dei Foundation Models, Source: CMA 2024

La grande varietà di opzioni di *deployment*, fa emergere un ampio spettro di possibili strutture aziendali. Ad un estremo di questo spettro troviamo una azienda completamente integrata verticalmente che possiede la propria potenza computazionale senza necessità di ricorrere a servizi cloud di terzi, è dotato di dati proprietari oltre che quelli pubblici, e infine, possiede l'*espertise* umana e i *development tools* necessari allo sviluppo dei modelli di base che affina e implementa nelle proprie applicazioni, prodotti o servizi. Incontriamo una struttura diametralmente opposta ogni qual volta ogni anello della catena del valore viene svolto da diverse imprese.

Ad oggi nel mercato si osserva un significativo grado di integrazione verticale con grandi *players* presenti a più livelli della *value chain.* Molti *FM developer,* come Microsoft, Amazon, Google e Meta possiedono internamente di infrastrutture chiave per lo sviluppo di *FM* quali *data-center* e *server.* Queste aziende sono inoltre presenti in diversi mercati *user-facing* dove i modelli posso essere facilmente integrati: online shopping, motori di ricerca, software e altro (Competition & Market Authority, 2023).

Inoltre, in un mercato come quello dell'AI, in cui le diverse tipologie di input non sono facilmente riproducibili sotto vincoli di capitale o di tempo, forte e dinamica è la presenza di partnership sia con finalità di investimento *R&D,* sia di tipo commerciale, che ne costituisce una caratteristica chiave per le dinamiche competitive. Si delineano così diverse possibili strutture di *value chain* (CMA , 2024):

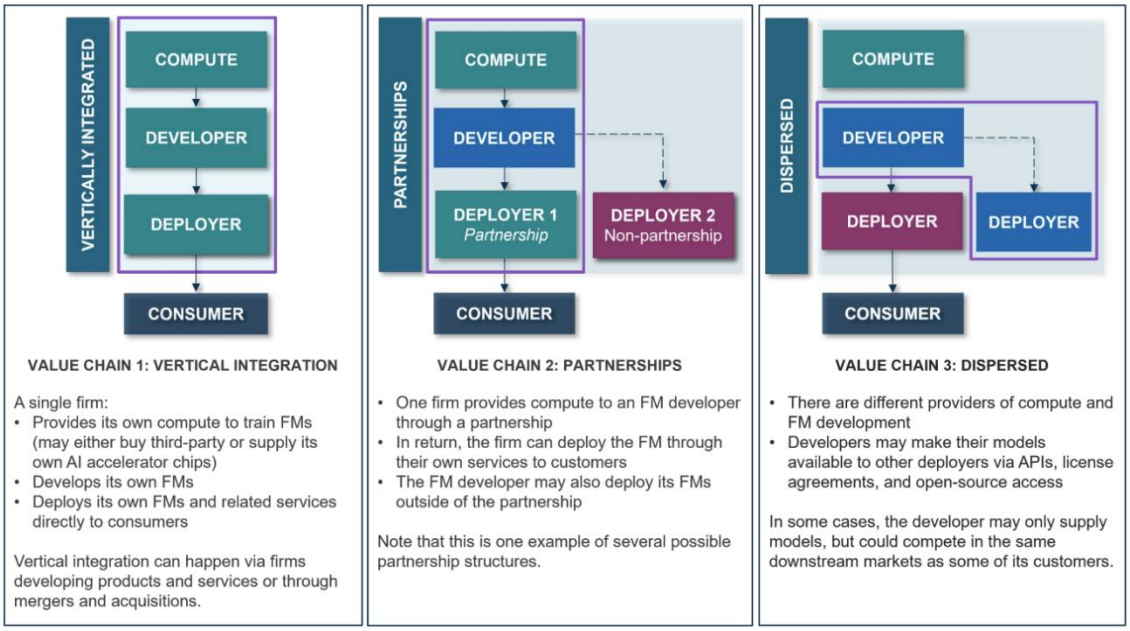


Figura 5: Diverse strutture di value chain, Source: CMA 2024

La complessità della catena del valore, l'elevato numero di *stakeholder* coinvolti e le *features* proprie del prodotto che richiedono enormi dataset, competenze tecniche di personale specifico e una potenza computazionale senza precedenti, uniti all'importanza strategica di questa tecnologia e la necessità di garantirne il progresso costituiscono una importante sfida per i *policymakers* adibiti alla salvaguardia della concorrenza nel mercato.

2.2 Caratteristiche e *snapshot* del mercato

La missione che guida la ricerca in campo Ai è la creazione di una *artificial general intelligence* (AGI), capace di eseguire ogni task cognitivo al pari, o meglio, di un essere umano. Se questa missione dovesse essere raggiunta la AGI potrebbe sostituire, se equipaggiata con i giusti strumenti e *hardware,* ogni lavoratore di qualsiasi settore. Questa visione estremista è sicuramente speculativa ma, data la rapida crescita della curva di adozione di questa tecnologia, mette economisti e regolatori nella condizione di dover considerare come mercato potenziale per i *foundation models*, l'intera economia (Korinek & Vipra, 2024).

Inoltre, in un contesto del genere, la concentrazione del mercato risulterebbe in una centralizzazione del potere nelle mani degli AGI *providers* senza precedenti storici: questo potere si estenderebbe, infatti, ben oltre i confini dell'economia, sfociando nel panorama sociale e politico mondiale (Korinek & Vipra, 2024).

Ciò implica la necessità di una elevata attenzione in ambito di concorrenza per evitare, l'insorgere di fenomeni anti-competitivi e un consolidamento eccessivo di potere, assicurandosi che l'alto potenziale tecnologico sia allineato agli interessi della società (Anton Korinek, 2023) (Korinek & Vipra, 2024).

Secondo gli autori Jai Vipra e Anton Korinek, (Korinek & Vipra, 2024) e Jon Schmid et al. nel *paper* dal titolo "*Evaluating natural monopoly conditions in the Ai foundation model market*" (Jon Schmid, 2024) infatti, i principali modelli hanno tutte le caratteristiche per sfociare in monopoli naturali. Da questa affermazione ne deriva la necessità di un intervento delle autorità Antitrust nel (i) garantire la contestabilità del mercato assicurandosi che eventuali tendenze monopolistiche non si propaghino verticalmente lungo la *value chain* e che (ii) date le difficoltà nel disciplinare il mercato, almeno vengano rispettati adeguati standard di qualità per contribuire efficacemente al *social welfare* (sicurezza, *privacy*, non-discriminazione, interoperabilità e affidabilità) (Anton Korinek, 2023).

Vi sono comunque anche altre visioni meno prospero per il futuro del mercato Ai. Un recente report di Goldman Sachs dal titolo: *"GenAi: Too much spend, too little benefit?"*

(Goldman Sachs, 2024) afferma di aver già raggiunto il limite tecnologico dei modelli e che quindi non si registreranno significativi incrementi di performance nei prossimi anni, con conseguente limitato impatto sull'economia. Ad esempio Daron Acemoglu stima che dal 2025 in avanti la crescita dell'Ai generativa sarà solamente dello 0,07% all'anno nel decennio successivo (Acemoglu, 2024). Anche i ricercatori dell'Università dei Stanford, nel documento annuale Ai Index report 2024 (Stanford University, 2024, p. 82) osservano come alcuni benchmark di performance utilizzati gli scorsi anni abbiano ormai raggiunto la saturazione e come le performance dei modelli si stiano sempre piu avvicinando a quelle umane senza però superarle in maniera radicale suggerendo quindi un appiattimento della curva di sviluppo.

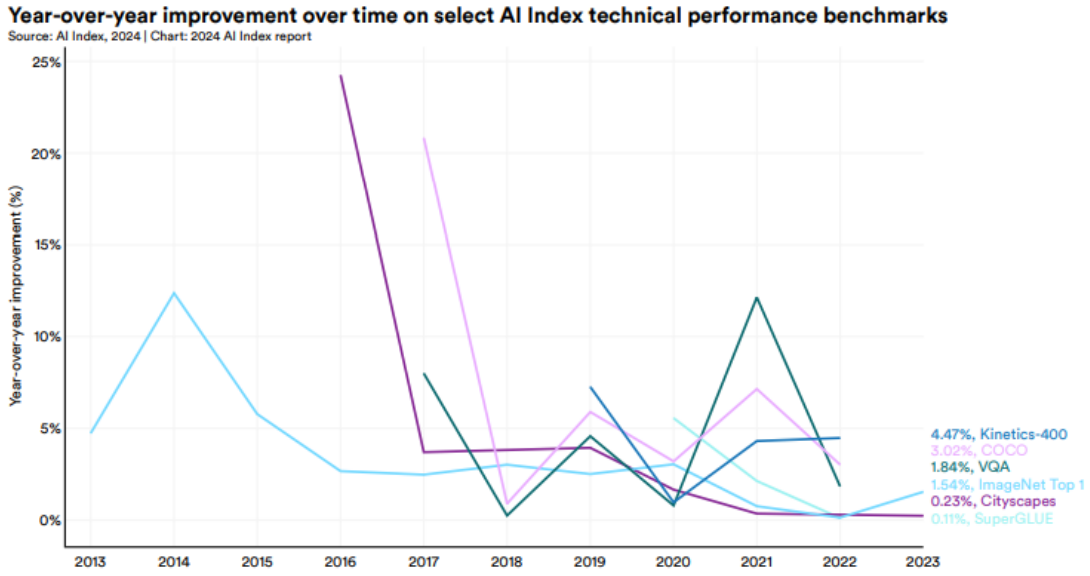


Figura 6: Miglioramento percentuale anno su anno di alcuni benchmark di performance. Source: Stanford 2024

Se così fosse le preoccupazioni circa la competizione nel mercato sarebbero solo un "*footnote*" nella storia economica (Korinek & Vipra, 2024).

Innovazione

Prima di addentrarsi nell'analisi della letteratura e dei driver competitivi a monte, come *input,* e a valle come *deployment,* è necessario sottolineare quale sia il modello di competizione in questo mercato.

Nell'ecosistema dei *foundation models* l'innovazione, ad oggi, è il principale motore della competizione (Thibault Schrepel &. A., 2024). Le aziende sviluppatrici, infatti, non competono per migliorare leggermente il prodotto rispetto ai concorrenti ma investono nel creare un modello che "distrugga" i presenti *competitors* e conquisti velocemente vaste quote di mercato grazie a prestazioni di gran lunga migliori, attraverso la così detta *disruptive innovation.* Non si tratta di un fenomeno sconosciuto, tuttavia, se nei mercati tradizionali un'innovazione distruttiva, di fatto, creava un nuovo mercato, nel nostro settore in esame, così come in altri mercati digitali (e.g. il mercato dei social network), questa dinamica è la caratteristica principale delle dinamiche competitive (Yang, 2023) (Petit, 2021).

In accordo con questo principio, il professore Thibault Schrepel dell'università di Amsterdam, affiliato a Stanford autore di numerosi articoli sulla concorrenza nei mercati digitali e non, propone un principio di *policy* che sia "*innovation first"*, ossia che dia all'innovazione la priorità in caso di *trade-offs.* Secondo questo principio i *policymakers* dovrebbero limitare l'operatività delle aziende e frenare l'innovazione solo quando si è in presenza di un concreto e documentato rischio per la società e non in maniera precauzionale (Thibault Schrepel &. A., 2024). Ciò si traduce in una principale raccomandazione: nuovi standard e regolamenti dovrebbero essere introdotti solo a seguito di un profondo *assessment* della minaccia monopolistica (Singla, 2023). Infatti, il peso della eccessiva regolamentazione viene spesso sottovalutato dai *policymaker* che però può essere così elevato da costituire una vera e propria barriera all'ingresso per start-up che non hanno il capitale o le competenze per sostenere tali costi di *compliance* (O'Brien, 2023) (Thibault Schrepel &. A., 2024). Inoltre, secondo Schrepel, le analisi sulle minacce alla concorrenza nel mercato AI dovrebebro essere condotte da un consorzio che unisca un comitato tecnico e diverse agenzie come il Digital Regulation Cooperation Forum ("DRCF") che racchiude l'Information Commissioner's Office

("ICO"), il Competition and Markets Authority ("CMA"), l' Office of Communications ("Ofcom") and the Financial Conduct Authority ("FCA") (Vanberg, 2023).

L'innovazione nel settore è estremamente dinamica; secondo il *Centre of Research on Foundation Models* (CRFM) di Stanford, dal settembre 2023 alla primavera del 2024 sono stati rilasciati nel mercato oltre 120 modelli, portando il numero totale di modelli offerti, conosciuti al pubblico e non privati, a 330 (ecosystem graphs, s.d.).

Se ad oggi, GPT-4 continua ad essere il più diffuso numerose aziende affermano di aver prodotto modelli migliori per determinate task e che pertanto cercano di surclassare l'egemonia di OpenAI, come GeminiUltra di Google, Cloud 3 di Anthropic o il più recente DeepSeek, modello *open-source* cinese rilasciato a fine gennaio 2025 (CMA , 2024).

Altra importante frontiera per l'innovazione è il segmento dei così detti *small foundation models*, modelli più "piccoli" in quanto a numero di parametri e, di conseguenza aventi un inferiore fabbisogno di potenza computazionale, che li rende adatti ad essere installati su piccoli *devices* (CRFM, 2024). La spinta innovativa in questa direzione è dovuta non solo al grande mercato potenziale, quello dei dispositivi personali, ma soprattutto dalla necessità di ridurre i costi di deployment a parità di prestazioni. Se infatti, come descritto in precedenza, la correlazione tra performance e numero di parametri è positiva, è altrettanto positiva la correlazione con i costi, da cui nasce un *trade-off* tra la dimensione del modello e i costi di *training* (CMA , 2024).

Esempi di modelli *small* sono Gemma 7B di Google e Zephyr 7B di Hugging Face che affermano di superare i grandi LLMs sotto diversi *benchmarks* (Google, 2024) (Lewis Tunstall, 2023), oppure lo *small language model* Phi-2 di Microsoft che con soli 2.7B di parametri dichiara superare il modello Llama2 di Meta da 70B parametri nel *common sense reasoning* (Mojan Javaheripi, 2023).

La ricerca nel campo dei modelli *small* segue tre linee principali: (i) la riduzione delle dimensioni attraverso il *fine tuning* (ottenendo modelli più *domain-specific*), ottenendo così modelli più performanti con meno parametri, (ii) lo sviluppo di architetture (numero di layers e tipologia di calcolo) innovative come alternativa ai *transformer* che riducano, a parità di parametri, la memoria e la potenza computazionale richiesta, (iii) la

produzione di nuove CPUs e GPUs con chip innovativi capaci di supportare un modello di AI su dispositivi piccoli come i *personal laptops* o *smartphones* (CMA , 2024).

Open Source

Fondamentale per comprendere le dinamiche di mercato lungo la *value chain* dei *foundation models* è la distinzione tra le due opzioni di *release* di questi modelli che determinano, insieme ai contratti di licenza, tipologie di API e di *partnerships*, le modalità con cui i modelli vengono resi disponibili ai *deployers* a valle della catena: *closed* o *open source.*

A partire dall'infrastruttura Linux, sino ai più moderni ecosistemi digitali come WordPress o Android, il ruolo dell'*open source* è stato significativo nel dare forma allo sviluppo tecnologico, slancio all'adozione di nuovi strumenti e nella scelta degli standard *de facto.*

Importando l'idea dal mondo del *software open source* dove apertura significa che il codice sottostante una data applicazione sia reso disponibile al pubblico per suo scrutinio ed eventuale modifica, nel settore dei modelli fondativi una definizione di apertura è che gli "*open systems*" sono quelli che forniscono "*transparency, reusability, and extensibility—they can be scrutinized, reused, and built on*" (David Gray Widder, 2023). Le tecnologie *open source* giocano un ruolo chiave nei così detti "*innovation commons",* ambienti collaborativi dove conoscenza, *skills,* strumenti e risorse vengono apertamente condivisi tra persone appassionate e con diversi *backgrounds* per supportare la nascita di nuove soluzioni (e.g. GitHub) (Potts, 2012) (CMA , 2024).

Pertanto, anche in questo nuovo settore digitale, gli *open foundation models,* se inizialmente faticano a competere con i grandi *players* privati, sono di grande aiuto nel promuovere l'innovazione nel *long term*, limitando le preoccupazioni Antitrust per i sotto-investimenti in ricerca e qualità tipici di situazioni monopolistiche o oligpolistiche. Per questa duplice funzione, i professori Schrepel e Potts, sottolineano l'importanza di analizzare il livello di apertura dei modelli e monitorare l'ecosistema internazionale dei *commons* con attenzione. A tal proposito, mancando ancora una precisa definizione

tecnica internazionale di "*open",* propongono una puntuale metodologia per indicizzare il livello di apertura di un modello considerando 18 principali caratteristiche tecniche, di trasparenza, legali e di governance dei contratti di licenza (Thibault Schrepel J. P., 2024). Il risultato delle loro analisi sottolinea quanto possa essere sfumata la distinzione tra *closed* e *open models* e quanto il livello di apertura possa variare tra un modello ed un altro introducendo un problema di definizioni per i *policymakers.*

La figura sottostante offre alcuni esempi di modelli posizionandoli in una scala di *openess level,* da quelli completamente proprietari, o *closed*, accessibili esclusivamente agli stessi *developer* (e.g., Flamingo di Google DeepMind), a quelli *black box* accessibili attraverso una API (e.g., GPT-4 di OpenAi), a quelli *cloud-based* che è possibile sia utilizzare sia affinare sempre attraverso una API (e.g., GPT-3.5 di OpenAI). Successivamente si posizionano i modelli comunemente definiti *open* caratterizzati da una maggiore *disclosure* circa parametri, pesi, dati utilizzati e codice, sia con restrizioni, come BLOOM di BigScience che proibisce l'utilizzo del modello nella stesura di contratti e altri documenti legali (HuggingFace, 2022), o senza limiti come GPT-NeoX di EleutherAI.

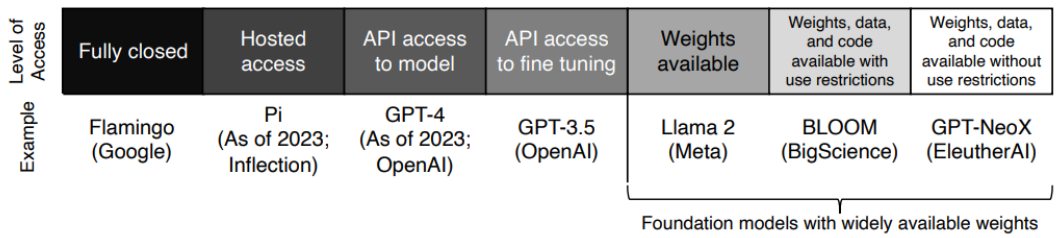


Figura 7: Openess Spectrum dei foundation models. Source: Stanford 2023

Similmente, altri ricercatori, così come Liesenfeld et al. (Andreas Liesenfeld, 2023) propongono indici diversi per misurare il livello di apertura considerando la quantità di codice condiviso, le informazioni rilasciate circa i dati e il loro completo trattamento, la trasparenza circa i parametri e meta parametri utilizzati e la relativa documentazione scientifica. Tutto ciò si allinea anche con il *Foundation Model Transparency Index* introdotto da Bommasani et al. del *Center for Research on Foundation Models* dell'Università di Stanford (Bommasani, et al., The Foundation Model Transparency

Index v1.1, 2024), che codifica per ogni modello un livello di trasparenza tenendo in considerazione cento diversi fattori sull'utilizzo delle risorse *upstream*, sulle tecniche di costruzione del modello e sul deployment a valle.

Sotto queste assunzioni, in questa tesi, con il termine *model openess* si vuole indicare un indice aggregato che tiene conto di tutte le variabili in gioco, permettendoci di distinguere tra le diverse tipologie di modelli. Seguendo la scelta dei professori di Stanford, consideriamo *open* i modelli appartenenti ai tre riquadri finali, ossia quelli con una disponibilità di informazioni circa pesi e parametri maggiore.

Come precedentemente accennato, l'*open source* permette ai modelli di essere costantemente revisionati da esperti del settore e studenti che collaborano tra loro e con l'azienda sviluppatrice nel correggere errori e nello sviluppo di nuove funzionalità, catalizzando, di fatto, il processo innovativo e competitivo e decentralizzando il potere di mercato dei leaders (Sayash Kapoor, 2024). A titolo di esempio si può citare il caso DeepSeek, modello cinese *open source,* che con un costo di *training* di appena $6M (1/20 della cifra per l'allenamento di GPT-4) è riuscito a raggiungere le performance di OpenAi, scuotendo l'intero mercato[2] con le sue innovazioni tecniche e di costo (DeepSeek-AI, 2024) (Xiao Bi, 2024).

Questi molteplici benefici hanno portato la comunità economico-scientifica a ricercare una sempre maggiore trasparenza e a sostenere l'*open source* come elemento essenziale nella competizione: "*make transparency, fairness and accountability the core of AI governance … [and] consider the adoption of a declaration on data rights that enshrines transparency*" queste le parole di Antonio Guterres, Segretario Generale delle Nazioni Unite[3]. Lo stesso Schrepel, unitamente al collega A. Pentland suggeriscono, quindi, di esimere la ricerca *open* dalle rigide leggi Antitrust in modo che "almeno un *provider* di *general public foundation model* rimanga nella porzione *open source* dello spettro" (Thibault Schrepel &. A., 2024). In particolare, suggeriscono che le aziende *open source*

[2] Vedi : https://www.deepseek.com/

[3] Press Release: Secretary-General Urges Security Council to Ensure Transparency, Accountability, Oversight, in First Debate on Artificial Intelligence, 18 Luglio 2023 (https://press.un.org/en/2023/sgsm21880.doc.htm**)**

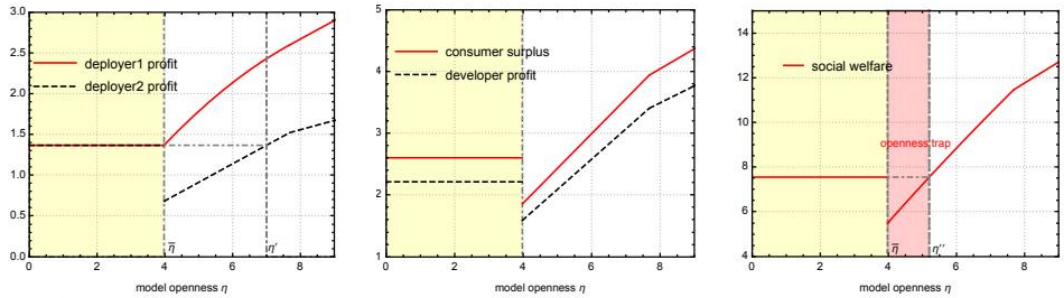
siano lasciate libere di collaborare e unire le risorse in *joint ventures* senza dover aspettare approvazioni dalle autorità nazionali e internazionali sulle fusioni o, in alternativa, essere soggette a procedure di *merge control* semplificate e più rapide. (Thibault Schrepel &. A., 2024).

Tuttavia, come dimostra la storia delle innovazioni tecnologiche, la correlazione tra l'apertura intesa come *technical disclosure* e la incapacità dei *leader incumbents* di appropriarsi delle rendite derivanti dalla loro invenzione è molto bassa. Così ad esempio nel settore farmaceutico, dove anche dopo l'estinzione dei brevetti e la pubblicazione della formula il leader di mercato continua a mantenere la sua *market share* attraverso il *brand,* la rete distributiva consolidata o altri assets vari relativi (e.g., Eli Lily e NovoNordisk con l'insulina). (Pierre Azoulay, 2024).

Pierre Azoulay, Joshua L. Krieger e Abhishek Nagaraj (Pierre Azoulay, 2024) considerano infatti le "filosofiche discussioni" circa la definizione di *openess* superflue e una distrazione dalla reale analisi della competizione nel mercato che, secondo gli autori, verte sul possesso degli assets complementari allo sviluppo dei modelli ancor più che sul modello stesso e le sue caratteristiche.

I professori Fasheng Xu dell'Università del Connecticut e Xiaoyu Wang della Facoltà di Business all'Università Politecnica di Honk Hong, il cui lavoro verrà ripreso e approfondito nel seguente capitolo di questa tesi, analizzano matematicamente come il livello di *openess* nel mercato dei *foundation models* impatti le dinamiche competitive *upstream* e *downstream,* tra le applicazioni, definendo la "*Openess Trap*" : un range di livelli intermedi di apertura che rendono il *social welfare* complessivo inferiore al caso in cui esistano solamente modelli chiusi.

PROPOSITION 2. *There exists a threshold η'', such that social welfare is reduced compared to zero openness if and only if $\bar{\eta} < \eta < \eta''$. We refer to this range of openness levels as the "openness trap."*



(a) two deployers **(b)** consumer & developer **(c)** social welfare

Figura 8: Impacts of model openness on each party's profit and social welfare (μ = 2, w = 0.4, γ = 0.3, c = 0.5)

La *openess trap* rappresenta una situazione paradossale in cui aumentare il livello di trasparenza oltre una certa soglia, conduce ad un risultato subottimale per tutti gli *stakeholders* coinvolti.

Dimostrano infatti come il livello di apertura impatti gli investimenti in *fine tuning,* la decisione temporale dei *deployers* di quando entrare sul mercato, la scelta sull'integrazione a valle per i *developers* e la profittabilità delle strategie di *free trials.*

Questo risultato sfida il pensiero comune della maggiore apertura come sempre benefica per *deployers* e utenti finali grazie alla riduzione dei costi di *fine tuning* e introduce una nuova prospettiva ai *policymakers*, i quali, devono accuratamente monitorare le dinamiche di mercato in modo da assicurarsi di incentivare l'*open source* senza però oltrepassare la soglia dell'*openess trap.*

La scelta dell'*open source* può avere quindi diverse motivazioni: etiche, tecniche o finanziarie. Quando ancora l'AI generativa era confinata all'ambito della ricerca ed aveva poche applicazioni, molte aziende hanno intrapreso il percorso di sviluppo di questi modelli proprio attraverso un paradigma *open source* in modo da attingere ad una ampia rete di AI *experts* senza gli alti costi del personale interno e con l'obiettivo di diffonderne l'utilizzo e educare l'*user.*

L'anno seguente la pubblicazione dei dettagli tecnici dell'infrastruttura *Transformer* nel 2017 (Ashish Vaswani, 2017), Google rilasciò il primo modello *transformer-based* BERT in una modalità *open source* cui seguirono, RoBERTa di Meta, GPT – 1 e GPT – 2[4] di OpenAI e LaMDA di Google, modelli che diedero slancio all'innovazione e al consolidamento delle prime *community*. La privatizzazione dei modelli è poi avvenuta nel tempo in concomitanza con gli ingenti capitali raccolti e la necessità (o volontà) delle aziende di controllarne lo sviluppo sia per questioni di "*competitive landscape"*, sia per "*safety concerns*" (OpenAI, 2023) (Anton Korinek, 2023).

L'*open source* incentiva gli utenti ad approfondire l'ecosistema dell'azienda stessa permettendole di estrarre valore dagli altri servizi e prodotti accessori offerti. Nel caso dei *foundation models* il mercato ausiliario include gli "*hosting service*": la potenza di calcolo offerta tramite i servizi cloud, servizi di *data storage* ecc.. per i quali, spesso, i modelli sono ottimizzati. In questo modo l'*open sourcing* può venire adottato come tattica di pubblicità e per aumentare le *revenues* generali (Ferrandis, 2022)[5].

Per concludere l'*overview* è necessario accennare ai rischi connessi all'*open source,* argomento di grande importanza per i *policymakers* oltre alla concorrenza. Come ogni potente strumento, se utilizzato per scopi illeciti o malevoli può causare non pochi danni. In un ambiente così aperto e dinamico come quello delle piattaforme *open source,* il rischio che questi modelli vengano impiegati per diffondere false notizie, designare email di *phishing* o altre truffe di vario tipo o, ancor peggio, costruire nuove armi biologiche o effettuare cyber-attacchi a istituzioni di rilievo, è sempre maggiore e necessita azioni di mitigazione *ad hoc* (Bommasani, Kapoor, Klyman, & Longpre, 2023) (Seger, 2023).

Economie ed effetti di rete

Le economie di scala dipendono dalla funzione di costo di una azienda e si verificano quando si hanno alti costi fissi e costi variabili relativamente più bassi. Nel contesto dei

[4] GPT – 1 nel 2028, GPT – 2 nel 2019 (Alec Radford, 2019)

[5] Ovviamente tutto ciò può anche essere venduto per i modelli proprietari. Ad esempio, OpenAI fornisce capacità computazionale dedicata e supporto ingegneristico a pagamento (Wiggers, 2023)

modelli fondativi i principali costi fissi sono quelli richiesti per la potenza computazionale del *training*, l'acquisizione di *dataset* per l'allenamento, l'acquisizione dei migliori talenti per lo sviluppo dell'algoritmo e altri costi accessori come energia e infrastrutture di diversa tipologia (Jon Schmid, 2024). Questi costi sono per lo più costi affondati, o *sunk costs,* in quanto non facilmente recuperabili in caso di bancarotta dell'azienda o fallimento del *training.*

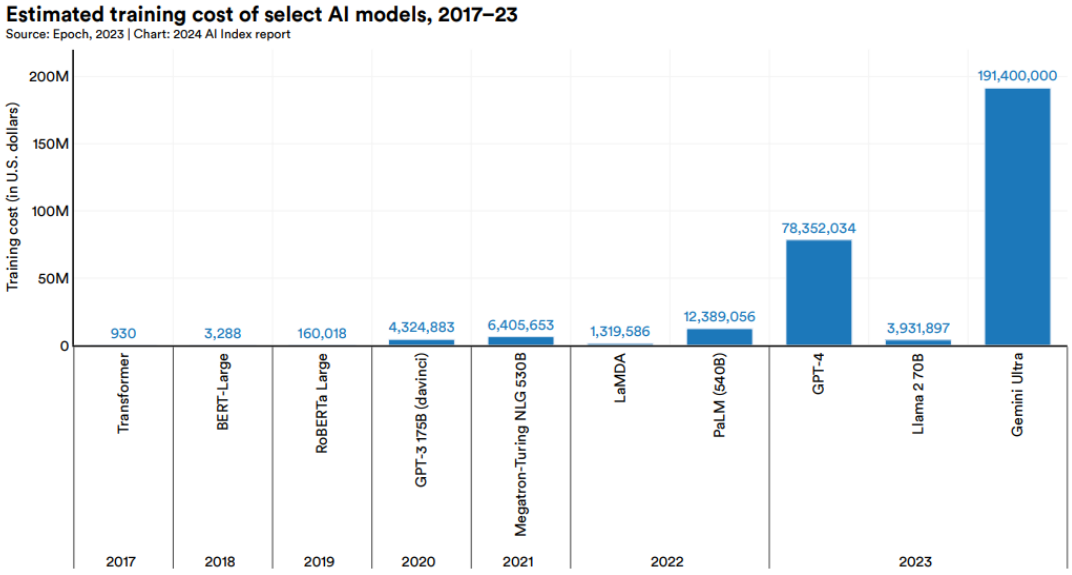


Figura 9: Costi di training stimati per alcuni dei modelli più importanti 2017-2023 Source: Stanford AI Index report 2024

I costi variabili invece riguardano i costi della fase di inferenza ossia quando il modello viene interrogato e produce un *output*, che vengono calcolati in base al numero di *token* in *input* e di *output* generato e richiedono potenza computazionale limitata alla singola esecuzione, che è tendenzialmente, molto bassa. Questi costi variabili non sono semplici da analizzare ma comunque, per una azienda sviluppatrice di *foundation models* sono, in proporzione, limitati[6]. Al momento della scrittura di questa tesi, ad esempio, OpenAi applica un prezzo di $1 ogni 100.000 sillabe (*tokens*) di testo generato usando GPT-4o.

[6] Si vedano i diversi marketplace per i FM, il prezzo è definito a numero di token in input e output, ad esempio AmazonBedRock: https://aws.amazon.com/it/bedrock/pricing/

La domanda sottostante questi costi è tuttavia quante inferenze per ora siano necessarie per sostituire completamente un essere umano lavoratore, e quanto questo, quindi, costi. (Korinek & Vipra, 2024).

I *foundation models* sono quindi un classico esempio di prodotto con forti economie di scala che costituiscono una potenziale barriera all'ingresso per nuovi entranti nella porzione *upstream* della *value chain* (Autorité de la concurrence, 2024). A valle, infatti i costi di *fine tuning* dei *deployers* sono di gran lunga inferiori e includono i costi del personale interno, del personale esterno utilizzato nel caso di *Reinforcement Learning based on Human Feedback* (RLHF) e i costi computazionali. Il *fine tuning* richiede meno dati, meno tempo e molta meno potenza computazionale rispetto alla fase di *general pre-training* e le aziende che se ne occupano tipicamente si affidano a servizi *cloud,* abbattendo così i costi di mantenimento dell'infrastruttura *hardware* (Carugati, 2023) (Korinek & Vipra, 2024).

Il settore è definito anche da economie di scala lato domanda, anche chiamate *network effects* che si presentano in varie forme, anche se, ad oggi, ancora inferiori come impatto rispetto alle economie lato *supply* (Anton Korinek, 2023).

Gli effetti di rete si verificano quanto il valore di un prodotto aumenta all'aumentare del numero di persone che lo utilizzano e sono tipiche dei servizi e delle piattaforme digitali. Nel mondo dei modelli fondativi essi si manifestano attraverso l'incremento della qualità del modello a seguito della raccolta dei dati dagli utenti che lo utilizzano, che attira nuovi clienti da cui si estraggono ulteriori dati e così via, creando un circolo virtuoso anche chiamato *data feedback loop*[7]. Si consideri ad esempio Google Maps. Utilizza l'intelligenza artificiale per consigliare il percorso più veloce per raggiungere una destinazione confrontando istantaneamente decine di diverse possibilità e tenendo in considerazione parametri come il traffico, la presenza di semafori o rallentamenti dovuti

[7] E' comunque importante notare che le definizioni, anche in questo caso, sono sfumate e oggetto di discussione. Ad esempio, spesso viene utilizzato soprattutto da regolatori e avvocati il concetto di "*data network effect"* con cui si intende il processo di miglioramento delle prestazioni del modello con l'utilizzo. Tuttavia, seppur vero, ciò non è automatico (come lo è ad esempio il raffinamento di un *recommendation system* sul singolo utente di un *social network*), ma richiede ulteriori investimenti nella preparazione dei nuovi dati e nel *re-training* del modello generale. Di conseguenza, alcuni autori non considerano questo effetto un effetto "di rete" in quanto manca l'automaticità del processo, ma lo definiscono più comunemente "*learning by doing"* o ancora, "curva di apprendimento" (Varian, 2018)

a lavori su strada o altro; più persone utilizzano l'app, più dati essa raccoglierà per affinare le proprie capacità predittive (Sheen S. Levine, Come gli effetti di rete rendono l'IA più intelligente, 2023). Tuttavia, non tutti i *data feedback loops* sono uguali (Hagiu & Wright, 2023 a). Uno sguardo ravvicinato rivela che la forza di questi *feedbacks* dipende dalla tipologia di modello (Thibault Schrepel &. A., 2024). Seguendo la tassonomia proposta da Schrepel, i modelli possono essere divisi in tre categorie caratterizzate da differenti livelli di *increasing returns*:

- *General public foundation models:* accessibili da chiunque online, possono essere di due tipologie a seconda dei dataset di allenamento, *general purpose* se allenati su vasti e vari dataset come GPT (accessibile attraverso l'interfaccia ChatGPT) e Gemini (utilizzabile online tramite il sito ufficiale o l'app) o *domain specific* come BloombergGPT.

  Godono di forti *increasing returns* provenienti da più direzioni. In primo luogo, vi è un forte *ecosystem effect:* maggiore è la diffusione di questi modelli maggiore è il numero di applicazioni che vengono sviluppate al di sopra e che attraggono ulteriori utenti e sviluppatori interessati sia al modello sia alla complessiva infrastruttura dell'azienda, la quale quindi può raccogliere dati tramite gli altri servizi offerti.

  Inoltre, questi modelli si migliorano a seguito della computazione di ogni *user input* con un andamento a ritorni decrescenti. Tuttavia, a differenza delle classiche piattaforme digitali che affinano il proprio *recommendation system* per ogni user in maniera continuativa, i *general purpose foundation models* non vengono costantemente riallenati e adattati ad ogni user ma, mentre i dati vengono raccolti, le interazioni avvengono con lo stesso modello sino al rilascio della versione successiva. La differenza tra queste due strutture di apprendimento e le relative ripercussioni sul mercato verranno approfondite nel capitolo 3 della presente tesi.
  Terzo, più utenti si hanno, più aumentano le *revenues* dell'azienda che può quindi investire in *dataset* esclusivi come quelli in licenza di YouTube o Reddit o di testate giornalistiche come il *Financial Times, Le Monde, Vox Media* o il *World Association of Newspapers and News Publishers* (Morris, 2024).

  Maggiore è la *user base* di un modello, maggiore è la reputazione di cui il modello gode, rendendolo un *partner* strategico per altre compagnie in altri progetti.

Importanti nell'ambito della reputazione e della fama di un modello sono tutti i servizi accessori sviluppati dall'azienda sviluppatrice o dagli *users* stessi come librerie di *prompt design* e *tutorials* che, semplificando l'utilizzo dei modelli, attirano altri utenti (Anton Korinek, 2023).

Infine, l'abilità delle *Big Tech*, principali sviluppatrici di *foundation models* generali, di raggiungere facilmente miliardi di users genera significativi *increasing returns.* Possono infatti integrare i propri modelli in altri prodotti aumentando i ritorni sugli investimenti e giustificando lo sviluppo di modelli *open source*, o far leva sulla esistente *user-base* per promuovere applicazioni *stand alone* sui propri modelli.

Queste aziende sono, pertanto, ben posizionate per affrontare la sfida della distribuzione e scalabilità di questi modelli sfruttando altri canali distributivi in cui sono presenti e riconosciuti godendo al massimo dei *feedback loops* che si creano in tutta l'infrastruttura.

- *Network foundation models:* accessibili a specifici *network* di aziende e persone, sono allenati su *dataset* privati spesso derivanti dalla cooperazione[8] tra questi enti. Possono essere creati da zero o basarsi su preesistenti *general FM* attraverso un primo *fine tuning.* Un esempio è Watsonx.ai di IBM, allenato su un "*curated, enterprise-focused data lake"* (IBM, 2023).

  Gli *increasing returns* per questa tipologia di modelli, non essendo facilmente trasferibili da una industria ad un'altra, sono limitati al singolo caso d'uso, quindi al settore e al grado di diffusione. All'interno di un settore specifico continua ad essere valido quanto detto per i modelli *general purpose:* una azienda che già offre servizi ad una industria sarà la meglio posizionata per offrirne l'infrastruttura Ai. Seppur, ad oggi, i modelli generali siano i più utilizzati dal grande pubblico, si pensa che nei prossimi anni emergeranno *dominant foundation models* per ogni settore grazie ai *feedback loops* di questi anni di sviluppo (Thibault Schrepel &. A., 2024).

[8] La condivisione di dati privati di aziende introduce sfide legali in quanto potrebbe condurre ad alcuni tipi di collusione o coordinamento nei *markets behaviors.* Inoltre, potrebbero venire infrante regole di privacy per i clienti di queste aziende, nel condividere dati all'interno del *network*

- *Personal foundation models:* creati *ad hoc* per una persona, azienda, istituzione o governo sono spesso derivanti da un modello generale e poi affinati su dati proprietari (e.g. Personal GPT). Hanno sperimentato, fino ad oggi, una curva di adozione più lenta ma alcune iniziative di Microsoft e Nvidia stanno diffondendo questa tipologia di *PersonalAi* (Nirmal Kumar Juluru, 2024) (Beatman, 2023).
  Questi modelli godono di scarsi *increasing returns* dovuti alla natura personale del modello e dalla non trasferibilità.
  Enti e individui producono una grande quantità di dati che vengono spesso immagazzinati attraverso servizi *cloud* di *data storage.* I principali *players* che offrono questo servizio possono quindi sviluppare e offrire *personal foundation models* ai propri clienti e influenzare negativamente la competizione nel settore limitando l'accesso ai dati proprietari degli *users* attraverso strategie di *tying* ad altri loro servizi, sistemi operativi o altri dispositivi. I dati personali sono centrali nel mercato del *self development,* stimato circa $47B nel 2022 con un CAGR atteso 2023-2030 del 5.5% (Grand View Research, 2023), che sta attirando l'attenzione di numerosi *players* nel fornire agli utenti analisi e consigli sartoriali circa la salute, le finanze, il lavoro, le attività ricreative, etc. Lo stesso Google in un "*leaked memo"* descrive le "*scalable Personal Ai"* come principale motivazione per cui né loro, ne OpenAi, hanno la garanzia di vincere la "gara" dei modelli fondativi (Dylan Patel, 2023).

(Thibault Schrepel &. A., 2024).

Grazie all'aumento delle prestazioni dovute ai *feedback loop,* i *foundation models* godono anche dei più tradizionali effetti di rete che accomunano la maggior pate dei servizi digitali. La diffusione di un modello infatti comporta la nascita di numerose applicazioni che lo implementano e che attirano altri utenti e sviluppatori. Più un modello attrae attenzione più vengono create librerie di prompt e di istruzioni, viene alimentato lo scambio di idee tra appassionati ed esperti e, in generale, l'azienda aumenta la sua notorietà. Il tutto comporta, quindi, lo sviluppo dell'ecosistema aziendale che può, potenzialmente, tradursi in nuovi profitti.

Il capitolo 3 approfondirà l'argomento dell'impatto sulla concorrenza dei *network effect* e dei *data feedback loop* per determinare se siano sufficientemente forti da costituire un indiscusso vantaggio competitivo.

Le economie di scopo sono significative a causa della natura, più o meno, generale dei modelli che vengono utilizzati per *tasks* molto diversi tra loro in una ampissima varietà di settori industriali, potenzialmente, tutta l'economia. Grazie a questa cross-settorialità e cross-funzionalità, le aziende riescono ad incrementare i ritorni sugli investimenti e ad abbattere i costi, sfruttando una stessa tecnologia su più aree di business.

Accanto a queste economie di scopo lato offerta, ve ne sono altre lato domanda derivati dal *bundling* di servizi diversi. Molte aziende sviluppatrici, anche quelle *open source*, infatti, "*runnano"* i propri modelli sui propri server, legando la licenza del modello all'affitto di potenza computazionale in *cloud* (Anton Korinek, 2023).

Riassumendo, seppur ad oggi l'ambiente dei *foundation models* sia dinamico e competitivo, presenta, a livello strutturale, forti economie di scala e scopo, effetti di rete, e alti costi di sviluppo tipici dei mercati altamente concentrati. Ricercatori e regolatori devono, infatti, continuare a monitorarne le dinamiche in modo tale da deviare quella che sembra essere la direzione naturale.

Snapshot del mercato

Il 2022 è stato l'anno di svolta nella storia dell'intelligenza artificiale. Nel dicembre di quell'anno è stato infatti rilasciato al pubblico la prima versione di ChatGPT, il sito (e successivamente app) *chatbot* basato sui modelli GPT di OpenAi. L'adozione da parte del pubblico è stata pressochè immediata: 1 milione di utenti in appena 5 giorni e circa 100 milioni di utenti attivi il mese successivo (Autoridade da concorrencia, 2023). Seppur, in termini di diffusione e monetizzazione, OpenAi goda ancora tutt'oggi dei benefici del *first mover*, nel corso del 2024 il mercato per i *foundation models* "di frontiera" (ossia i modelli generali più grandi e con le prestazioni migliori sul mercato) è stato altamente dinamico e caratterizzato da un buon livello di competizione con 16

nuove aziende emergenti i cui modelli, alcuni rilasciati con un paradigma *open source,* hanno superato GPT-4 in diversi test (Korinek & Vipra, 2024).

Il miglioramento delle performance dei modelli ha reso importante capire quali siano i modelli preferiti dal pubblico (Stanford University, 2024). Lanciato nel 2023, *Chatbot Arena Leaderboard (Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots)* è la piattaforma *open* più usata ad oggi per comparare le performance dei vari modelli, sia in *overall,* sia su *task* più specifici. Sono infatti disponibili i *ranking* per le capacità matematiche, la scrittura creativa, codice, etc. (Wei-Lin Chiang, 2024). La tabella sottostante mostra una lista dei *AI labs* con i loro modelli meglio valutati dal pubblico alla data del 17 Febbraio 2025, ordinati sulla base dello *score* LMSYS (*Language Model System*). L'indicatore si basa sul sistema di *rating* Elo, in origine sviluppato per il gioco degli scacchi, con il quale i modelli vengono messi a confronto nel rispondere a reali richieste umane e poi valutati direttamnete dall'*user* che sceglie la risposta di qualità maggiore (Wei-Lin Chiang, 2024). I punteggi dei modelli vengono continuamente aggiornati per riflettere la probabilità relativa del modello di vincere attraverso la formula :

$$P(A\ beats\ B) = 1/\left(1 + 10^{-\Delta/400}\right)$$

dove Δ rappresenta la differenza tra gli *score* LMSYS tra i due modelli concorrenti A e B.

Lab	Country	Best Model	Released	LMYSY
Google DeepMind	USA/UK	Gemini 2.0 – Flash Thinking Exp	21 gennaio 2025	1384
OpenAi	USA	GPT – 4o latest	29 gennaio 2025	1377
DeepSeek	CHINA	DeepSeek – R1	20 gennaio 2025	1361
Alibaba	CHINA	Qwen 2.5 - Max	28 gennaio 2025	1332

Tabella 2: Migliori foundation models providers secondo il LMSYS. *Source: leaderboard su https://lmarena.ai/?leaderboard. Accesso il 17 Febbraio 2025*

Come evidenzia la vicinanza di punteggi LMSYS tra i modelli *leaders* di mercato, essi hanno performance molto simili e, dal punto di vista economico, sono pertanto (quasi perfetti) sostituti. I primi due modelli Gemini 2.0 – Flash Thinking Exp e GPT – 4o differiscono solamente di 7 punti il chè si traduce in una probabilità di *outperformance* di Gemini di appena il 51,0%. Anche in competizione con il quarto modello in classifica, Google ha solamente il 57,4% di probabilità di vittoria.

Un altro *benchmark* molto utilizzato nel settore è l'indicatore MMLU ideato da Hendrycks et al. che misura la conoscenza generale e l'abilità del *problem solving* in campi scientifici dei modelli (Dan Hendrycks, 2021). I risultati di questo *ranking* sono in linea con i precedenti anche se qui a vincere la competizione è Claude 3.5 Sonnet di Anthropic [9] (rilasciato il 22 ottobre 2024) a parimerito con GPT – 4o (di ottobre 2024).

A conferma dell'attuale scenario altamente competitivo è inoltre importante notare che tutti i modelli sopra citati sono stati rilasciati nell'ultimo mese dai giorni di scrittura di questa tesi (o ultimi quattro se consideriamo anche Claude 3.5 Sonnet). Dato il veloce ritmo di sviluppo di questo mercato le aziende *developer* rilasciano regolarmente versioni aggiornate dei loro modelli con la finalità di superare l'ultimo modello competitor e riconquistare i primi posti di una ipotetica classifica di performance e diffusione. Così come accadde nell'agosto del '24 quando, non appena Google Gemini superò GPT-4o nel *ranking* LMSYS, OpenAi rilasciò la nuova versione riottenendo in pochi giorni la prima posizione (Korinek & Vipra, 2024). Questo, in realtà è diventato negli anni un *trend* consolitado nel mondo *tech* dell'AI che deve però essere ben monitorato dalle autorità garanti della concorrenza. Di per sé, infatti, il parallelismo dei comportamenti tra imprese non è illegale ma semplicemente frutto delle caratteristiche e della struttura stessa del mercato: le imprese monitorano le azioni dei propri concorrenti per non perdere quote di mercato e comportarsi nella maniera strategicamnete migliore. D'altra parte il parallelismo è tipico dei mercati oligopolistici dove il rischio di collusione esplicita o tacita o pratiche facilitanti è reale e molto alto (Università di Palermo, -). Come esempio della velocità del mercato si evidenzia come OpenAi abbia aggiornato il suo ultimo

[9] Dati rilevati sul sito MMLU Benchmark (Multi-task Language Understanding) | Papers With Code, il 17 febbraio '25 uniti al documento di Anthropic "Claude 3.5 Sonnet Model Card Addendum" (Anthropic, 2024)

modello GPT – 4 sette volte dal lancio della prima versione nel marzo del 2023. Questi *update* hanno notevolmente migliorato la qualità di risposta del modello (l'attuale punteggio LMSYS della prima versione di GPT – 4 è 1186 (LMArena, s.d.)), aumentato la sua velocità di calcolo di tre volte e il numero di parole che è capace di processare di sedici volte, riducendo i costi di generazione di certo output del 92%; il tutto nell'arco di 20 mesi[10] (Korinek & Vipra, 2024).

Come analizzato nei paragrafi precedenti, il principale *driver* della competizione è l'innovazione e le aziende *developer* competono per ottenere più *market share* possibile sia per un guadagno diretto, ma principalmente per quello indiretto derivante dagli altri prodotti e servizi offerti nell'ecosistema aziendale. Il prezzo non è quindi, in questo scenario, un elemento centrale del *business model* di queste aziende; molti osservatori hanno infatti sottolinato che il *pricing* applicato dagli *AI Labs* appena permette loro di coprire i costi variabili, non portando di fatto, alcun profitto diretto (Knight, 2024): le dinamiche competitive sembrano quindi essere vicine alla competizione di *Bertrand* (Korinek & Vipra, 2024).

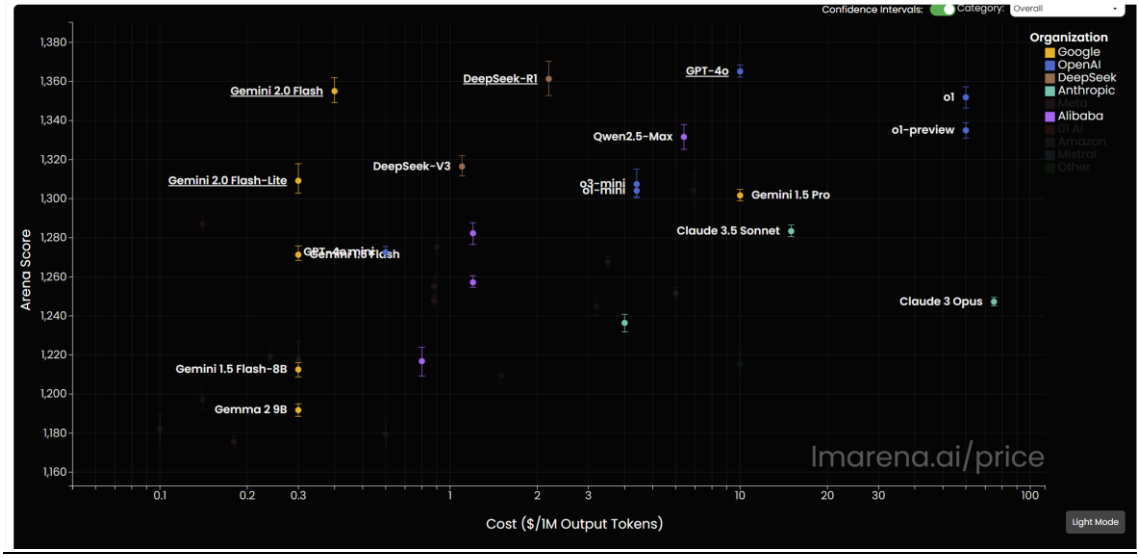


Figura 10: Grafico prezzo-performance dei modelli principali. Source: LMArena.ai il 17/02/2025

[10] Questi valori si riferiscono al periodo marzo 2023 - novembre 2024, mese in cui Korinek e Vipra li scrivono. Da novembre 2024 a febbraio 2025 c'è infatti stata una ulteriore *release*; i multipli e la percentuale di costo potrebbero, quindi, essere leggermente migliori

Conoscere il *background* dei principali *players* del settore può essere utile per meglio comprendere le loro strategie e motivazioni.

OpenAi è stato tra i primi a scommettere sul futuro dell'intelligenza artificiale generativa e sul successo dei *large language models.* Fondata nel 2015 con la *mission* di rendere l'AI accessibile a tutta l'umanità, è stata l'azienda *leader* di mercato sin dal lancio del primo modello GPT – 1 nel 2018. Nel 2019 creò una sussidiaria *non-profit* con l'obiettivo di raccogliere fondi per sviluppare il modello e sostenere gli enormi costi computazionali. Il principale investitore fu, e ad oggi ancora è, Microsoft che sino ad oggi ha investito circa $14B. L'ultimo *round* di finanziamenti è stato lo scorso autunno per $6,6B che aveva portato l'azienda ad una valutazione di $157B ma si vocifera di una possibile nuova raccolta di capitale di $40B che la farebbe salire ad una *post money valuation* di $340B[11] (Stefano, 2025), probabilmente in vista dei nuovi GPT – 4.5 e GPT – 5 e il processo di unificazione dei modelli annunciato dal CEO di OpenAi, Sam Altman, sulla piattaforma X il 12 febbraio 2025[12].

Altro *main player* del settore è Google DeepMind di proprietà di Alphabet (Google) con la sua serie principale di modelli *Gemini.* DeepMind è il risultato della fusione tra Google Brain, il dipartimento di AI avanzato di Google creatore del *trasformer* nel 2019, e DeepMind, una società inglese di ricerca in AI acquisita da Alphabet nel 2014. I modelli Gemini sono più recenti ma hanno da subito raggiunto le vette delle classifiche prestazionali sfidando il primato di GPT – 4 e superandolo in svariati test (Stanford University, 2024), soprattutto in quelli legati al ragionamento matematico; ad esempio, AlphaProof e AlphaGeometry si sono aggiudicate la medaglia d'argento alle olimpiadi internazionali della matematica (Korinek & Vipra, 2024).

Alla diciottesima posizione della scala LMSYS e primo in quella MMLU, con Claude 3.5 Sonnet si trova Anthropic. L'azienda fu fondata da un ex dipendente di OpenAI in disaccordo con la nuova direzione commerciale e *for-profit* di quest'ultima. Inizialmente

---

[11] Articolo del 11 febbraio '25

[12] Per leggere il post : Sam Altman su X: "OPENAI ROADMAP UPDATE FOR GPT-4.5 and GPT-5: We want to do a better job of sharing our intended roadmap, and a much better job simplifying our product offerings. We want AI to "just work" for you; we realize how complicated our model and product offerings have gotten. We hate" / X

creata come una *public benefit corporation* (PBC) come OpenAi, raggiunse i vertici della classifica prestazionale nella primavera del 2024 confermandosi una agguerrita rivale dei *labs* sopra citati. La principale *partnership* della società è quella con Amazon che permette di avere accesso all'infrastruttura *cloud* e hardware di AWS in cambio di un accesso preferenziale ai clienti di AWS per nuove *features* del modello (Biagio, 2024). Il duo ha come obiettivo lo sfruttamento di sinergie operative e di costo e la creazione di un forte *brand* che possa competere con il colosso OpenAi-Microsoft che detiene complessivamente circa il 69% del mercato.

La figura 11 mostra, infatti, la *market share* dei principali *players* a dicembre '23. Seppur ad oggi i numeri possano essere leggermente diversi grazie all'ingresso di nuovi *players,* la *leadership* è chiara e inequivocabile. In questa immagine Claude di Anthropic si trova incluso in "*Others*" oppure in "AWS" se raggiunto tramite *Amazon Web Services.* Non sono presenti in questa infografica neanche i modelli *open source* in quanto non hanno *revenues.*
A conferma di questa *share distribution* è la valutazione finanziaria di queste aziende con OpenAi in prima linea seguito da xAI[13] con $50B (Altchek, 2024), Anthropic con $60B (Konrad, 2025).. si vedano ulteriori dettagli nell'*annex A.*

[13] Azienda fondata da Elon Musk, in origine co-fondatore di OpenAi, nel 2023 e che occupa ad oggi con il suo modello Gork-2 (13/08/2024) la 18esima posizione nel ranking di LMArena.Ai con un LMSYS di 1288. Ad ottobre '24 occupava la terza posizione con un LMSYS di 1290 (Korinek & Vipra, 2024)

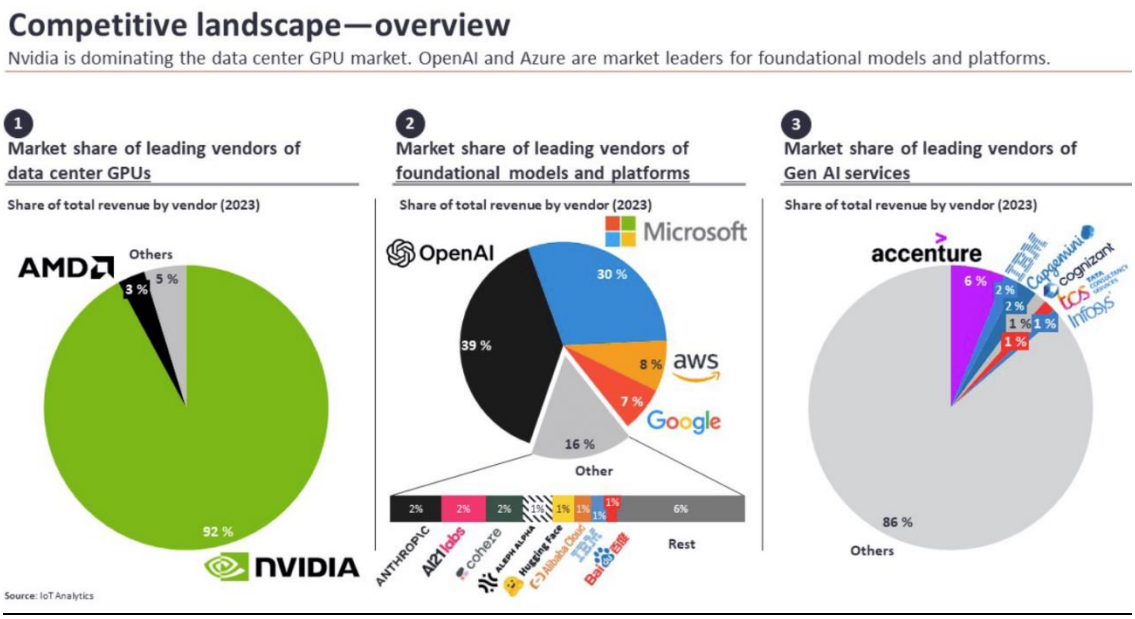


Figura 11: Market share nei mercati FM, GPU e servizi a fine 2023 Source: Generative AI Market Report 2023–2030 (Fernandez, 2023)

Nel già citato *leaked memo* Google afferma di non avere, così come neanche OpenAi, alcun "fossato" di separazione tecnologica rispetto agli altri *players.* Con questa affermazione sottolinea la dinamica competitiva del mercato che cerca l'innovazione "*disruptive*", ottenibile in diversi modi, anche da aziende *open source* che godono di una forte e dinamica community di esperti (Dylan Patel, 2023). L'*open source,* infatti, ha raggiunto livelli eccellenti paragonabili ai modelli proprietari come si nota nella classifica LMSYS.

L'azienda cinese DeepSeek, fondata nel 2023 in Cina, ha introdotto lo scorso fine gennaio il suo ultimo modello DeepSeek – R1 con cui, con un investimento di appena 6 milioni, ha raggiunto le prestazioni di OpenAi mostrando al mercato come sia possibile raggiungere certi livelli di *performance* con una ridotta potenza computazionale, e pertanto investimenti di gran lunga inferiori; notizia che si è riflessa in un crollo del 18% del valore delle azioni di Nvidia, leader mondiale dei *chip* e delle GPU per l'Ai (Picchi, 2025) e di altre aziende del settore *tech-Ai.*

Pochi giorni dopo il rilascio di DeepSeek – R1, il colosso cinese del commercio Alibaba ha annunciato il lancio del suo modello Qwen 2.5 – Max, completamente *open source*,

che vuole sfidare i due *main comptitors* DeepSeek e Llama 3.1 – 405B di Meta. Il successo di Alibaba nel campo dell'Ai generativa si è notato anche durante il rilascio del modello precedente che sulla piattaforma HugginFace ha raggiunto i 94 milioni di download superando il modello più scaricato di Meta, affermandosi come modello *open source,* più scaricato al mondo (Servidio, 2025).

Non citato nelle varie classifiche nei paragrafi precedenti ma importante presenza nello scenario competitivo attuale[14] è Meta con la gamma di modelli Llama. Meta è stata la prima tra le *Big Tech* americane a scegliere una differente *business strategy:* rendere i propri modelli scaricabili e modificabili a piacimento sui propri computers. Questa strategia è guidata da due motivazioni principali. In primo luogo, il *core business* di Meta è l'*advertising* sulle piattaforme *social*; pertanto, permettere l'utilizzo gratuito del proprio LLM non impatta la principale fonte di *revenues*, anzi, può solamente aiutare e supportare i *creators,* con conseguente aumento dell'*engagement* con l'ecosistema aziendale. (Thibault Schrepel &. A., 2024). In *secundis,* nonostante il *Chief AI scientist* sia Yann Lecun, un luminare dell'Ai, Meta è stata un *late entrant* sul mercato dei modelli fondativi, rilasciando il primo modello Llama 1.0 a febbraio '23, che era comunque al di sotto del livello prestazionale dei modelli di frontiera di quel periodo. L'unica modalità di attrazione utenti era quindi la scelta dell'*open source* (Korinek & Vipra, 2024). Nel tempo, comunque, l'impegno della comunità *open source* ha portato Meta a rilasciare nel 2024 Llama 3.1 – 405B, modello avanzato dalle prestazioni eccellenti.

L'accesa competizione e il miglioramento delle prestazioni hanno spinto verso il basso il prezzo tollerato dal mercato agevolando così la diffusione dei modelli Llama e degli altri modelli *open source*, gratuiti, portando con sé notevoli vantaggi dal punto di vista socioeconomico. I *foundation models* sono, infatti, beni non rivali la cui distribuzione gratuita corrisponde ad una soluzione di *pricing* di *first best* e al massimo valore di *consumer surplus* (Korinek & Vipra, 2024).

In generale, l' LMSYS *leaderboard*, elenca oltre un centinaio di modelli degni di nota rilasciati a partire da inizio 2023, riflettendo la dinamicità del mercato. Alcuni competono

[14] 24esima nella classifica di LMArena con un LMSYS di 1269 (al 18/02/2025)

sulle prestazioni come quelli elencati nella tabella 2, altri si differenziano per la dimensione ridotta implementabile su *laptop* o cellulari o altre caratteristiche. Per non appesantire la trattazione si vedano ulteriori dettagli nell'annex B.

2.3 *Asset* complementari e barriere all'ingresso

Uno dei principali vantaggi dell'essere *first mover* è la possibilità di assicurarsi un accesso privilegiato agli *input* della produzione e ad altre risorse necessarie ad operare in quel mercato. I *late entrants* potrebbero, infatti, trovare complesso attingere alle medesime risorse, potrebbero dover pagare prezzi più alti o avere difficoltà nell'ottenere l'accesso agli *assets* complementari (Anton Korinek, 2023). L'insieme degli ostacoli che disincentivano l'ingresso di nuovi entranti a mercato vengono definite barriere all'ingresso e, avendo un impatto diretto sul livello di concentrazione, devono essere opportunamente monitorate e indagate dai *policymakers.*

Nel documento "*Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy*" del 1986, Teece suggerisce che l'abilità degli innovatori di appropriarsi dei ritorni dei propri prodotti è determinata da due forze fondamentali: il controllo sul *core know how* dell'innovazione, anche detto "*appropriability*", e il controllo sugli *asset* complementari necessari per convertire questa conoscenza in una *value proposition* per cui i clienti sono disposti a pagare (Teece, 1986). Questi *assets* possono essere tangibili come gli strumenti R&D, la rete di distribuzione o *asset* produttivi, o intangibili come l'esperienza in ambito di regolamentazione, *skills* specifiche, relazioni nell'industria o più semplicemente il riconoscimento del *brand.* Il mercato dei modelli fondativi può quindi essere analizzato sotto queste lenti studiando le caratteristiche e gli *asset* che ne costituiscono barriere all'ingresso che, pertanto, plasmano la competizione (Pierre Azoulay, 2024).

Di seguito vengono quindi analizzate le tre principali risorse limitate che corrispondono a queste due forze in gioco: personale specializzato, dati e infrastruttura di calcolo (Korinek & Vipra, 2024) (Carugati, 2023).

Talenti e *appropriability*

Vi sono due principali modi con cui le aziende possono proteggere la conoscenza generata dal loro *team* di ricerca: diritti di proprietà intellettuale (IP) come marchi, brevetti e diritti d'autore, e vari meccanismi per ostacolare il *reverse engineering* come

segreti industriali, accordi di non divulgazione o di non concorrenza e la conoscenza tacita (Pierre Azoulay, 2024).

Per quanto riguarda i brevetti, non sembrano essere la soluzione migliore per i progressi in campo *foundation models* sia per le più pratiche questioni burocratiche che non permettono alla procedura brevettuale di essere al passo dell'innovazione, sia perché, spesso, gli sviluppi in campo AI/ML rimangono in un perimetro di concetti astratti lontani dai requisiti di brevettabilità (i.e. l'industrialità) (Pierre Azoulay, 2024). In generale, nel settore AI il numero di richieste di brevetto ha registrato negli ultimi anni un *trend* in forte crescita[15] (Stanford University, 2024), accompagnato però da un aumento della discrepanza tra i *patents filed* e quelli *granted,* a causa delle criticità sopra descritte. Inoltre, è importante sottolineare come la maggior parte delle aziende depositi brevetti per l'approvazione che sono relativi a tecnologie confinanti il campo dei *FM* o inerenti a specifiche *features* di *software* che incorporano l'output di un modello generativo[16] (Pierre Azoulay, 2024).

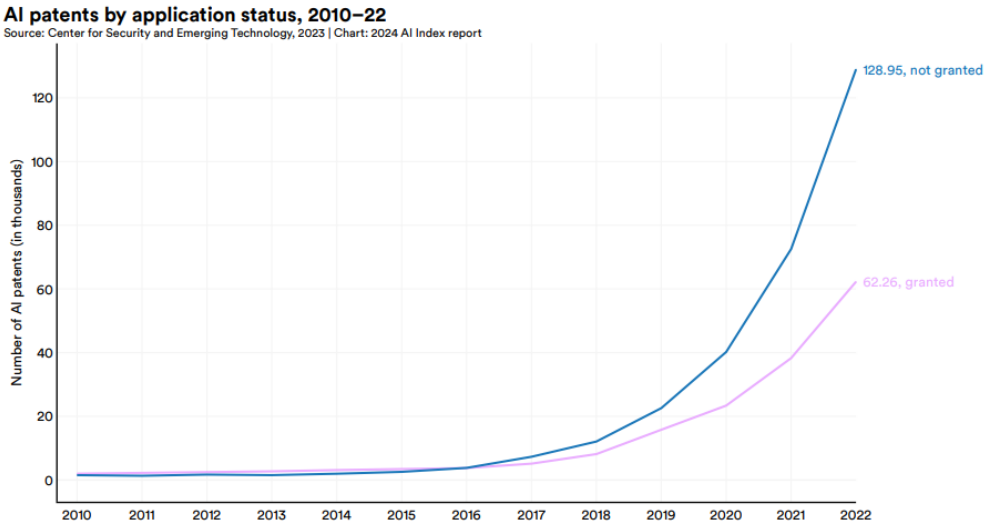


Figura 12: Stato dei brevetti depositati nel settore Ai nel decennio 2010-2020 Source: AI Index report, Stanford

[15] solamente dal 2021 al 2022 il numero di brevetti in campo AI concessi nel mondo è aumentato del 62.7% (Stanford University, 2024)

[16] Si veda ad esempio il brevetto di Microsoft: "Neural network categorization accuracy with categorical graph neural network" che descrive una nuova modalità per la categorizzazione grazie all'utilizzo di un graph neural network (Tianchuan Du, 2024) o l nuova feature di Adobe per Photoshop che consente la ricostruzione automatica di immagini in caso parte di una immagine sia cancellata.

Lo sviluppo dei *foundation models* si fonda su principi statistico-matematici e di *computer science* ben fondati e conosciuti su cui si basano: il *design* degli esperimenti per testare le ipotesi circa il comportamento del modello, il test delle *performance* su *task* e *dataset* reali e la formalizzazione rigorosa dei risultati ottenuti rendendoli implementabili come migliorie. Queste *skills* più tecniche sono tendenzialmente presenti in tutte le aziende, tuttavia, la definizione di un nuovo modello porta con sé una buona dose di intuito ed esperienza che difficilmente sono acquisibili dalla mera teoria. Questioni importanti che definiscono il modello quali la scelta dell'architettura migliore e degli iper-parametri, trovare il modo migliore per evitare l'*over* e *under fitting* e per preparare i dati di allenamento, oppure ancora, definire i programmi di *human reinforcement learning* richiedono necessariamente molti errori e iterazioni al posto di un processo predeterminato. Questo tipo di conoscenza che Pierre Azoulay definisce "*craft knowledge*" per sottolineare la reale (e quasi paradossale) artigianalità di un modello di AI, è tacita e incorporata nel processo stesso di *training* e sviluppo e costituisce un primo mezzo di appropriabilità.

Un'altra modalità per mantenere il controllo sulla conoscenza è la semplice non divulgazione dei dettagli del modello; non solo dei parametri ma anche del *data corpus* utilizzato, delle tecniche di *pre processing* dei dati e delle caratteristiche tecniche dello *user interface layer.*

Le aziende pioniere di questa tecnologia sembrano quindi avere un regime di appropriabilità più rigoroso grazie alla conoscenza cumulata in anni e anni di ricerca e alla possibilità di mantenerla proprietaria e tacita (piuttosto che attraverso IP) (Pierre Azoulay, 2024). Tuttavia, a possedere questa così detta "conoscenza tacita", più che le aziende in sé, sono le persone che con esse lavorano o collaborano. Di conseguenza i *leaks* sono frequenti e difficili da prevedere o gestire; una volta che i dettagli interni di un modello vengono divulgati, non vi è modo di fare un passo indietro. A ciò si aggiungono le vibranti e talentuose *community open source* che immediatamente sfruttano queste informazioni per migliorare i propri modelli (Pierre Azoulay, 2024).

In generale la domanda per ingegneri e ricercatori che possano costruire i modelli di frontiera e l'infrastruttura server necessaria di gran lunga supera l'offerta. Ciò è evidenziato nel fatto che le aziende produttrici di *FM* sono spesso disposte ad assumere

personale senza esperienza pregressa in progetti AI per un successivo *in house training,* e un conseguente maggior costo del personale (Anton Korinek, 2023). Inoltre, vi è una stretta relazione tra talenti e risorse computazionali dell'azienda: i migliori ricercatori e sviluppatori preferiscono gli *employer* con la più alta disponibilità di calcolo che possa permettere loro di sviluppare modelli migliori senza limitazioni di costo. L'accesso a queste risorse, congiuntamente ai professionisti e ricercatori che su esse lavorano, viene infatti utilizzato come metodo di assunzione per attrarre talenti da tutto il mondo. Come conseguenza, nuovi entranti nel mercato dei *foundation models*, volendo assumere personale con previa esperienza in aziende *leader*, si vede costretta ad offrire un generoso *premium* aumentando notevolmente il costo di ingresso a mercato[17]. Migliori informatici e ingegneri conducono a importanti migliorie al codice e all' intera infrastruttura permettendo notevoli risparmi in termini di costo computazionale, il ché contribuisce ad aumentare la domanda per professionisti di livello e, quindi, i salari.

L'alta domanda di *expertise* e gli alti costi computazionali disincentivano i migliori ricercatori e professori a rimanere in ambito accademico, dove potrebbero insegnare ed ispirare le nuove generazioni di studenti, per orientarsi verso l'ambiente *corporate.* A. Korinek e J.Vipra riportano, infatti, che la percentuale di Ph.D in campo AI che lascia le università è salito dal 21% del 2004 al 73% nel 2022 e sottolineano, in merito, l'importanza dei fondi pubblici a sostegno della ricerca universitaria[18] (Korinek & Vipra, 2024).

Per ovviare alla difficoltà nell'assunzione di esperti, molti *developer* hanno intrapreso pratiche di "*acqui-hiring*", con cui riescono ad avere accesso al personale di un'altra azienda senza realmente acquisirla (Carugati & Kar, 2024). Queste *partnership* spesso coinvolgono anche la licenza non esclusiva di determinati diritti di proprietà intellettuale in modo che i dipendenti possano continuare a lavorare sulle tecnologie e codici sviluppati senza infrangere alcuna *policy.* Importanti esempi includono gli accordi tra

[17] Ad esempio lo stipendio medio di un ingegnere di OpenAI è di 121 k€/year secondo il sito OpenAI AI Engineer Salaries | Glassdoor, concorde con quanto riportato da Work | 2024 Stack Overflow Developer Survey e Stanford (Stanford University, 2024)

[18] Si veda ad esempio il *National Artificial Intelligence Research Resource* (NAIRR) del *US National Science Foundation*

Microsoft e Inflection (Nadella, 2024) oppure Google e Character.Ai (Character.AI, 2024).

L' *Ai Index Report 2024* di Stanford, a conferma di questa grande richiesta di personale qualificato, mostra come, in generale nel mondo, il tasso *year over year* di assunzione di personale[19] con qualche forma di esperienza o conoscenza di intelligenza artificiale, sia molto alto, sopra il 10% nella maggior parte delle nazioni (Stanford University, 2024, p. 232). Hong Kong, Singapore e il Lussemburgo sono i tre Paesi che registrano il più alto valore di questo *ratio,* nonostante il primato assoluto per affluenza di talenti rimanga, ad oggi, agli Stati Uniti (MacroPolo.org, 2023). Invero, come conferma il *Global AI Talent Tracker 2.0*[20], sviluppato dall'organizzazione MacroPolo, nel 2023 gli USA continuano ad essere la principale destinazione per i ricercatori internazionali di AI di alto profilo (il *top 2%*) ed è sede del 60% delle istituzioni dedicate all'innovazione in questo settore nonostante la provenienza dei professionisti che ivi lavorano sia cinese per il 47%. La Cina ha, infatti, negli ultimi anni velocizzato gli investimenti in ricerca e sviluppo AI dando la possibilità alle università di creare un forte e dinamico *pool* di studenti, ricercatori e professori che contribuiscono sempre di più anche all'offerta nazionale interna di queste figure professionali. Anche l'India, seppur rimanendo una significativa esportatrice di *top-tier AI researchers*, sta aumentando la sua capacità di trattenere i propri talenti per allinearsi alle principali potenze mondiali.

Le dinamiche cinesi e indiane appena descritte evidenziano un definito *pattern* di staticità degli ultimi anni: solo il 42% dei *top* ricercatori al mondo sono originari di una nazione diversa da quella in cui portano vanti i propri studi, 13 punti percentuali in meno rispetto al 2019 (MacroPolo.org, 2023).

In generale, comunque, lo sviluppo di *large language models* non richiede *team* numerosi e molte *startup* sono infatti state fondate da un gruppo ristretto di esperti ottenendo in

---

[19] Il tasso viene calcolato con i dati degli annunci e delle assunzioni della piattaforma social LinkedIn, pertanto, è utile analizzare il macro-*trend* piuttosto che soffermarsi sui singoli valori nazionali che potrebbero essere falsati da un diverso grado di diffusione e utilizzo del social.

[20] La metodologia utilizzata considera i *paper* ammessi alla conferenza *Neural Information Processing Systems* (NeurIPS), la più selettiva del settore, per individuare i migliori ricercatori al mondo (Ruihan Huang, 2023). Per più dettagli si consulti il sito: Methodology for Global AI Talent Tracker 2.0 - MacroPolo

breve tempo ottimi risultati, come testimonia la storia di MistralAI, *startup* francese che pochi mesi dopo la sua fondazione ha rilasciato il suo primo modello creato da un *team* di soli 22 dipendenti[21] .

Per riassumere, riprendendo quanto commentato nel Capitolo 2.2, nonostante la crescente segretezza circa i dettagli di sviluppo sia diventata negli anni una barriera all'ingresso del mercato, la proliferazione delle piattaforme di ricerca *open source* e la dinamicità del personale qualificato mondiale non rende i *trade secrets* la miglior modalità di appropriazione dei ritorni sui risultati tecnologici (Pierre Azoulay, 2024). Inoltre, seppur per definizione costituiscono una risorsa scarsa e potenzialmente molto costosa, la presenza di comunità *open source* e la crescente attenzione universitaria al mondo AI, non rendono l'acquisizione di talenti una insormontabile barriera all'ingresso. Per creare o preservare il proprio vantaggio competitivo le aziende potrebbero, quindi, focalizzarsi sulla seconda forza evidenziata da Teece: il possesso e controllo degli *asset* complementari, l'"ecosistema", spesso ottenibile grazie ad integrazioni o *merge* (Anton Korinek, 2023).

[21] Le Monde Informatique Mistral lève 385 M€ et devient une licorne française - Le Monde Informatique, 11 dicembre 2023

Dati

I modelli fondativi vengono allenati su enormi quantità di dati con natura differente: testo, immagini, audio, video o codice che vengono appositamente raccolti, filtrati e, in generale, pre-processati al fine di eliminare eventuali *bias* o contenuti sensibili e aggiornati per evitare ogni qual forma di "*data staleness"*. Ad oggi, febbraio 2025, il modello che gode della più vasta *training pool* è Qwen2.5-72B di Alibaba, allenato su un *corpus* di 18 trilioni di *token*[22] (Epoch Ai, 2025).

Nei mercati *data driven,* le ricerche hanno evidenziato come essi siano un *asset* competitivo caratterizzato dalle "4V": volume, varietà, velocità e valore[23] (Maurice E. Stucke, 2016).

Nel 2020 un *paper* di ricerca degli ingegneri di OpenAI correlava, infatti, la dimensione del *dataset,* il suo volume, alle crescenti *performance* dei modelli dando avvio ad una vera e propria corsa verso *dataset* sempre più vasti (Jared Kaplan, 2020).

Un importante studio di Epoch.ai nel 2022 avvertiva i professionisti del settore della limitata disponibilità di dati per il futuro dell'AI e prevedeva, ad un costante tasso di crescita del volume utilizzato nel *training* e dei nuovi *data-token* prodotti, che l'AI terminasse il testo disponibile *online* entro il 2040, di cui, quello professionalmente editato entro il 2024 (Pablo Villalobos J. S., 2022). Fortunatamente trattandosi di un modello dalle forti assunzioni semplificative e di un risultato *ab origine* incerto, ciò non si è verificato. La versione aggiornata della ricerca, pubblicata a giugno 2024, stima uno *stock* complessivo di *human-generated public text* di circa 300 trilioni di *tokens*[24] che,

---

[22] Dati consultati il 27 febbraio 2025 sul sito https://epoch.ai/trends , ultimo aggiornamento 13 gennaio 2025

[23] Per una applicazione della teoria delle 4V in un caso di competizione si veda l'*assessment* dell'acquisizione di Shazam da parte di Apple redatto dalla Commissione Europea il 06/09/2018: https://ec.europa.eu/competition/mergers/cases/decisions/m8788_1279_3.pdf

[24] Questa stima, calcolata con un intervallo di confidenza del 90%, considera solamente i dati di una qualità adeguata all'allenamento dei modelli di intelligenza artificiale

seguendo il *trend* attuale, verrà completamente utilizzato all'incirca entro il 2028[25] (Pablo Villalobos A. H., 2024).

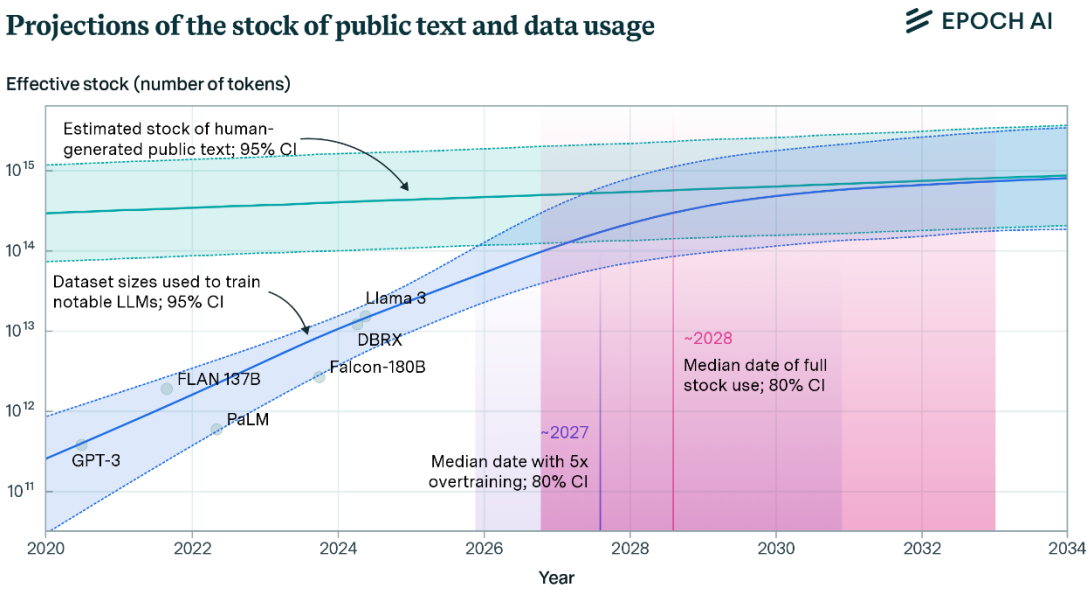


Figura 13: Proiezioni circa la quantità di dati prodotti online e utilizzati nel training dei LLMs. Source: Epoch.ai (Pablo Villalobos A. H., 2024)

I dettagli circa i *dataset* utilizzati in fase di *training* non vengono tendenzialmente pubblicati per i modelli proprietari, le cui aziende *developer* si limitano a garantire l'utilizzo coerente con le finalità di pubblicazione del *creator* originale (Pierre Azoulay, 2024). L'assenza di tali informazioni rende difficile stimare un ordine di grandezza opportuno o confrontarli con i modelli *open source.* A titolo di esempio, nella primavera del 2024, Meta ha riportato di aver utilizzato un set di circa 15.000B di *tokens* per dare origine al suo modello Llama3, sette volte più grande di quello utilizzato per il predecessore Llama2, rilasciato meno di un anno prima e avente lo stesso numero di parametri (Meta, 2024).

[25] I calcoli prevedono, con un intervallo di confidenza dell'80% una finestra temporale dal 2026 al 2032. La data esatta dipenderà dalle modalità di *scaling* dei modelli; in caso di *overtraining,* ad esempio, gli sviluppatori tendono a diminuire il numero di parametri del modello ed aumentare i dati processati , accorciando l'intervallo agli anni 2025-2030 (https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data)

La varietà dei dati è altresì importante per la natura generale del modello che deve pertanto imparare a riconoscere diverse tipologie di dati e restituire *output* differenti nel formato più efficace.

Numerosi studi, come quello condotto da Anil et. al. circa i modelli PaLM nel 2023 (Anil, 2023) dimostrano, comunque, quanto la qualità dei dati utilizzati sia al pari di volume e varietà. Un *dataset* di qualità si ottiene attraverso processi di *screening* e *cleaning* che costituiscono la fase di *pre-training,* fondamentale per le corrette prestazioni del modello ed importante elemento differenziante. Inoltre, massimizzare la qualità dei dati rispetto alla sola quantità riduce i costi computazionali rendendo l'intero processo più efficiente, come dimostrano i *developer* del Berkeley Artificial Intelligence Research del modello di dialogo Koala, *fine-tuned* a partire da LLaMa 13B in sole sei ore ad un costo inferiore ai 100$ ma comunque con ottime prestazioni (Xinyang Geng, 2023).

Con il termine "di qualità" ci si riferisce a dati che siano formalmente corretti, veritieri, oggettivi, classificabili e, soprattutto, aggiornati. La velocità, o "freschezza" dei dati rappresenta un grande ostacolo per gli sviluppatori di *foundation models*, allenati tipicamente su dati raccolti fino ad una certa data. Per ovviare alla questione i principali *developer* allenano propri modelli in maniera costante utilizzando i *dataset* più aggiornati ricavati da processi di *crawling* o di *click-and-query*[26] (Carugati, 2023) con un notevole aumento dei costi.

Infine, con il termine "valore" ci si riferisce al valore economico, estratto grazie al miglioramento effettivo delle *performance* del modello riconducendoci al discorso, già affrontato, sui *data network effects*.

Per citare la stessa OpenAI, nel descrivere lo sviluppo dei propri modelli, la società afferma che "*OpenAI's foundation models, including the models that power ChatGPT, are developed using three primary sources of information: (1) information that is publicly*

---

[26] "*Crawling*" si riferisce al processo mediante il quale i *crawler*, o *bot*, navigano automaticamente su Internet per raccogliere informazioni da diverse pagine web. Durante questo processo, analizzano il contenuto delle pagine e seguono i collegamenti ipertestuali per estrarre dati utili. Queste informazioni vengono poi indicizzate, ovvero organizzate e memorizzate in un database, permettendone un facile accesso tramite motori di ricerca.
Il "click-and-query" indica invece un metodo in cui gli utenti interagiscono attivamente con un motore di ricerca o un database per ottenere dati specifici.

available on the internet, (2) information that we partner with third parties to access, and (3) information that our users or human trainers and researchers provide or generate" (OpenAi, 2024).

Figura 14: Diverse tipologie di dataset e la loro accessibilità. Source: (CMA , 2024)

Le fonti dei *training data* sono infatti le più diverse e dipendono dalla tipologia di *task* che il modello eseguirà: alcuni richiedono dati di alta qualità da *database* curati *ad hoc,* mentre gli sviluppatori dei *foundation models* più generali ricercano un ampio bacini di dati variegati e meno strutturati. Questi dati spesso provengono da *dataset* pubblici come i dati *web-scraped*[27]*,* ma anche da fonti proprietarie o di terzi *data-providers* e dati sintetici (Carugati & Kar, 2024).

Secondo quanto riportato dalla autorità per la concorrenza francese, dalla loro inchiesta è emerso che la maggior parte dei modelli vengono allenati utilizzando *database* pubblici o contenuti direttamente accessibili *scraping online,* nonostante alcuni possano comunque essere protetti da qualche forma di tutela legale come il *copyright* (Autorité de la concurrence, 2024). Alcuni esempi sono il Protein Data Bank, un *database* utilizzato per i modelli predittivi della struttura proteica come AlphaFold o RoseTTAFold, ImageNet, tra i maggiori *database* pubblici per la classificazione di immagini, Project Gutenberg per i libri o Common Crawl, *database* curato da una

[27] Il processo di *web-scraping*. Questo processo implica l'uso di *software* o *script* di codice per raccogliere automaticamente dati da pagine web, come testi, immagini, link e altre informazioni. I dati raccolti possono poi essere utilizzati per analisi, ricerca, marketing o altre finalità.

organizzazione *no-profit* che a maggio 2024 conteneva le informazioni di oltre tre miliardi di pagine *web* (Autorité de la concurrence, 2024). Altre fonti *open source* derivano, invece, dalle *community* come i *dataset* HuggingFace OpenAssistant[28] e Amazon Massive Dataset[29].

Accanto a queste, vi si collocano le fonti classificate come "*semi-public*", ossia contenenti quei dati non esplicitamente privati, come e-mail personali o SMS, ma che esistono in una *gray-area* per quanto riguarda le aspettative e le *policy* di *privacy.* Alcuni esempi includono i *post* su *social media*, *forum* o altri siti di discussione, biblioteche di codice, recensioni di prodotti, articoli accademici o di giornale, o qualsiasi contenuto creato e caricato in rete (Pierre Azoulay, 2024). Nell'utilizzare quest'ultima tipologia di dati, i *foundation models developer* dovrebbero essere attenti nel considerarne gli aspetti legali, come *copyright* o altre leggi a tutela dei dati, considerazioni etiche e il potenziale impatto sulla propria reputazione tra gli utenti.

I modelli possono inoltre essere allenati con dati proprietari dell'azienda *developer* stessa o provenienti da terze parti attraverso contratti di licenza, partnership o acquistati, che diventano sempre più importanti verso le fasi più di *fine-tuning* per circoscrivere il modello ad un settore o tipologia di elaborazioni, il che comporta numerosi spunti di riflessione per i *policymakers* in termini di *governance* e concorrenza.

Il *framework* legale che governa l'utilizzo dei dati nei processi di allenamento delle reti neurali per l'AI generativa non è ancora stato universalmente definito (Gans, 2024) per quanto le iniziative europee siano un primo approccio al problema. Ad esempio, il *Copyright in the Digital Single Market Directive* (CDSM)[30] e il *EU AI Act*[31] prevedono il diritto per *copyright holder* di proibire l'utilizzo delle proprie opere per

---

[28] HuggingFace offre un totale di oltre 158.000 *datasets* di diverse tipologie, come, ad esempio OpenAssistant OASST1 disponibile al sito : https://huggingface.co/datasets/OpenAssistant/oasst1

[29] Dataset disponibile sulla piattaforma GitHub al link : https://github.com/alexa/massive

[30] Disponibile per la consultazione al link : https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng

[31] Disponibile per la consultazione al https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

l'addestramento di algoritmi e sistemi di *machine learning* e AI generativa (Wouter van Wengen, 2024).

Mentre le aziende sviluppatrici invocano il principio del *fair use,* ossia l'utilizzo dei dati, soprattutto quelli *semi-public,* in linea con l'obiettivo originario del *creator,* o almeno, non in contrasto e non per scopi pericolosi; questo è ampiamente contestato dai *copyright holder* attraverso una vasta serie di cause legali come quella tra il New York Times e OpenAi[32] per l'utilizzo illecito dei suoi articoli *copy-protected* o le problematiche connesse all'utilizzo del *database* Common Crawl. Fletcher, giornalista dell'Università di Oxford afferma, infatti, che il 79% di tutti i siti di notizie negli Stati Uniti hanno bloccato il *crawler* di OpenAI[33], così come anche oltre la metà dei nuovi diti *web* delle dieci nazioni più evolute (Fletcher, 2024).

Una conseguenza potrebbe essere che solamente le aziende con una forte *expertise* legale e risorse finanziarie adeguate riescano ad appropriarsi di *dataset* sufficientemente ampi e dettagliati e allo stesso tempo, gestire nella maniera più opportuna dati personali e/o sensibili escludendo le giovani *startup* dall'ingresso nel mercato. L'evoluzione delle *policy* a riguardo costituirà un elemento importante nella competizione nell'Ai generativa. I regolatori, infatti, devono bilanciare gli interessi di diversi *stakeholders* quali (i) le aziende *developer,* (ii) i *content creators* e (iii) gli *user,* intesi come intera società, e considerare il *trade off* tra i diritti di proprietà intellettuale e l'incentivo alla continua pubblicazione lato *creator* e il beneficio sociale che si genera da migliori performance dei modelli lato user, nonché il profitto delle aziende e la loro concentrazione. Pamela Samuelson della *Berkeley Law School,* in linea con i risultati del modello matematico di Joshua S. Gans[34] (Gans, 2024), evidenzia come, negli Stati Uniti, se i querelanti per i diritti di *copyright* dovessero prevalere, gli unici sistemi di AI

---

[32] Le Monde, 27 dicembre 2023 "Le « New York Times » poursuit en justice Microsoft et OpenAI, créateur de ChatGPT, pour violation de droits d'auteur" https://www.lemonde.fr/pixels/article/2023/12/27/le-new-york-times-poursuit-en-justice-microsoft-et-openai-createur-de-chatgpt-pour-violation-de-droits-d-auteur_6207946_4408996.html

[33] OpenAI, "Overview of OpenAI Crawlers" Disponibile al link: https://platform.openai.com/docs/bots

[34] (Gans, 2024) fornisce un modello economico che può essere utilizzat per valutare alcuni aspetti del dibattito e propone alcune soluzioni per bilanciare gli interessi dei *copyright holder* e gli sviluppatori dei grandi modelli di AI

generativa ad essere legali sarebbero quelli addestrati esclusivamente attraverso fonti di pubblico dominio o sotto licenza, il ché si ripercuoterebbe su coloro che integrano l'AI nei loro prodotti o la utilizzano per la ricerca scientifica o scopi privati in quanto ciò porrebbe un freno al miglioramento delle performance (Samuelson, 2023). Considerato ad esempio il caso delle riviste accademiche e della ricerca scientifica, la questione è se editori come Elsevier, Springer e Wiley abbiano il diritto di controllare e limitare l'accesso ai propri articoli ad aziende *developer* di *LLMs*, o se l'estrazione di dati al fine di *training* debba essere considerato un *fair use*. Mentre l'utilizzo di modelli di IA per riprodurre articoli protetti da *copyright* a scopo di lucro violerebbe chiaramente i principi del *fair use*, la situazione diventa più complessa se si considera il potenziale dell'IA nell'ambito della ricerca scientifica nel sintetizzare informazioni provenienti da più fonti. Tale apprendimento non riflette forse il modo in cui la scienza ha sempre progredito, "*standing on the shoulders of giants*"? Inoltre, è accettabile che editori come i sopracitati controllino il futuro dello sviluppo scientifico limitando l'accesso ad una così preziosa risorsa o, eventualmente, attraverso partnership esclusive con specifici fornitori di IA? (Hagiu & Wright, 2024)

Anche altri autori, in Europa, facendo riferimento alla causa legale tra NewsCorp e Perplexity sottolineano l'importanza di assicurare un corretto bilanciamento tra la protezione dei *rightholder* e il sostegno all'innovazione e alla competizione senza creare eccessive barriere legali (Carugati & Kar, 2024).

All'inizio dell'era della *generative AI* nel 2023, un numero sempre crescente di *content provider* hanno ristretto l'accesso ai propri dati ai *bot* per il *data-scraping* con finalità di *taining* rendendo più complesso o semplicemente più costoso il processo (Korinek & Vipra, 2024). Alcune piattaforme di *community* come Reddit e X nel 2023 hanno aumentato il prezzo delle loro API, temendo infatti che esse venissero usate per il *training* di modelli di AI (Collinas, s.d.). Per evitare rincari o azioni legali, alcuni sviluppatori stanno creando *partnership* con editori e *right holder*. Ad esempio, Google ha firmato accordi con Reddit e StackExchange[35] e OpenAI con una serie di *content creator* e di

[35] Stack Overflow blog, https://stackoverflow.co/company/press/archive/google-cloud-strategic-gen-ai-partnership , 29 febbraio 2024.

editori di stampa in diversi Paesi, tra cui Associated Press negli Stati Uniti[36] e Le Monde in Francia[37]. In generale il fenomeno delle *data partnership* è diffuso e consolidato nel settore[38] (Carugati & Kar, 2024).

La tabella sottostante elenca le principali *data partnership* annunciate da OpenAI. Il *deal* con NewCorp dovrebbe essere il più caro, ad oggi, circa 250$m per cinque anni (Alcaraz, 2024).

Content provider	Country	Agreement date
The Atlantic	United States	29/05/2024
Vox Media Inc.	United States	29/05/2024
News Corp	United Kingdom	22/05/2024
Reddit	United States	16/05/2024
Dotdach Meredith	United States	07/05/2024
Financial Times	United Kingdom	29/04/2024
Le Monde	France	13/03/2024
Prisma Media	Spain	13/03/2024
Axel Springer	Germany	13/12/2023
Associated Press (AP)	United States	13/07/2023

Tabella 3: Data partnership di OpenAI fino al 20 giugno 2024 Source: Autorité de la concurrence

Si cita per completezza di trattazione il portale "Platforms and Publishers: AI Partnership Tracker" che traccia gli accordi tra *developers* e *data providers* (o altri fornitori di diverso tipo)[39] e ne descrive la finalità.

---

[36] Associated Press, AP, https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/ , 13 luglio 2023

[37] Le Monde https://www.lemonde.fr/le-monde-et-vous/article/2024/03/13/intelligence-artificielle-un-accord-de-partenariat-entre-le-monde-et-openai_6221836_6065879.html , 13 marzo 202

[38] Ad esempio, OpenAI ha una sezione esclusiva alle richieste per partnership sui dati nel prprio sito web: https://openai.com/form/data-partnerships/

[39] Diponibile per la consultazione al link: https://petebrown.quarto.pub/pnp-ai-partnerships/

Oltre alle questioni legali, un altro elemento che può costituire una significativa barriera all'ingresso riguarda i *dataset* proprietari che contengono dati di maggiore qualità, aggiornati e, talvolta, manualmente controllati, in grado di offrire un reale vantaggio competitivo ai modelli che li utilizzano. L'accesso privilegiato ai dati rappresenta, infatti, un elemento cardine del dominio dei grandi *player* tecnologici che controllano una larga frazione dei dati raccolti online circa le interazioni sulle piattaforme o con i motori di ricerca, le e-mail, foto, video e altri documenti e che ne limitano l'accesso e l'utilizzo. Alphabet, ad esempio, oltre ad essere l'unica ad essere totalmente integrata, possedendo sia l'infrastruttura di calcolo, sia modelli proprietari (Gemini), sfrutta per il *training* la ricchezza di dati generati dal motore di ricerca Google, dal browser Chrome, da YouTube e da Google Books; Meta, grazie alle piattaforme Facebook e Instagram, dispone di un'enorme quantità di immagini e video condivisi dagli utenti, che costituiscono una risorsa preziosa per i suoi LlaMa. Microsoft, infine, attinge ai dati del motore di ricerca Bing e della piattaforma di sviluppo GitHub, per alimentare i propri modelli e migliorare i servizi offerti come MicrosoftCopilot (Autorité de la concurrence, 2024). Inoltre, la posizione delle grandi aziende digitali viene rafforzata dall'accesso preferenziale a metadati e dati di utilizzo dei loro servizi, considerati dati indiretti come la raccolta di feedback sull'esperienza utente durante la fase di inferenza che genera un circolo virtuoso e sostiene il consolidamento della loro posizione.

Aziende innovative e *startup* nel settore potrebbero incorrere in pratiche di discriminazione o *refusal of access* da parte di aziende con un accesso ai dati significativo, come un *web index.* In Europa gli sviluppatori di sistemi di ricerca possono fare affidamento al Digital Market Act, articolo 6(11)[40] (Regulatio EU, 2022) che obbliga i grandi motori di ricerca, denominati *gatekeepers,* a condividere i *search data* con i competitors (Carugati, 2023). I modelli che invece, non necessitano di dati in *real time* possono usufruire delle classiche API connesse a *dataset* proprietari o *open source.*

[40] Il testo del DMA è consultabile al link: Regolamento - 2022/1925 - EN - EUR-Lex, in particolare l'articolo in questione cita: "Il *gatekeeper* garantisce alle imprese terze che forniscono motori di ricerca online, su loro richiesta, l'accesso a condizioni eque, ragionevoli e non discriminatorie a dati relativi a posizionamento, ricerca, click e visualizzazione per quanto concerne le ricerche gratuite e a pagamento generate dagli utenti finali sui suoi motori di ricerca online. I dati relativi a ricerca, click e visualizzazione che costituisce dati personali sono resi anonimi."

Altri rischi per la concorrenza riguardano il pericolo di cartelli o abuso di posizione dominante nel momento in cui i grandi *players* stipulano accordi con i *content creator* o i grandi *data storage* a condizioni economiche proibitive per realtà più piccole al fine di escluderne l'accesso a quel determinato *asset*[41]. Questa teoria rimane comunque molto difficile da dimostrare empiricamente poiché non vi sono espliciti ostacoli per potenziali concorrenti a negoziare accordi sotto nuovi termini. Inoltre, tutto ciò pone le autorità nella complessa posizione di dover stabilire il prezzo dei dati e le relative soglie (Autorité de la concurrence, 2024).

Rimanendo in tema di *training data,* una strategia che si sta affermando nell'addestramento di modelli fondativi avanzati è l'utilizzo di dati sintetici. Tale approccio, che sostituisce dati reali con dati generati computazionalmente, mira a ridurre i costi di acquisizione dei dati e a mitigare problematiche legate a *privacy*, *copyright* e, spesso, la scarsa disponibilità di dati proprietari. Nuove tipologie di reti neurali come le GAN e le *variational autoencoders* (VAE) consentono la creazione di contenuti che emulano dati di input, modificandone dei *feature,* offrendo un'alternativa economicamente vantaggiosa, come dimostrato dall'addestramento, a Stanford, di Alpaca-7B avvenuto con dati sintetici generati con ChatGPT ad un costo di appena $600 (Rohan Taori, 2023). Data la loro recente importanza, l'esperto in AI e scrittore Alan Thompson, nel suo ultimo studio, stima che circa il 70% dei dati utilizzati per GPT – 5 sarà di origine sintetica (Thompson, 2024). Tuttavia, la validità e l'efficacia su larga scala dei dati sintetici sono ancora oggetto di studio in quanto permangono, rischi quali la propagazione di *bias* e l'incremento del tasso di errore, richiedendo un'attenta valutazione dell'impatto complessivo (CMA , 2024).

Anche il *fine tuning* avviene con dati proprietari, curati e circoscritti ad un particolare settore, azienda, utente o *task.* Inoltre*,* i dati raccolti durante la fase di inferenza, a valle, rimanendo in mano ai *deployer* o agli stessi *user* (a seconda del singolo caso e possono essere ceduti in licenza agli sviluppatori a monte con appositi contratti) costituiscono un principale *driver* dell'integrazione verticale (Anton Korinek, 2023).

[41] Ad esempio, Alphabet ha accettato di pagare $60 milioni all'anno per l'accesso ai dati di Reddit. "I contenuti del social Reddit serviranno ad allenare l'IA: accordo da 60 milioni di dollari - la Repubblica, 19 febbraio 2024

In conclusione, sebbene i dati rappresentino un input cruciale per lo sviluppo dell'AI generativa e il loro controllo rappresenta uno strumento importante con cui i grandi colossi provino a mantenere il proprio vantaggio competitivo creando modelli più performanti; è fondamentale riconoscere che la loro centralità non preclude necessariamente l'ingresso di nuovi attori sul mercato. L'emergere di modelli *open source*, la crescente disponibilità di dataset pubblici, dati sintetici e la possibilità di effettuare *fine tuning* su modelli pre-esistenti stanno progressivamente abbattendo le barriere all'ingresso (Carugati, 2023). Questi elementi, combinati con la capacità di specializzarsi in nicchie di mercato specifiche, offrono opportunità significative per aziende di minori dimensioni e per la comunità *open source*, promuovendo un ecosistema dell'AI generativa più diversificato e competitivo.

Risorse computazionali

Con il termine risorse computazionali, frequentemente abbreviato con il neologismo inglese *compute*, si intende la complessiva infrastruttura *hardware* e *software* necessaria allo sviluppo e al *training* dei *FM.* Questa include (i) l'*hardware* a supporto della potenza di calcolo, (ii) l'insieme di *software* e sistemi operativi necessari alla programmazione, (iii) l'infrastruttura di rete per facilitare il veloce trasferimento dei dati tra *computers,* (iv) l'integrazione con i servizi *cloud* che permettono una gestione flessibile del carico computazionale che assecondi le esigenze del *training* e (v) diversi *tool* per il monitoraggio e controllo dei sistemi tecnico-informatici, dei progressi dell'allenamento e dell'utilizzo di risorse e per la gestione del *deployment* e operatività del modello (Pierre Azoulay, 2024). Considerata la vasta scala, l'ottimizzazione dell'intero ecosistema, ottenuta dal bilanciamento tra le capacità *hardware*, il supporto *software* e la gestione dell'infrastruttura, risulta fondamentale per garantire un addestramento e un collaudo efficienti dei *foundation models*, con costi complessivi nell'ordine dei miliardi di dollari. La letteratura accademica di settore analizza come i *developer* dei modelli debbano, infatti, valutare il compromesso tra la dimensione del modello, la grandezza dei *dataset* di addestramento e il *budget* di calcolo disponibile (Hoffmann, et al., 2022).Queste ricerche identificano precise leggi di scalabilità, "*scaling laws*", riprendendo la teoria della legge di Moore e stabiliscono la correlazione positiva tra dimensioni e complessità,

performance e risorse computazionali, e quindi, finanziarie necessarie al *training* (Kaplan, et al., 2020). Queste regolarità sono molto importanti per le aziende *developer* in quanto permettono di ridurre l'incertezza di fronte alle decisioni di investimento. Per meglio chiarire l'ordine di grandezza a cui ci si riferisce, il CEO di Meta, Mark Zuckerberg, in un comunicato stampa di inizio 2024 ha annunciato che, nel suo impegno per addestrare la successiva generazione della famiglia Llama[42] e competere nella corsa per l'AGI, *artificial general intelligence,* l'azienda sta ingrandendo la sua infrastruttura di calcolo con l'acquisto di 350.000 GPU H100 di Nvidia, che, in aggiunta a quelle già presenti nei *server* Meta, porterebbe il totale a 600.000 GPU entro la fine dell'anno. Valutato al prezzo di vendita al pubblico (dai 30 ai 70 mila dollari), l'investimento in GPU da solo si aggirerebbe attorno ai 10 miliardi di dollari[43]. Cifre simili hanno riguardato gli investimenti di Microsoft, in *partnership* con OpenAI, per lo sviluppo di nuovi *supercomputers* (285.000 CPUs e 10.000 GPUs (Black, 2020)) a sostegno di Microsoft Azure, pari a $1B nel 2019[44] e aggiuntivi $10B nel 2023 (Bass, 2023). Nonostante l'apertura del modello Llama, che rimane, come i precedenti, *open source*, Meta eserciterebbe comunque un rigoroso controllo sulle risorse computazionali e, di conseguenza, su come il modello verrà addestrato nelle future iterazioni.

In linea con quanto teorizzato dalle leggi di scala, i ricercatori di Epoch AI hanno stimato una crescita complessiva di 4,1x/*year* nelle risorse computazionali (misurate in FLOP[45]) utilizzate negli ultimi 15 anni per lo sviluppo dei principali modelli di *machine learning* e *deep learning* (epoch, 2024). Tuttavia, se si considerano esclusivamente i modelli rilasciati a seguito dell'introduzione del Transformer nel 2017, la tendenza per i LLMs

---

[42] A gennaio 2024, il CEO di Meta si riferiva in particolare a Llama 3, uscito nella primavera seguente e che ha visto il rilascio di altri due aggiornamenti nel corso dello stesso anno. A dicembre 2024, la *release* dell'ultima versione Llama 3 (https://www.rivista.ai/2024/12/06/meta-lancia-il-nuovo-modello-ai-llama-3-3-piu-economico-piu-leggero-e-competitivo-con-i-big-del-settore/)

[43] https://www.theverge.com/2024/1/18/24042354/mark-zuckerberg-meta-agi-reorg-interview

[44] https://openai.com/index/microsoft-invests-in-and-partners-with-openai/

[45] I FLOP, acronimo di "*Floating Point Operations Per Second*" (operazioni in virgola mobile al secondo), rappresentano l'unità di misura della potenza computazionale di un sistema, misurando il numero di operazioni eseguite in un secondo. Ad esempio, un sistema con una potenza di calcolo di 1 teraflop è in grado di effettuare 1 trilione di operazioni in virgola mobile al secondo. I FLOP consentono di confrontare in modo efficace le capacità di calcolo di diverse architetture hardware.

cresce in modo significativamente più rapido aumentando del 9,5x/*year* nel periodo 2017-2024[46]. Questa tendenza potrebbe riflettere il fatto che i modelli di linguaggio sono partiti da un livello inferiore rispetto ai modelli di ML in essere, e hanno scalato la frontiera di settore dopo aver acquisito popolarità, per poi subire una decelerazione a partire dal 2020. Gli ultimi 4 anni hanno infatti registrato una crescita di *compute* impiegata del 5,0x/*year*. Data la concentrazione di imprese che posseggono in via proprietaria questa infrastruttura si analizzano i *trend* delle principali *leader:* Google DeepMind, OpenAI e Meta. I risultati evidenziano un tasso di crescita maggiore per quest'ultima anche se, la rapida evoluzione del mercato e gli investimenti recenti potrebbero aver modificato questa classifica risalente a maggio 2024. Di seguito i grafici per miglior chiarezza.

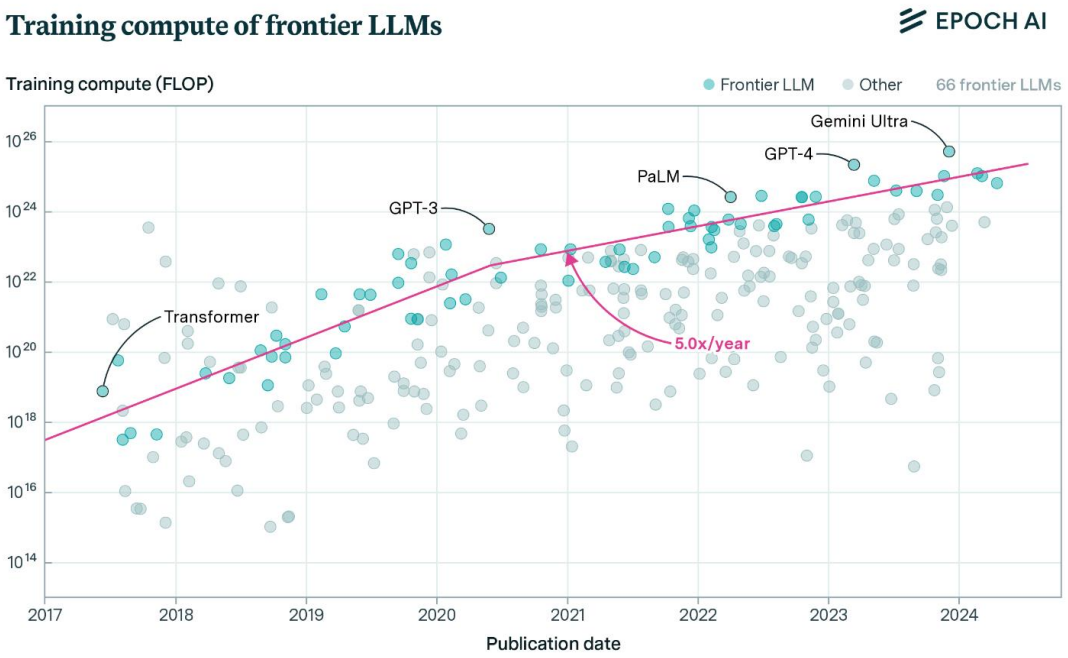


Figura 15: Crescita nella potenza computazionale utilizzata per i LLMs. Decelerazione evidente dal 2020. Source: EpochAI

[46] Con un intervallo di confidenza del 90%

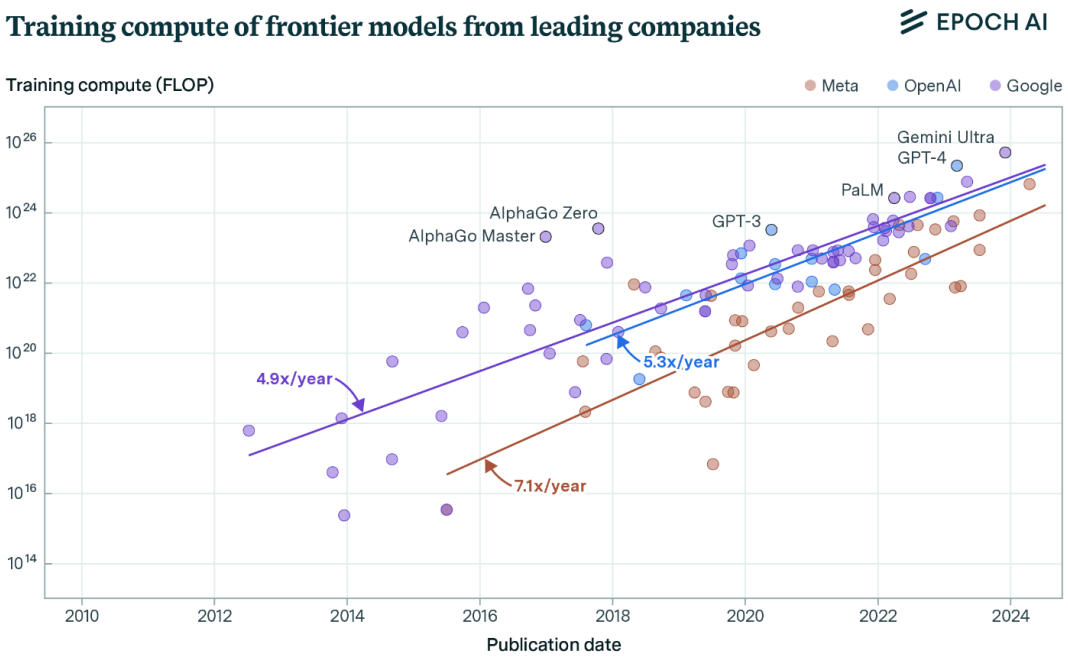


Figura 16: Compute delle maggiori aziende svilupparici di FM, OpenAi, Meta e Google DeepMind. Source: EpochAi

Tenendo in considerazione la progressiva riduzione dei prezzi dei *chip,* la cui evoluzione segue anch'essa la legge di Moore, si è calcolato che il tasso di spesa computazionale per i modelli di frontiera è cresciuto solamente del 2,4x/*year;* prendendo come riferimento il costo ammortizzato dell'*hardware* utilizzato e dell'energia, e del 2,5x/*year,* considerando il costo di affitto medio per l'infrastruttura in *cloud* (Cottier, 2024). Questa ultima metodologia di calcolo è più semplice in quanto i *cloud provider* definiscono a priori il prezzo sulla base del costo energetico e di un costo fisso *chip-hour;* tuttavia, il risultato di questa stima in valore assoluto è circa il doppio della precedente in quanto, nella realtà, molti dei modelli di frontiera sono sviluppati con infrastruttura proprietaria che, non includendo il margine per il *cloud provider,* ne abbatte i costi. Il *trend* attuale è molto importante per i *policymaker* in quanto stima, se il *trend* prosegue, il raggiungimento di un costo computazionale per modello di oltre un miliardo di dollari entro il 2027, prezzo che solo le organizzazioni più facoltose potranno permettersi. Di seguito i costi di *training* stimati per le principali aziende e, in Annex C il *breakdown.*

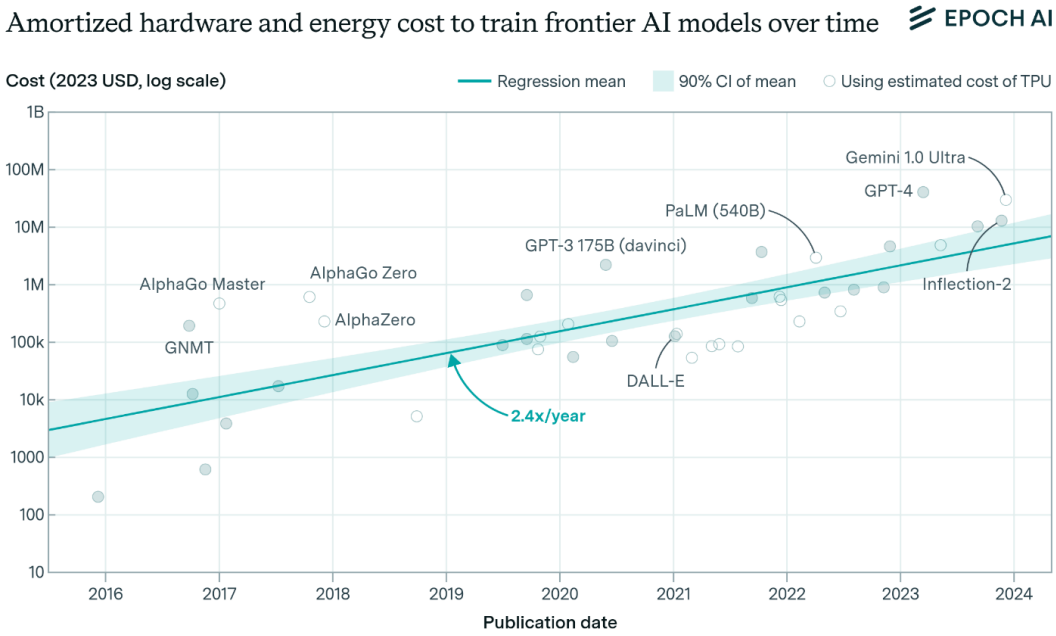


Figura 17: Costo di training simato per i principali FM di frontiera[47]*. Source EpochAi*

Sono necessarie due importanti considerazioni. In primis le analisi sopra descritte si riferiscono a stime, e pertanto sono vittime di errori dovuti alla non divulgazione dei dettagli tecnico-economici dei modelli. In particolare, si riferiscono ai costi per l'ultima fase di *training* del modello più complesso di quella precisa famiglia, non considerando le prove e gli esperimenti precedenti necessari per arrivare ad una architettura ottimale (Heim, 2021). I modelli rilasciati al pubblico, infatti, includono diverse varianti che differiscono principalmente per il numero di parametri e la dimensione del dataset di allenamento. Ad esempio, GPT-3 è stato pubblicato in otto varianti, Llama1 in quattro, moltiplicando i costi. Inoltre, il processo di sviluppo richiede diverse prove ed esperimenti prima di raggiunge l'architettura ottimale e il modello rilasciato è il frutto di un lungo, e costoso, processo di selezione (Autoridade da concorrencia, 2023). Le ricerche sui costi rimangono, comunque, fondamentali per individuare i *trend* e l'ordine di grandezza in questione, necessari a politici e regolatori per comprendere il mercato e

[47] I modelli selezionati rientrano tra i 10 più intensivi in termini di calcolo nel periodo. I costi dell'*hardware* sono calcolati come il prodotto delle ore di utilizzo dei *chip* di addestramento e di un costo di ammortamento, con un sovraccarico del 23% aggiunto per l'infrastruttura di rete. I cerchi aperti indicano costi che si basano su un costo di produzione stimato per l'hardware TPU di Google, il cui valore non è divulgato ufficialmente. Questi costi sono generalmente più incerti rispetto agli altri, che si fondano su dati di prezzo reali piuttosto che su stime (Cottier, 2024).

meglio quantificare possibili investimenti pubblici e/o a sostegno di *starup*. In secondo luogo, va precisato che il *trend* di crescita proseguirà esclusivamente se i benefici economici dei *foundation models* continueranno a scalare in linea con i costi (o più) (Korinek & Vipra, 2024). Nel ciclo di sviluppo di una innovazione è comune che inizialmente la tecnologia cresca più velocemente del mercato in quanto parte da zero; ma nel medio lungo termine nessun settore cresce più velocemente dell'intera economia. Ciò implica che ci si può aspettare una decelerazione del tasso di crescita dei costi di *training* man mano che l'AI generativa occuperà un ruolo sempre più significativo nella società. Esperti di settore come Suleyman, CEO di Microsoft AI e *Co-Founder* di DeepMind e Inflection (2023) e altri analisti prevedono comunque un *trend* costante per i prossimi 3-5 anni e, date le ottime premesse della tecnologia, anche qualche anno in più. In linea con questa idea anche le stime di Goldman Sachs sull'impatto dei futuri investimenti *hardware* in rapporto al GDP statunitense (Goldman Sachs, 2025).

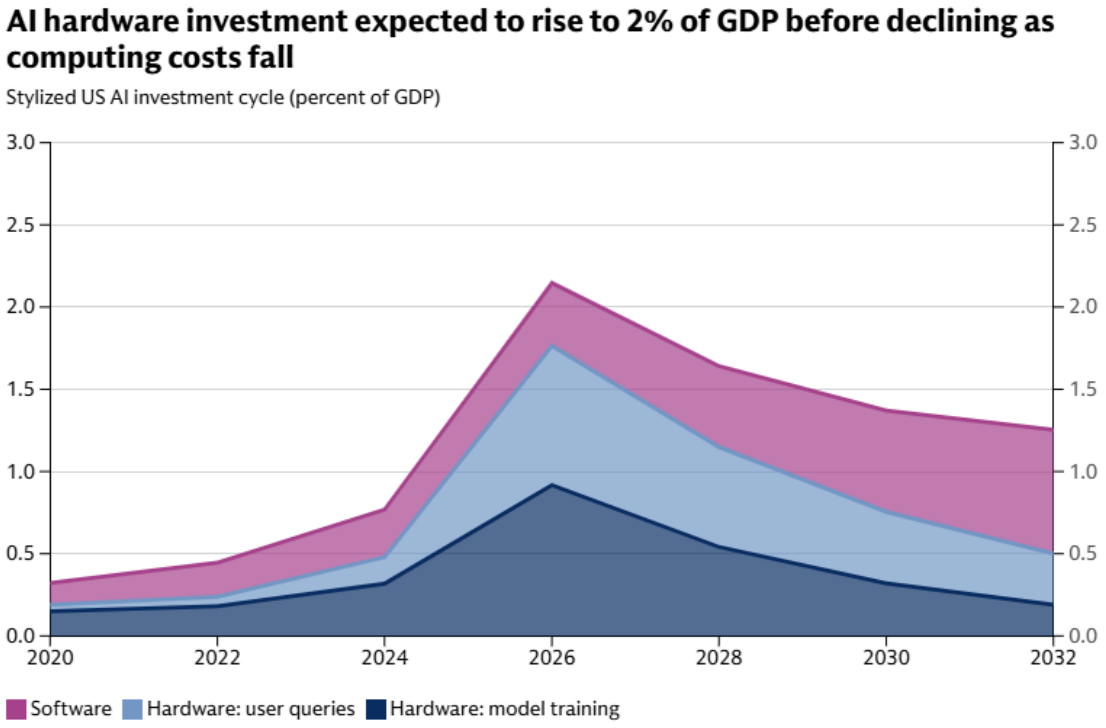


Figura 18: Percentuale degli investimenti Capex per l'infrastruttura AI del GDP statunitense. Source:Goldman Sachs Research

La recente crescita della domanda computazionale trainata non solo più dalle aziende private ma anche dalle istituzioni pubbliche che mirano a creare domini tecnologici nazionali potrebbe comunque mantenere spingere i prezzi verso l'alto. Ulteriori *driver* di prezzo sono la limitata e concentrata offerta mondiale di *chip* e semiconduttori e le strategie competitive dei *cloud provider.*

L'infrastruttura di calcolo è accessibile al pubblico e agli sviluppatori di *FM* attraverso i servizi *cloud* che si inseriscono subito al di sopra nella filiera di sviluppo. *FM* e *cloud provider* sono interdipendenti in quanto i primi necessitano dell'infrastruttura *cloud* per allenare, utilizzare e distribuire i loro modelli e applicazioni, e i secondi considerano la GenAI un motore di crescita importante per ampliare la loro quota di mercato nel settore *cloud computing* e nei mercati correlati tra cui quello *software*, motori di ricerca, *browser* o il settore pubblicitario (Carugati, 2023b). Questo *layer* è dominato da tre grandi colossi tecnologici, chiamati in questo contesto "*hyperscalers*", Amazon Web Services (AWS), Microsoft Azure e Google Cloud Platform, ma include anche *player* minori come DGX Cloud di Nvidia, IBM Cloud, Alibaba Cloud e CoreWeave Cloud, una recente *startup* partecipata da Nvidia e quotatasi in borsa lo scorso venerdì 22 marzo 2025[48] rappresentando la più grande *tech IPO* americana dal 2021, a voler sottolineare l'importanza del *cloud* nell'attuale *tech economy*[49]. I maggiori *cloud provider* dispongono, spesso, anche dell'infrastruttura sottostante e pertanto sono i primi ad investire in intelligenza artificiale tramite lo sviluppo di *foundation model* proprietari o stringendo *partnership* strategiche con *developer* terzi affinché implementino modelli e applicazioni sulla loro infrastruttura e nei loro servizi, come prevedono gli accordi tra Microsoft e OpenAi o tra Amazon/Google e Anthropic.

Nel contesto AI gli *hyperscalers* offrono una grande varietà di servizi suddivisi in tre macrocategorie:

---

[48] https://www.cnbc.com/2025/03/28/coreweave-starts-trading-on-nasdaq-at-per-share.html

[49] Secondo un recente report di Statista.com le aspettative di crescita del settore *cloud* stimano un mercato da €176B entro il 2027 https://www.statista.com/forecasts/1235161/europe-cloud-computing-market-size-by-segment

i. ***Infrastructure as a Service*** **– IaaS**

Modalità con cui il *cloud provider* fornisce direttamente, ed in maniera flessibile e scalabile, la potenza computazionale richiesta dal *developer* a seconda delle necessità permettendo di customizzare molte delle caratteristiche tecniche dell'architettura informatica sottostante. In questo modo il *developer* riesce ad avere le risorse necessarie e personalizzate senza dover investire in infrastruttura. Esse si adattano automaticamente al carico di lavoro, che sarà elevato e costante durante le settimane di *training* e più variabile una volta che il modello viene implementato e utilizzato dal pubblico.

Numerosi sono gli accordi di *partnership* tra *FM developer* e fornitori *cloud* che regolano l'utilizzo di una data infrastruttura. Citata più volte, la *partnership* tra Microsoft e OpenAI rientra in questa categoria; Microsoft ha, infatti, investito oltre $11B in una *partnership* esclusiva con OpenAI volta a sostenere lo sviluppo delle tecnologie *hardware* di Microsoft Azure dedicate ad OpenAI (Bass, 2023).

L'esclusività di queste "*compute partnership",* tuttavia, può sfociare in pratiche anticoncorrenziali nel momento in cui si limiti la possibilità al *developer* di cambiare fornitore o si verifichi un effetto *lock-in* troppo stringente a causa di *switching cost* troppo elevati (Carugati & Kar, 2024). Collaborazioni non esclusive, per contro, abbattono i rischi competitivi incentivando un sano e dinamico ecosistema, consentendo a *cloud provider* di ospitare diversi *developer* e garantendo a questi ultimi di avere più potere negoziale. Alcune di queste *partnership* non esclusive comprendono AWS con Hugging Face nel 2023[50], AWS con Anthropic nel 2023[51] , Google Cloud

[50] AWS, 'AWS and Hugging Face collaborate to make GenAI more accessible and cost efficient', 21 February 2023, https://aws.amazon.com/blogs/machine-learning/aws-and-hugging-face-collaborate-to-makegenerative-ai-more-accessible-and-cost-efficient/

[51] Amazon, 'Amazon and Anthropic Announce Strategic Collaboration to Advance GenAI', 25 September 2023, https://press.aboutamazon.com/2023/9/amazon-and-anthropic-announce-strategic-collaboration-to-advancegenerative-ai

con Anthropic nel 2023[52] o la più recente *partnership* tra Microsoft e MistralAI nel 2024[53]. In questo contesto i *FM developer* sono, quindi, i clienti dei *cloud provider*.

ii. *Platform as a Service* – PaaS

I *cloud provider* che offrono soluzioni PaaS consentono ai loro clienti di accedere ad un ecosistema di modelli AI, *software* e strumenti *ready-to-use,* necessari per lo sviluppo delle loro applicazioni, quali *database*, *tool* di *data analytics* o di *data preparation*, strumenti per il *fine tuning* o il *deployment*. Con questa modalità di accesso i clienti possono quindi usufruire della potenza computazionale di cui hanno bisogno ma senza la libertà di personalizzazione dell'infrastruttura fisica sottostante, a differenza del caso precedente.

In questo contesto i *FM developer* mettono a disposizione i propri modelli ai clienti del *cloud provider* attraverso la sua piattaforma affinché possano utilizzare e sviluppare applicazioni di intelligenza artificiale. Ad esempio, OpenAI fornisce i suoi modelli di *machine learning* su Microsoft Azure tramite il servizio Azure OpenAI[54]. AWS, da parte sua, offre accesso a vari modelli di AI21 Labs, Anthropic, Cohere, Meta, Stability AI e Amazon attraverso il servizio AWS Bedrock[55].

In questo scenario la relazione tra *foundation model* e *cloud provider* è la classica delle piattaforme digitali. Il fornitore di servizi *cloud*, attraverso la sua piattaforma *marketplace* mette in comunicazione i fornitori di modelli e utenti finali con l'importante dettaglio che, in questo mercato, i principali *cloud provider* sono anch'essi fornitori di modelli.

---

[52] Lizette Chapman, Katie Roof and Julia Love, 'Google Bets $2 Billion on AI Startup Anthropic, Inks Cloud Deal', Bloomberg, 27 October 2023, https://www.bloomberg.com/news/articles/2023-10-27/google-to-invest-2-billion-in-ai-startup-anthropic-wsj-says#xj4y7vzkg

[53] Microsoft, https://azure.microsoft.com/en-us/blog/microsoft-and-mistral-ai-announce-new-partnership-to-accelerate-ai-innovation-and-introduce-mistral-large-first-on-azure/ 26 February 2024.

[54] https://azure.microsoft.com/en-us/products/ai-services/openai-service

[55] https://aws.amazon.com/it/bedrock/

Provider di *cloud* e altri fornitori di *FM,* sulla stessa piattaforma competono per la stessa clientela, comportando potenziali rischi competitivi, come meccanismi di *self-preferencing* o comportamenti scorretti circa i dati raccolti dagli sviluppatori finali. Ad esempio, la piattaforma Amazon Bedrock potrebbe utilizzare i dati di acquisto e *deployment* dei diversi modelli per avvantaggiarsi rispetto ai concorrenti oppure promuovere i propri strumenti a discapito di quelli di terzi limitandone l'accessibilità. (Carugati, 2023b).

iii. ***Software as a Service* – SaaS**

Con questa modalità di accesso cloud, i clienti finali possono utilizzare online i software e le applicazioni offerti dal cloud provider senza (o quasi) possibilità di modifiche. I developer di FM abilitano, quindi, i cloud provider a integrare i loro modelli all'interno di applicazioni SaaS proprietarie da offrire agli utenti finali. Ad esempio, Microsoft integra le soluzioni di OpenAI nel proprio software di produttività, Microsoft 365 Copilot, nel browser web Microsoft Edge, nel motore di ricerca Microsoft Bing e nel sistema operativo Microsoft Windows Copilot (Shaw, 2023). Analogamente, Google adotta una strategia simile con il suo software di produttività, Google Workspace, e il suo motore di ricerca, Google Search implementando Gemini (CMA , 2024). Alcuni SaaS operano come piattaforme che permettono a sviluppatori terzi di creare applicazioni che completano e interagiscono con i loro servizi. A tal proposito, Microsoft consente a sviluppatori esterni, come OpenTable ed Expedia, di realizzare plugin che interagiscono con ChatGPT, Microsoft Bing, Microsoft 365 Copilot e Microsoft Windows Copilot (Carugati, 2023b). Queste applicazioni testimoniano come i *foundation model* possano essere *fine-tuned* e integrati all'interno dei presenti ecosistemi *software* per migliorarne la produttività, creatività e processi decisionali tra diversi settori. Inoltre, sottolineano come, spesso, la linea di demarcazione tra *FM* e il suo *deployment* può essere molto sfumata (Hagiu & Wright, 2024).

In questo scenario, i fornitori di *cloud* e di *FM* risultano verticalmente integrati anche con il *layer* di *deployment* a valle, il che comporta il rischio di pratiche di *tying* e *bundling* e di abuso di posizione dominante o *leveraging*

che possono negativamente influenzare la competitività del mercato e la libertà di scelta degli utenti (Carugati, 2023) .

Va comunque sottolineato che queste tre strutture, nel caso della relazione *cloud - foundation model* coesistono e si sfumano l'una nell'altra, talvolta assumendo anche altre denominazioni più specifiche come *Model as a Service* – MaaS, utilizzato per sottolineare come il modello sia utilizzato come base per lo sviluppo di applicazioni, sia direttamente del *cloud provider* o dei clienti della piattaforma (Autorité de la concurrence, 2024).

Il molteplice ruolo del *cloud* nello sviluppo tecnologico dell'intelligenza artificiale lo rende un elemento sempre più importante in ambito investimenti e sviluppo ottenendo grande attenzione e finanziamenti da parte del settore pubblico e privato. Secondo l'ultima proiezione di Gartner, le aziende globali investiranno $723,4B in servizi *public cloud* nel 2025 contro i $595,7B spesi nel 2024, pari a un incremento del 21,5%, principalmente trainato dalla modalità IaaS (Licata, 2024).

A differenza del *layer* dei dati o dei modelli in sé che rimangono, ad oggi, due mercati competitivi, il mercato dei servizi *cloud* ha una alta probabilità di proseguire la tendenza di continuo consolidamento nelle mani di pochi *player*; più precisamente AWS, Google Cloud e Microsoft Azure. La ragione principale risiede nelle ingenti economie di scala e degli alti costi di sviluppo, imputabili principalmente all'infrastruttura miliardaria di GPUs necessaria allo sviluppo dei modelli di intelligenza artificiale (Hagiu & Wright, 2024).

Le maggiori preoccupazioni riguardo la concentrazione dei *cloud provider* è che possano fare leva, *leveraging,* della propria posizione dominante in un settore negli altri della *value chain.* Infatti, gli *hyperscaler,* insieme alle altre due *big tech,* sono coinvolte in tutti i livelli:

- Tutte le aziende stanno attivamente sviluppando dei *chip* proprietari per ridurre la dipendenza da *provider* terzi. In particolare, Google guida gli sforzi con le sue *Tensor Processing Units* (TPUs)[56]
- Tutte hanno accesso a enormi quantità di dati proprietari, sebbene non necessariamente unici
- Tutte, ad esclusione di Apple, hanno sviluppato *foundation models* di frontiera in vari domini, oltre che agli investimenti in altre aziende e *start up* di settore
- Tutte sono attive a livello *application* con sistemi *chatbot,* assistenti virtuali, *plug-in software*, etc…

L'abilità con cui queste aziende riusciranno a traslare il loro dominio dal settore cloud agli altri segmenti dell'AI *stack,* dipenderà dalla loro abilità di strutturare strategie di integrazione verticale, *lock in, tying, bundling, exclusive dealing* e limiti all'interoperabilità per aumentare gli *switching cost.* Tutte pratiche ben monitorate dalle autorità garanti per la concorrenza in diverse nazioni. Lo scorso 28 gennaio 2025, ad esempio, l'autorità garante inglese, la CMA, ha rilasciato le prime e provvisorie conclusioni circa una profonda indagine di mercato in campo *cloud services* che ha evidenziato le alte barriere all'ingresso nel settore e un alto livello di concentrazione e ha sollevato preoccupazioni riguardo a presunti comportamenti anticoncorrenziali da parte di Microsoft e AWS. La discussione continuerà fino all'agosto 2025, mese in cui è prevista la decisione definitiva dell'ente (CMA, 2025).

Ci si sarebbe potuti aspettare che, anche in assenza di comportamenti di *leveraging* espliciti, la forza degli *hyperscalers* in termini di potenza di calcolo e accesso ai dati li posizionasse naturalmente come leader a livello di *foundation model*. Tuttavia, tutte e cinque le grandi aziende sono state lente nell'affermarsi nel settore soprattutto se confrontate con startup come OpenAI e Anthropic. Tre i fattori chiave che si ritengono essere alla base di questo fenomeno. In primo luogo, le GAMMA (Google, Amazon, Meta, Microsoft e Apple), impegnate con le loro molteplici *business unit*, non si sono da subito focalizzate sullo sviluppo di nuovi *FM*, diversamente da alcune *startup* pioniere. In secondo luogo, i vantaggi di cui godevano in termini di *compute*, talenti e dati sono

[56] https://cloud.google.com/tpu?hl=it

stati parzialmente erosi dalle *startup* e dai loro investitori, disposti a effettuare investimenti significativi per acquisire tali risorse. Infine, la minaccia di regolamentazione e antitrust rivolta alle grandi aziende tecnologiche (ad esempio i vincoli del DMA riguardo i *gatekeepers* e il GDPR per la *privacy*) potrebbe averle dissuase dall'utilizzare in modo aggressivo i dati proprietari ottenuti da altri servizi per affermarsi tra i primi *foundation model*.

Alla base della filiera dell'AI generativa troviamo, come detto, i *chip* e *l'hardware* per il calcolo.
Il *processing hardware,* costituente l'infrastruttura *cloud,* include le *central processing unit* (CPUs), *chip* per l'interpretazione e l'esecuzione dei calcoli, le *graphic processing unit* (GPUs), per la parallelizzazione delle operazioni, la *random access memory* (RAM), per memorizzarne i risultati intermedi e le *tensor processinng unit* (TPUs) necessari ad accelerare complessivamente l'elaborazione (Carugati, 2023). L'*hardware* di elaborazione così formato costituisce i *server,* i *super-computer* che, essendo sede del calcolo, permettono l'allenamento e l'utilizzo del modello, memorizzandone dati e parametri. L'importanza strategica dei *server* è testimoniata dagli enormi investimenti nel 2019 e 2023 di Microsoft e OpenAi già menzionati che hanno l'obiettivo di ampliare lo *stock* di *server* di Microsoft Azure.

Ad oggi, i *server* adibiti all'allenamento dei grandi modelli di AI sono composti per il 51% da GPUs (Global Market Insight, 2024) di cui il fornitore globale per eccellenza è Nvidia che, nel 2024, registra oltre il 90% di *market share* nel settore AI (Qi, 2024). È comunque importante notare che nonostante gli sforzi delle *big tech* nella ricerca e produzione di nuovi *chip* per l'allenamento di grandi LLMs, la maggior porzione delle vendite di Nvidia deriva proprio dagli stessi *hyperscalers* (Hagiu & Wright, 2024). Il *business model* di Nvidia prevede la progettazione e la vendita di questi *chip*, mentre la produzione è affidata ad importanti fonderie di semiconduttori, principalmente: Taiwan Semiconductor Manufacturing Company (TSMC) e, in misura minore, Samsung Electronics. Oltre all'hardware, Nvidia ha consolidato la sua posizione nell'industria dell'AI fornendo il *framework software* più popolare per l'utilizzo delle GPU, noto come *Compute Unified Device Architecture* (CUDA). CUDA consente agli sviluppatori di sfruttare la potenza di elaborazione parallela delle GPU per il *general-purpose computing*, che è fondamentale per lo sviluppo di *FM* e applicazioni. Essendo diventato

lo *standard de facto* nel mercato, CUDA crea attorno ai *chip* di Nvidia, effetti di rete simili a quelli creati da un sistema operativo: maggiore è il numero di sviluppatori che lavorano e creano *software* per CUDA, più aumenta l'attrattività della *Big Tech* americana. In generale, ciò crea significative barriere all'ingresso a favore di Nvidia nel mercato dei *chip* per AI e ne aumenta il potere di mercato.

Gli alti margini del settore e la necessità dei governi, principalmente americano ed europei, di creare una indipendenza tecnologica dalla Cina[57], ha portato ad una abbondanza di sussidi pubblici[58] (Stanford University, 2024) che potrebbero spingere il mercato verso una maggior diversificazione nella *supply chain* dei semiconduttori e, di conseguenza, abbattere i costi per il *training* dei modelli (Korinek & Vipra, 2024). Tuttavia, ad oggi, nonostante siano numerose le *start up* attualmente impegnate nella progettazione e produzione di nuovi *chip,* il mercato rimane concentrato nelle mani di Nvidia e crescenti sono le preoccupazioni circa un suo *leveraging* in altri settori o per potenziali comportamenti escludenti o discriminatori (Autorité de la concurrence, 2024). La società potrebbe infatti, per mantenere la propria posizione monopolista, stringere contratti che le impongano l'esclusività, accordarsi con prezzi strategici come sconti o *rebates,* oppure potrebbe sfruttare gli effetti di rete del suo *software* CUDA ottimizzandolo per il proprio *hardware* e limitandone l'interoperabilità (Hagiu & Wright, 2024). Tutte azioni, sotto l'attento scrutinio delle autorità. Come esempio di *leveraging,* la società sta espandendo la sua presenza negli altri livelli della catena del valore con una sua prima famiglia di modelli di fondazione *open source* NVLM e una relativa piattaforma di sviluppo Cosmos[59].

E' altresì importante evidenziare che, spesso, gli sforzi di una azienda dominante in un settore nel traslare in un altro adiacente non sono spinti dalla volontà di affermarsi nel

---

[57] In risposta agli investimenti americani la Cina ha allocato $47B nella ricerca e nello sviluppo nazionale di semiconduttori e *chip* https://forbes.it/2024/05/27/cina-creato-nuovo-fondo-47-miliardi-sviluppo-chip/

[58] In America, il *CHIP and Science Act* del 2022 si pone l'obiettivo di supportare l'industria americana domestica dei semiconduttori allocando capitali per oltre $50B https://www.pwc.com/us/en/library/chips-act.html. In Europa, l'*European Chip Act,* entrato in vigore nel 2023, imposta un *budget* di circa €43B finalizzato allo sviluppo di impianti produttivi e all'ottimizzazione delle partnership tra governi, industrie e le università per la ricerca https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-chips-act_en

[59] https://www.nvidia.com/en-us/ai/cosmos/

secondo quanto più da ragioni difensive mirate a ridurre la dipendenza da altri *player* (Autoridade da concorrencia, 2023).

Sulla base dell'analisi condotta, si stimano almeno sette attori principali a più livelli della *value chain*, tra cui le sei grandi GAMMAN (Amazon, Apple, Google, Meta, Microsoft e Nvidia) e OpenAI. Rispetto all'elevata concentrazione osservata in altri servizi digitali come i *search engine,* dove Google è quasi monopolista o le piattaforme di *e-commerce* con Amazon, ecc… un mercato con sette concorrenti ben finanziati, unitamente ad imprese specializzate in specifici *layers*, rappresenta un significativo miglioramento. Inoltre, un'applicazione efficace delle leggi sulla concorrenza potrebbe favorire l'ingresso di nuovi entranti, nonostante le economie di scala continuino a generare una certa concentrazione nei servizi cloud (Hagiu & Wright, 2024)

La figura sottostante raffigura le principali relazioni tra le GAMMAN, e i *FM developer*[60].

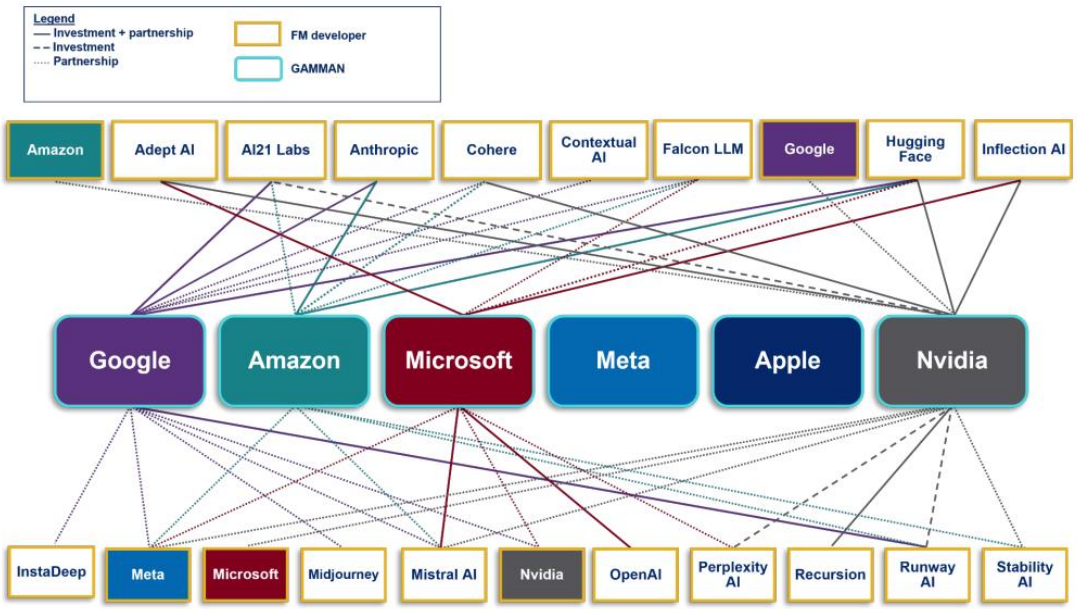


Figura 19: relazioni tra le GAMMAN e i developer di FM. Source CMA, 2024

[60] Per ulteriori dettagli si consulti il "AI Foundation Models technical update report", rilasciato dal CMA nell'aprile 2024

Nella prossima figura, invece, per chiarezza, sono descritti i principali player del settore e collocati lungo la *value chain.* Ulteriori dettagli nell'Annex D.

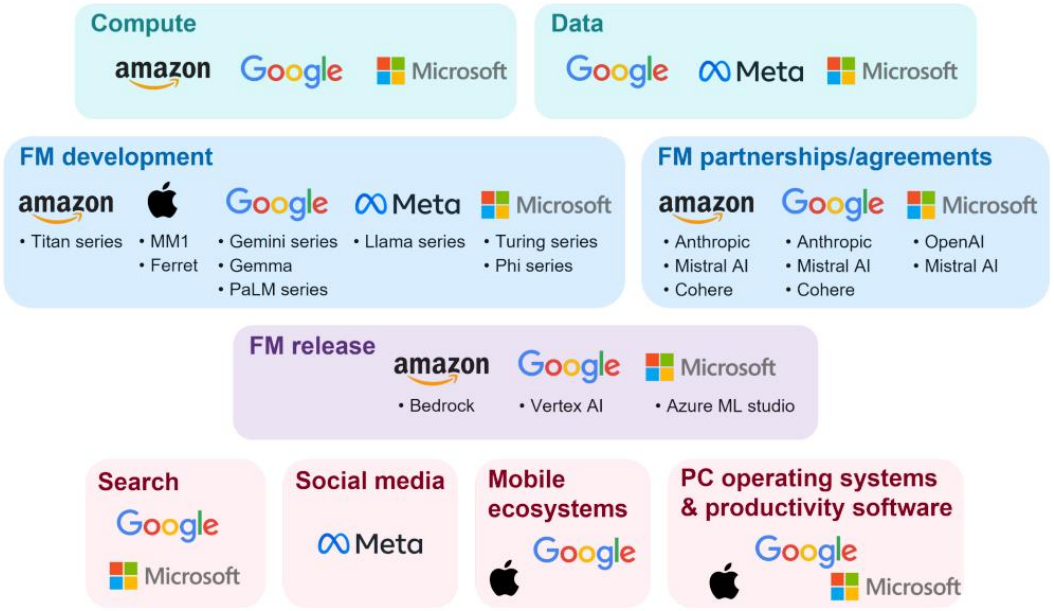


Figura 20: Illustrazione del collocamento delle aziende GAMMAN nella value chain della Gen.AI. Source: CMA,2024

In conclusione, riprendendo il discorso di Teece, i vantaggi competitivi e la dominanza in un settore possono essere ottenuti attraverso il controllo, più che sul prodotto di riferimento, sugli asset complementari necessari alla sua produzione o commercializzazione. La teoria resta valida anche nel settore dei *foundation models* dove, non potendo fare leva sull*'appropriability*, i grandi *player* si concentrano sulla crescita nei settori adiacenti e di supporto. Questi *asset* complementari non riguardano solamente i dati e l'infrastruttura computazionale ma anche tutto l'ecosistema di *benchmarking* e *safety*: imporsi sul mercato come *standard de facto*, impegnarsi nella standardizzazione di indici e/o processi e avere le risorse finanziare e di *know how* per investimenti in *governance* e *privacy*, sono tutte azioni importanti per acquisire notorietà e autorità nel settore che contribuiscono, come un effetto di rete, ad attrarre più utenti o sviluppatori (Pierre Azoulay, 2024).

Quindi, seppur il settore foundation model in sé risulta, ad oggi, abbastanza competitivo con numerosi player che si sfidano quotidianamente in termini di diffusione e

innovazione, i settori adiacenti lungo la filiera dell'AI non lo sono altrettanto. Le attenzioni dei regolatori dovrebbero quindi maggiormente concentrarsi in questa direzione, continuando, comunque, a monitorare attivamente il mercato e la sua evoluzione.

Il prossimo capitolo, a conferma di questa tesi, analizza il settore in termini di dinamiche economiche focalizzandosi sul ruolo dei dati e dei feedback loop e su come questi possano o meno costituire un vantaggio competitivo per gli incumbent in grado di ridurre la contestabilità del mercato dei FM.

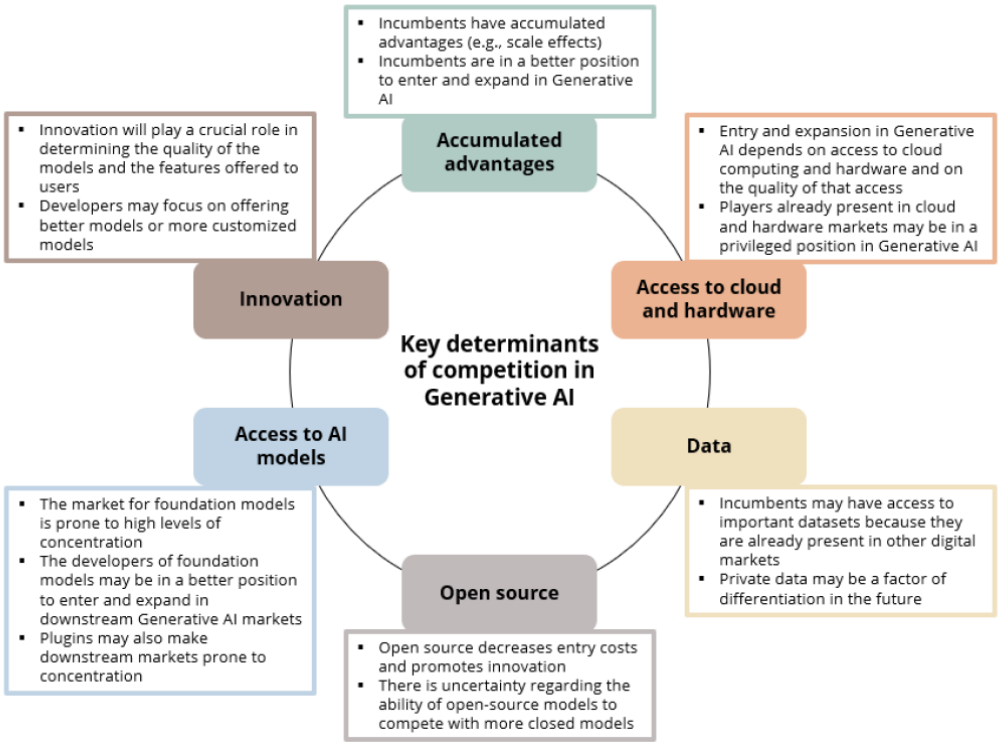


Figura 21: Determinanti della competizione nel mercato dell'intelligenza artificiale a valle. Source: (Autoridade da concorrencia, 2023)

Capitolo 3: Vantaggio competitivo dei dati

Nel precedente capitolo sono stati descritti i concetti di economie di scala e di *increasing returns* dei dati e di come questi giochino un ruolo importante nel settore. Sono state introdotte le principali tipologie e caratteristiche dei dati e del loro utilizzo concludendo che, per quanto centrali nello sviluppo dei modelli, non precludendo l'ingresso di nuovi entranti, non costituiscono una insormontabile barriera all'ingresso che aumenta la concentrazione nel settore.

Adesso, in questa sezione, si vuole sviluppare ulteriormente questa tesi per approfondire da un nuovo punto di vista, più tecnico ed economico come l'utilizzo dei dati nel settore dei *foundation models* impatti la competizione e se rafforzi, o meno, le tendenze monopolistiche (come un veloce ragionamento potrebbe suggerire).

3.1 *Network effect* e *Data-feedback loop*

Yuval Noah Harari, analizzando la Storia dell'umanità dagli albori ai giorni nostri per comprenderne la direzione e le tendenze attuali, nei suoi libri, introduce il concetto di "datismo" o "religione dei dati"[69] ipotizzandolo come principale teoria filosofica e socioculturale possibile in futuro. Secondo questa visione, l'universo è visto come un flusso costante di informazioni dove il valore di ogni fenomeno, individuo o entità è determinato dal suo contributo all'elaborazione di questi dati che, pertanto, sono considerati come unica fonte di verità per prendere decisioni informate. L'autore suggerisce come il datismo potrebbe influenzare profondamente il modo in cui viviamo e agiamo, con algoritmi che potrebbero arrivare a conoscerci meglio di noi stessi, sollevando numerose domande circa la libertà, la *privacy* e il ruolo della tecnologia nelle nostre vite.

[69] Dal libro "Homo Deus – A brief History of Tomorrow", 2015 cap. 11 'Homo Deus' by Yuval Noah Harari

Ripercorrendo gli ultimi decenni, a fine XX secolo il sociologo Luca Gallino parlò di "finanzcapitalismo"[70] per definire l'alleanza tra le principali industrie e i colossi finanziari; qualche anno dopo, lo sviluppo di internet e delle piattaforme digitali portò diversi autori, tra cui il saggista Nick Srniceck, a parlare di "capitalismo delle piattaforme"[71] analizzando gli impatti socio-economici di questo nuovo modello di *business* basato sulla raccolta e l'analisi delle informazioni che ogni utente lascia dietro di sé durante la navigazione in rete. Infine, negli ultimi anni si è diffusa la teoria di "capitalismo della sorveglianza"[72], raccontata nell'omonimo saggio della docente alla Harvard Business School Shoshana Zuboff. Esso è un capitalismo dell'estrazione e non della creazione di valore, in cui, secondo l'autrice, l'esperienza umana diventa materia prima gratuita trasformata in dati comportamentali e venduta come prodotto di previsione nel nuovo mercato dei *big-data* ad imprese affamate di conoscere il nostro comportamento.

Da questo quadro socioeconomico e culturale si vede come le idee "distopiche" di Harari, seppur non riferendosi ai giorni nostri, semplicemente proiettano sul futuro, enfatizzandoli, sentimenti e credenze già presenti oggi nella società.

I dati sono ormai argomento quotidiano nelle giornate di ognuno: *socials, cookies*, autorizzazioni al consenso del trattamento dei dati personali e normative *privacy,* notizie riguardanti estrazioni e/o vendita illegale di informazioni ecc…
L'attenzione delle persone al trattamento dei propri dati personali, con la paura che vengano sfruttati per scopi di manipolazione o malevoli è cresciuta molto, arrivando a coinvolgere oltre il 68% dei consumatori; il ché, potrebbe indurre molti a limitarne la condivisione (Gillespie, Lockey, Curtis, & Pool, 2023) (Fazlioglu, 2023).

---

[70] Dal libro "Finanzcapitalismo – la civiltà del denaro in crisi" , Einaudi 2011 https://www.einaudi.it/catalogo-libri/scienze-sociali/economia/finanzcapitalismo-luciano-gallino-9788806250102/

[71] Dal libro "Il capitalismo delle piattaforme" Nick Srnicek, pubblicato da John Wiley & Sons, 2016 Platform Capitalism - Nick Srnicek - Google Libri

[72] Dal libro "Il capitalismo della sorveglianza" di Shoshana Zuboff, pubblicato in Italia dalla Luiss nel 2019 Il capitalismo della sorveglianza. Il futuro dell'umanità nell'era dei nuovi poteri : Zuboff, Shoshana, Bassotti, Paolo: Amazon.it: Libri

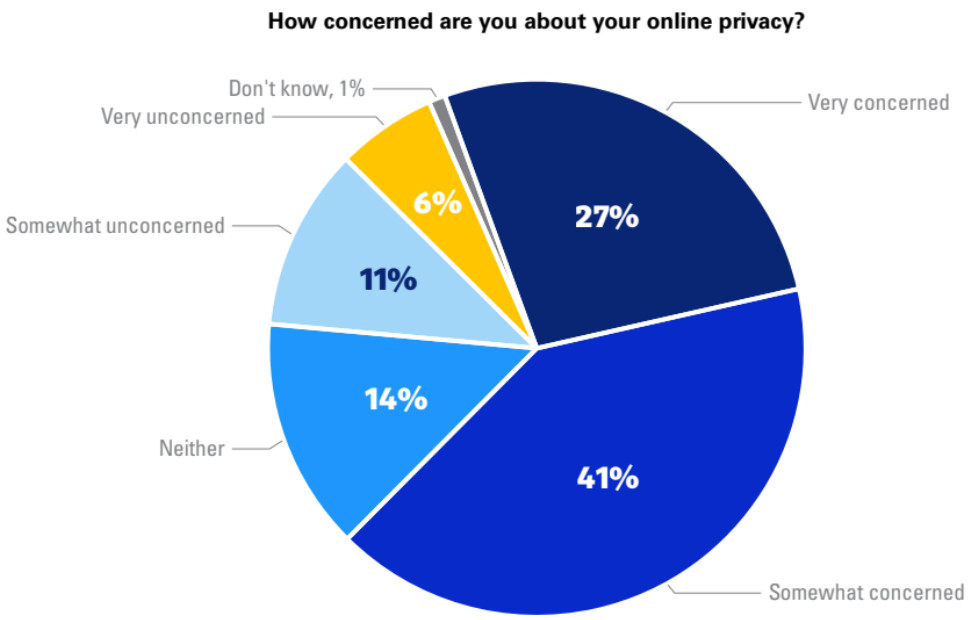


Figura 22: Interesse dei consumatori circa la privacy dei loro dati personali. Source IAPP

Negli ultimi anni poi, la diffusione dell'utilizzo dell'intelligenza artificiale o, in generale, le notizie circa i suoi rapidi progressi hanno notevolmente incrementato la pressione sociale sul tema e il potere che le persone attribuiscono ai dati, propri e degli altri.
Ciò detto, infatti, molti investitori e dirigenti d'azienda in campo AI credono che sia possibile utilizzare i dati estratti dal proprio *customer base* per ottenere un significativo e concreto vantaggio competitivo. La *ratio* sottostante questa assunzione è semplice: maggiore è il numero di clienti, maggiore è il quantitativo di dati estraibili che attraverso meccanismi di *machine learning* permettono migliorare l'offerta e attrarre ulteriori utenti. È quindi comune pensare di poter marginalizzare i propri *competitor* in una maniera paragonabile agli effetti di rete. Tuttavia, nella maggior parte dei casi questa assunzione è errata e dovuta ad una stima eccessiva del potere dei dati e del vantaggio che creano (Hagiu & Wright, When Data Creates Competitive Advantage, 2020).

I tradizionali effetti di rete, tipici delle piattaforme digitali, consistono nell'aumento di valore di un prodotto/servizio in relazione al numero complessivo di utenti che lo utilizzano. Si pensi ad esempio al telefono: il suo valore è strettamente correlato al numero complessivo di persone che ne possiedono uno e con cui sono possibili i contatti. Inoltre, maggiore è il numero di persone che ne possiede uno, più altri utenti saranno

attratti e portati all'utilizzo del dispositivo, generando un circolo virtuoso che si autoalimenta definito *network effect* diretto. Lo stesso accade con i *network effect* indiretti (o "di second'ordine") come *social media* o siti *web* come AirB&B o Booking il cui aumento di utenti da un lato stimola l'aumento del numero di *host* o strutture dall'altro con conseguente aumento di valore (e ricavi) della piattaforma per ambo le parti.

La figura sottostante rappresenta l'economia dell'esternalità di rete positiva diretta con il semplificato modello di Rohlfs (Rohlfs, 1974) in cui θ_0^L rappresenta il livello minimo di penetrazione nel mercato che si deve raggiungere per rimanere competitivi. In questo punto, anche detto di massa critica, il valore ottenuto dal bene è maggiore o uguale al prezzo pagato per acquistarlo.

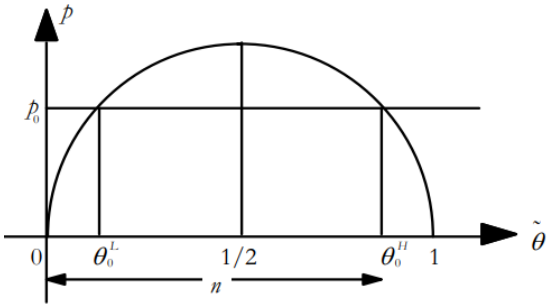


Figura 23: Esternalità di rete positive Source: Politecnico di Torino

Le economie di rete sono note per causare effetti di *lock-in.* Infatti, raggiunta una determinata massa di utenti quel prodotto potrebbe affermarsi come standard di mercato e far aumentare notevolmente gli *switching costs* per il passaggio ad un nuovo prodotto la cui rete, e i suoi conseguenti benefici, deve ancora consolidarsi. Basi pensare a Microsoft e alla tastiera QWERTY o a Facebook, Instagram e X.

Il ciclo virtuoso generato nei sistemi *data-enabled learning* come i *foundation models,* potrebbero sembrare, ad un primo sguardo, simili ai *network effect* tradizionali (comunemente, infatti, vengono definiti *data network effect*) in quanto il modello, attraverso *feedback loop* positivi, migliorando le *performance,* dovrebbe attirare nuovi utenti e raggiugere la massa critica per surclassare i competitors ed eventualmente affermarsi come standard. In questo caso, quindi, il valore non deriva dal numero di utenti (come nel caso del telefono) o dal numero di acquirenti e venditori (come nelle piattaforme di vendita) ma piuttosto dalla natura stessa della tecnologia: i modelli di

intelligenza artificiale migliorano attraverso l'apprendimento rafforzativo con previsioni seguite da *feedback* (Sheen S. Levine, 2023). Per contro numerosi *venture capitalist* e professionisti sottolineano come, in realtà, i *data feedback loop* del settore AI siano deboli e sovrastimati e forniscano vantaggi di molto inferiori rispetto ai tradizionali *network effect* delle piattaforme digitali, più duraturi e forti (Currier, 2020) (Korinek & Vipra, 2024). Per stabilire un reale vantaggio competitivo una azienda ha, infatti, necessità di sfruttare entrambi gli effetti per quanto sia complesso e, ad oggi, poche sono le realtà in grado di farlo (Hagiu & Wright, When Data Creates Competitive Advantage, 2020).

Il vantaggio competitivo dei dati

Raccogliere dati dai clienti per migliorare la propria offerta è una strategia collaudata e utilizzata da sempre ma il processo può essere molto lento, gli effetti limitati e difficile da realizzare su larga scala. Per case automobilistiche, aziende di beni di largo consumo e molti altri produttori tradizionali, ciò richiedeva l'analisi di dati di vendita, la conduzione di sondaggi tra i clienti e l'organizzazione di focus group, il chè rendeva il processo difficile e costoso.

Qualcosa ha iniziato a migliorare con l'avvento dei computer e del *cloud* grazie a cui è divenuto possibile elaborare in poco tempo enormi quantità di dati.

Prodotti e servizi connessi a Internet possono ora raccogliere informazioni direttamente dai clienti, inclusi i loro dettagli personali, il comportamento di ricerca e la cronologia *online*, le preferenze sui contenuti consumati e/o prodotti, la loro posizione GPS o anche una serie di dati più privati come abitudini alimentari o informazioni biometriche (Hacohen, 2023). Analizzando in *real time* questi dati, algoritmi di *machine learning* possono regolare le offerte per riflettere i risultati e persino adattarle ai singoli individui.

Questi sviluppi rendono l'apprendimento basato sui dati molto più potente delle intuizioni sui clienti che le aziende producevano in passato. Tuttavia, non garantiscono necessariamente un vantaggio competitivo stabile verso i *competitor* (Hagiu & Wright, 2020).

Gli autori Andrei Hagiu e Julian Wright propongono un *framework* di sette domande chiave per determinare a quale grado un vantaggio competitivo fornito dal *data-enabled learning* sia sostenibile nel tempo e di effettivo valore (Hagiu & Wright, When Data Creates Competitive Advantage, 2020).

1. Quanto valore aggiunto apportano i dati dei clienti rispetto al valore intrinseco dell'offerta? L'esempio di Mobileye[73], dove i dati di test sono cruciali per l'accuratezza dei sistemi ADAS (*Advance Driver Assistance System*), illustra un caso in cui il valore aggiunto dei dati è molto alto. Al contrario, per le smart TV, le raccomandazioni personalizzate basate sui dati dei clienti hanno un valore aggiunto relativamente basso per i consumatori.

2. Quanto rapidamente diminuisce il valore marginale dell'apprendimento? Seguendo l'esempio di Mobileye, i dati necessari per aumentare il livello di accuratezza dal 90% al 99% è di gran lunga superiore dei dati necessari inizialmente, tuttavia, date le implicazioni vita-o-morte del servizio, il valore di soli 9 punti percentuali aggiuntivi rimane estremamente alto. Più lentamente diminuisce il valore marginale, più forte è la barriera competitiva e ciò si nota anche nei sistemi di diagnosi di malattie rare, o nei motori di ricerca che continuano ad aver bisogno di grandi quantità di dati per produrre risultati affidabili. Ciò non accade invece in sistemi come, ad esempio, i termostati

[73]Sistema di guida autonoma dove l'accuratezza e la sicurezza sono imprescindibili e ottenibili attraverso i dati raccolti dai veicoli su cui viene installato https://www.mobileye.com/

intelligenti, a cui, una volta apprese le preferenze degli utenti in un paio di giorni, ulteriori dati non aggiungono valore.

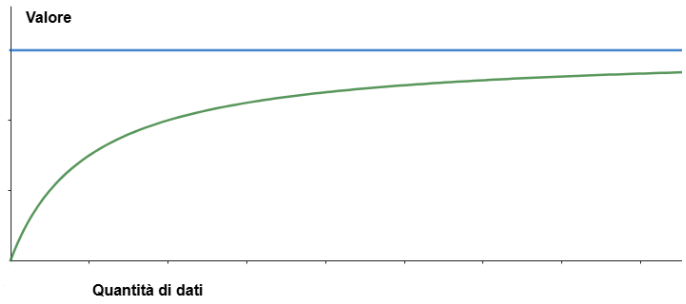


Figura 24:Il valore dei data network effect è, di solito, asintotico. Source: (Currier, 2020)

3. Quanto velocemente si svaluta il valore dei dati raccolti? Se i dati diventano rapidamente obsoleti, è più facile per un concorrente entrare a mercato senza dover accumulare anni di dati e, di conseguenza, minore è il vantaggio competitivo che forniscono. I dati accumulati da Mobileye e Google Search rimangono preziosi nel tempo; al contrario, nel mercato del *mobile gaming*, le preferenze degli utenti cambiano rapidamente come testimonia la storia di Zynga[74], famosa sul mercato per il lancio, nel 2009, di FarmVille che raggiunse nel 2012 una capitalizzazione di 10.4$B, crollata a 4$B a seguito dell'ingresso nel mercato di Epic Games[75] con Fortnite.

4. I dati sono proprietari? Oppure possono essere acquistati da altre fonti, facilmente copiati o *reverse-engineered*? Come abbiamo evidenziato nello scorso capitolo sui dati, l'avere dati proprietari a disposizione può essere un vantaggio competitivo solamente nella misura in cui gli stessi risultati non siano ottenibili grazie a *dataset* pubblici o a dati sintetici. Il mercato dei *software* di riconoscimento vocale, ad esempio, è stato dominato per molti anni dal sistema

[74] https://www.zynga.com/

[75] https://store.epicgames.com/it/p/fortnite

Dragon sviluppato dall'azienda Nuance[76] che migliorava con l'utilizzo del singolo utente e quindi, l'utilizzo dei dati privati raccolti. Negli ultimi anni il *machine learning* ha permesso la nascita di decine di sistemi simili e di pari livello allenati però su dati pubblici che si adattano in pochissimo tempo ad una nuova voce, mettendo a rischio la posizione di Nuance nel mercato.

5. Quanto è difficile per i competitors imitare i miglioramenti del prodotto ottenuti grazie all'elaborazioni dei dati dei clienti senza questi ultimi? Anche se i dati sono unici e forniscono intuizioni preziose, è difficile costruire un vantaggio duraturo se i miglioramenti risultanti possono essere facilmente copiati dai concorrenti senza dati simili. È importante, infatti, capire quanto i miglioramenti siano incorporati nel processo produttivo o, quantomeno, nascosti e la velocità con cui gli *insight* dei dati cambiano. Molte delle *feature* di Google Maps posso essere facilmente copiate ma il suo valore deriva dall'abilità di prevedere il traffico e raccomandare i tragitti migliori in *real time*, il ché risulta molto difficile da imitare, facendo leva su *user data* che diventano obsoleti in pochi minuti.

6. I dati di un utente aiutano a migliorare il prodotto per lo stesso utente o per tutti? Idealmente, entrambi, in quanto entrambi possono creare effetti di rete. Quando i dati di un utente migliorano il prodotto per sé stesso si parla di *within user learning,* che personalizzando l'esperienza utente crea per lui degli *switching costs* e ne aumenta la fedeltà al prodotto/servizio. Quando invece, i dati raccolti vengono uniti e usati per migliorare complessivamente l'offerta, si parla di *across user learning* che può portare un vantaggio chiave nella competizione per attrarre nuovi clienti e, potenzialmente, creare un *network effect* più potente. Le differenze di implicazioni tra i due stili di apprendimento verranno approfondite in seguito.

7. Quanto velocemente gli *insight* derivanti dai dati utente possono essere incorporati nei prodotti? Cicli di apprendimento rapidi rendono difficile per i concorrenti introdursi a mercato. Il vantaggio competitivo offerto dai *customer*

[76] https://www.nuance.com/it-it/dragon.html?srsltid=AfmBOopIm47DizwB47VOvnkkSH7ZVCqqNa04gYl26zVrNZKW3dzWfJ8R

data è tanto più forte quanto i clienti di oggi possono beneficiare degli *improvement* del prodotto, e non solo quelli di domani. Mappe, motori di ricerca e sistemi di gestione delle colture basati sull'IA possono essere aggiornati rapidamente. Al contrario, i *software* per la stima del profilo creditizio beneficiano dei dati storici dei clienti passati, analizzando *pattern* di comportamento e tasso di rimborso da cui deriva l'interesse applicato ai nuovi clienti, i quali, non beneficeranno in alcun modo dei propri dati.

Porgendo maggior attenzione alle domande 6 e 7, quando l'apprendimento attraverso i nuovi dati raccolti si riesce, in breve tempo a tradurre in un miglioramento tangibile della qualità dell'offerta, allora ai clienti interesserà l'estensione della *user base* e il valore offerto e percepito crescerà con essa, determinando un "*data network effect"*.

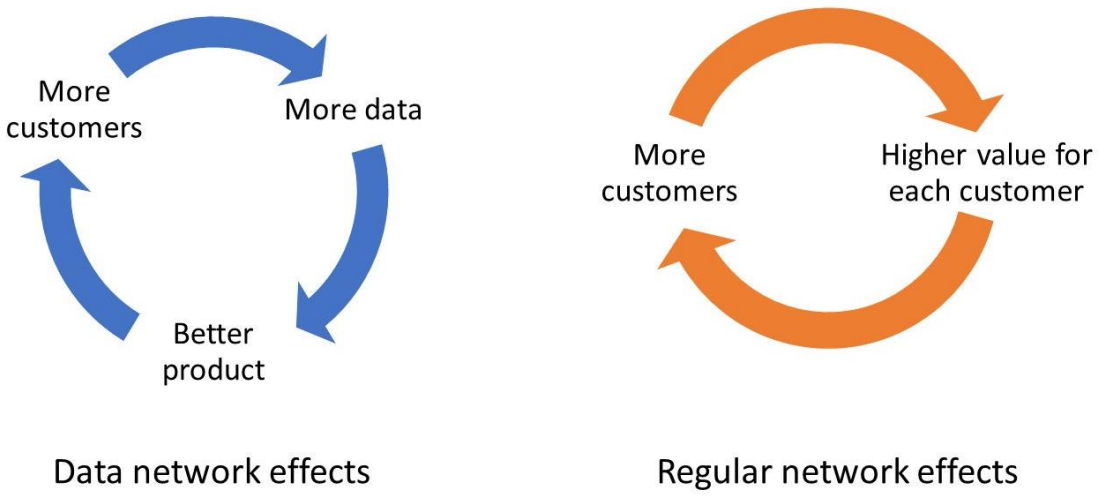


Figura 25: Differenze tra i Data network effect e i Network effect tradizionali. Source: Hagiu & Wright 2020 (https://platformchronicles.substack.com/p/why-data-network-effects-are-less)

In riferimento alle tradizionali esternalità di rete, vi sono alcune differenze importanti da sottolineare. In primis, se in entrambi i casi è necessario un numero minimo di utenti (per raggiungere la massa critica da un lato e avere una minima quantità di dati per iniziare il processo di miglioramento dall'altro), è pur vero che nel caso dei *data network effect* il problema è meno rilevante in quanto è più semplice reperire i dati necessari, comprandoli, che clienti. In secondo luogo, la presenza di *dataset* acquistabili, pubblici

o di dati sintetici e i miglioramenti tecnologici delle reti AI riducono la necessità di *customer data*, che già di per sé hanno una marginalità decrescente, e che, in alcune circostanze potrebbero non portare più, col tempo, alcun beneficio. Gli effetti di rete classici, per contro, sono più resilienti: un utente in più continua a portare valore anche se la *customer base* è già piuttosto vasta. Infine, la principale differenza tra i due fenomeni è che in presenza di *data network effect* i clienti sono interessati al numero complessivo di utenti in termini di qualità e utilità del prodotto, mentre con l'esternalità tradizionale sono interessati al numero crescente di interazioni possibili, a parità (o quasi) di prodotto sottostante. Se da un lato quindi il processo è "automatico" (Varian, 2018) e non richiede modifiche al prodotto, dall'altro può essere complesso e costoso prevedendo un reale miglioramento dell'ingegneria o del *design* del prodotto. Inoltre, la raccolta dei dati deve essere ben curata e mirata e riguardare una precisa tipologia: i *feedback.*

Per capire quanto i *data network effect* nel settore dei *foundation models* costituiscano un vantaggio competitivo o siano realmente sovrastimati occorre comprendere la natura dei dati in questione e quali segnali di *feedback* si riescano ad ottenere (Hagiu & Wright, 2023 a).

Alcuni prodotti hanno intrinsecamente forti *feedback loop.* Si prenda come esempio sempre Google Maps dove la scelta di percorso di ogni utente e il tempo impiegato a raggiungere la destinazione aiuta l'algoritmo a migliorare i tragitti proposti e le previsioni di traffico, innescando un *self reinforcing cycle.* Lo stesso vale per i *social network* o Spotify dove l'utente rivela chiari segnali di preferenza durante il naturale processo di consumo del prodotto/servizio. Altri prodotti, invece, hanno per natura dei *feedback loop* più deboli in quanto il loro utilizzo è difficile da tracciare o non rivela preziose informazioni di preferenza dell'utente. Ciò è ovvio per i prodotti non *digital connected* come auto o vestiti, per cui i *feedback* vengono raccolti tramite sondaggi e *focus group,* e per i prodotti con un ciclo di vita molto lungo che attendono anni prima di ricevere un riscontro*,* come i sistemi di previsione del merito creditizio o di *venture investing.* Lo è meno il fatto che anche alcuni prodotti digitali che raccolgono enormi quantità di dati possano soffrire di scarsi *feedback loop*. Classico esempio è Fitbit[77], pioniere dei *tracker*

[77] https://community.fitbit.com/t5/Community/ct-p/EN

fitness indossabili dal 2007 che ha collezionato negli anni una vastissima quantità di dati da milioni di utenti non riuscendo però a preservare il proprio vantaggio con l'arrivo nel mercato di Samsung, Garmin ecc… Fitbit, in particolare, raccoglie dati biometrici dagli utenti e li elabora per fornire loro informazioni sulla loro condizione fisica e consigli pratici. Tuttavia, il sistema si limita a comparare i dati osservati con i *pattern* e i valori pre-programmati fornendo informazioni che sono principalmente statistiche riassuntive o risultati di confronti con punti di riferimento preesistenti (Hagiu & Wright, 2023 a). Inoltre, l'algoritmo non ha modo di conoscere lo "stato vero" dell'utente per capire la differenza con la sua previsione e, di conseguenza, non ha modo di migliorare la propria accuratezza basandosi sui dati osservati. Il valore del prodotto risiede quindi nella pre-programmazione e non aumenta significativamente attraverso un *data-enabled learning.*

I LLMs ingeriscono una enorme quantità di dati dal web o altre fonti private e li utilizzano per generare risposte che altro non sono che previsioni di tipo statistico. La fase di *training* prevede sempre una fase di *testing* durante la quale il modello misura la differenza tra il proprio *output* e quello corretto, traendo quindi un *feedback.* L'interazione con gli utenti, invece, seppur riesca a raccogliere un enorme quantitativo di dati circa i task o le domande più frequenti, non sempre crea dei *feedback* utili limitando il valore offerto al *pre-training.* Vi sono comunque degli accorgimenti che possono essere implementati per ovviare al problema e rafforzare i *feedback loop.* Innanzi tutto, il prodotto deve essere progetto affinché l'utente, con il semplice utilizzo riesca a segnalare, in maniera più o meno esplicita, il proprio livello di soddisfazione, ad esempio, inserendo dei meccanismi *thumbs up/thumbs down* o di *ranking,* tracciando se gli utenti copiano le risposte, permettendo loro di creare una cartella con le migliori risposte o monitorando il tempo trascorso su ogni *output* (Amer). Inoltre, può essere utile inserire delle domande progressive e specifiche che guidino la risposta del modello per meglio comprendere la richiesta dell'utente oppure inserire umani nel *loop* di apprendimento (*Reinforcement Learning from Human Feedback*) che controllino e valutino le risposte. Costoso e poco scalabile, quest'ultima metodologia può essere molto efficace nelle fasi iniziali di sviluppo, durante le quali la *learning curve* è più ripida (Timo Kaufmann, 2024).

Nel mondo dell'intelligenza artificiale generativa, quindi, come in molti altri mercati, digitali e non, i *data feedback loop* hanno la potenzialità di generare *data network effect*

e rafforzare la dominanza nel mercato. Numerose sono, tuttavia, le motivazioni che sostengo questa tesi essere l'eccezione piuttosto che la regola. Il principale utilizzo di *dataset* pubblici per il training dei grandi *foundation models,* unito al tendenzialmente scarso o nullo volume di dei *feedback* ottenuti, non crea alcun vantaggio competitivo nei confronti dei *competitor,* i quali possono avere accesso agli stessi dati (o simili). Come sottolineato nei paragrafi precedenti l'unicità dei dati utilizzati è altrettanto importante alla qualità e alla quantità dei dati al fine di creare un distacco dai concorrenti. Se anche si riuscissero, infatti, a creare *feedback loop* ma i dati *ab origine* sono di pubblico dominio o reperibili da molte fonti diverse, il vantaggio creato si perde completamente. Le aziende che offrono servizi di radiologia supportata da intelligenza artificiale esemplificano ulteriormente questo concetto. Una compagnia che ha stabilito collaborazioni con numerosi ospedali può avere accesso a milioni di immagini e rapporti radiologici corrispondenti per addestrare e migliorare i propri modelli. Tuttavia, considerando il gran numero di ospedali che forniscono servizi di radiologia a livello globale, risulta poco realistico che un'unica realtà possa garantire l'accesso esclusivo a dati sufficienti per impedire ai rivali di addestrare modelli concorrenti. Pertanto, sebbene ogni azienda di radiologia AI possa migliorare i propri modelli nel tempo grazie ai dati ottenuti dai *feedback*, nessuna entità può monopolizzare il mercato esclusivamente attraverso questo meccanismo. Non sorprende, quindi, che attualmente ci siano molte startup di radiologia AI ben finanziate che competono in questo settore, tra cui Rad AI[78], Nuance Communications[79], Subtle Medical[80], DeepTek.ai[81] e Aidoc[82] (Hagiu & Wright, 2024).

Oltre alla natura dei dati disponibili, è necessario è necessario valutare se la curva di apprendimento continui a crescere o si stabilizzi ben prima che tutti i dati a disposizione

[78] https://www.radai.com/

[79] https://www.nuance.com/it-it/index.html?srsltid=AfmBOoqO9XPBAgaVP13bllN6USyM9Y1edvzCtOWk5gmWNJV4NnEw7tOe

[80] https://subtlemedical.com/

[81] https://deeptek.ai/

[82] https://www.aidoc.com/solutions/radiology/

siano esauriti. Alcuni sostengono che, nel contesto dei big data e dell'intelligenza artificiale, quest'ultima situazione sia più comune (Bajari, 2019); ma si tratta comunque di una questione empirica, e la risposta dovrà essere verificata per ciascuna applicazione. I motori di ricerca, ad esempio, pur avendo collezionato negli anni quantità considerevoli di dati, continuano ad imparare e affinarsi quotidianamente, mantenendo una *learning curve* in crescita. Gli studi empirici condotti su questi sistemi hanno infatti confermato la presenza di importanti *feedback loop* costanti e conseguenti effetti di rete. Questo riflette l'importanza degli *edge case* ("casi limite"). Infatti, il fattore distintivo tra un motore di ricerca adeguato ed uno eccellente risiede nella capacità di fornire risultati utili per *query* meno comuni. Pertanto, maggiore sarà l'utenza del sistema, maggiore sarà il numero di *feedback* per questi casi particolari e maggiore sarà l'accuratezza (Klein, 2022) (Schaefer & Sapi, 2023). Inoltre, la pertinenza dei risultati di ricerca cambia spesso nel tempo, rendendo necessaria la raccolta continua e costante di dati. Tutto ciò genera forti *feedback loop* che sono sicuramente tra i determinanti della struttura *winner-takes-all* del mercato dei *serch engine* (Hagiu & Wright, 2024). Al contrario, nel settore dei *FM,* sebbene alcune applicazioni possano essere progettate *ad hoc* con dati unici e proprietari e creare ottimi *feedback* per il *deployer,* non è sempre chiaro se e come ciò venga trasmesso al *developer* a monte. Molti applicativi, infatti, non condividono gli *user-data* con i proprietari del modello per ragioni di *privacy* o strategiche. Un esempio è Chat GPT Enterprise[83], utilizzato da molte aziende per sviluppare il proprio servizio *chatbot AI* personalizzato, che non condivide i dettagli dei dati utente ai modelli GPT. Questi segnali posso poi essere di limitata utilità o essere troppo dilazionati nel tempo come discusso in precedenza nel caso di Fitbit e dei programmi valutativi del merito creditizio o di *investing.*

Infine, l'ulteriore elemento che determina gli effetti sulle dinamiche competitive dei *data feeedback loop* e se essi costituiscano una reale barriera all'ingresso per nuovi entranti è la natura dell'apprendimento.

[83] https://openai.com/index/introducing-chatgpt-enterprise/

3.3 Across user & Within user learning

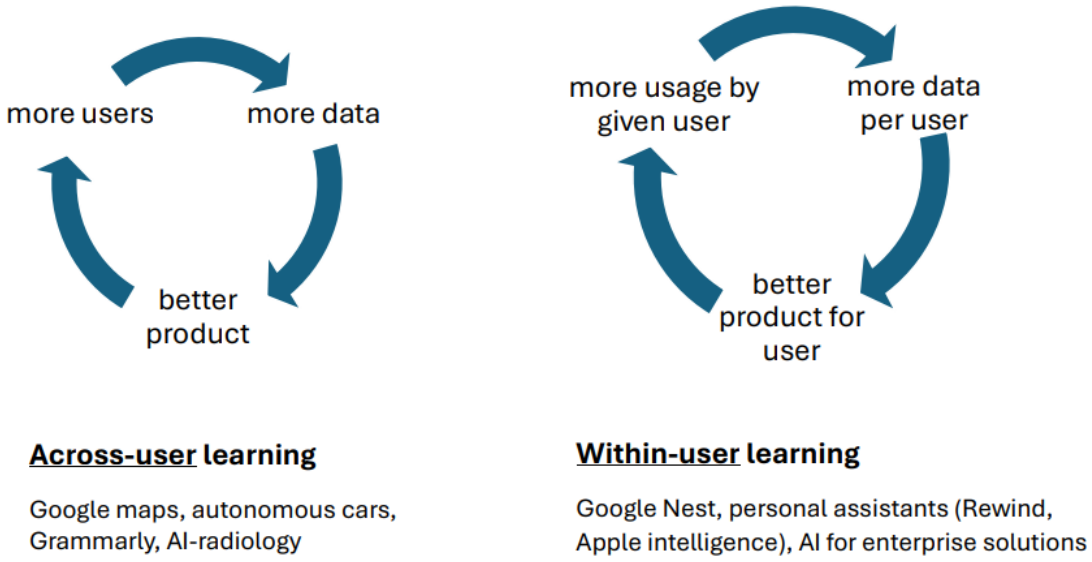


Figura 26: Across user learning and Within user learning. Source: (Hagiu & Wright, 2024)

L'apprendimento basato sui dati generati dagli utenti (*user generated data* UGD (Hacohen, 2023)) può essere di due tipologie distinte a seconda di quanto spazio venga affidato alla personalizzazione e quanto al miglioramneto del modello generale offerto. Nell'apprendimento *across-user* i dati e i *feedback* raccolti da tutti gli utenti vengono uniti ed elaborati congiunamente per migliorare il prodotto. I risultati di questo *upgrade* sono quindi fruibili da tutti gli utenti. Esempi classici di questi sistemi sono Google Maps, gli applicativi di radiologia AI *powered* o Grammarly[84] , un servizio *cloud-baased* per la correzione sintattica e grammaticale che suggerisce il corretto *wording* per ogni situazione. In ogni esempio, al crescere degli utenti il prodotto migliora per tutti.

Viceversa, il *within-user learning* si verifica quando un maggiore utilizzo di un prodotto o servizio da parte di un singolo utente consente all'intelligenza artificiale di migliorare il prodotto specificamente per quell'utente. In altre parole, il sistema impara dalle interazioni ripetute di un individuo e personalizza la sua offerta in base alle preferenze e

[84] https://www.grammarly.com/

alle esigenze di quel preciso utente nel tempo. Esempi per quest categoria includono gli applicativi personalizzati per le aziende che vengono affinati sui dati propietari come GPT – Enterprise o il più volte menzionato Fitbit: quanto più un cliente usa il proprio dispositivo Fitbit, tanto più il sistema è in grado di fornire consigli migliori e più utili per quella persona.

La distinzione è importante in fase di analisi delle ripercussioni economiche dei *feedback loop* associati ai due meccanismi.

I *feedback loop* basati su un apprendimento *across-user* agevolano la formazione di *network effect* in quanto il prodotto migliora le *performance* per tutti gli utenti, attirandone di nuovi. Per questo motivo le aziende facenti leva su questo meccanismo hanno una maggiore probabillità di sperimentare dinamiche di mercato *winner-takes-all.*

Diversamente, invece, un sistema *within-user learning* non crea effetti di rete tra altri utenti ma, customizzando l'esperienza, crea degli *switching cost* verso i rivali che crescono nel tempo. Questo crea un progressivo effetto di *lock-in* con i clienti esistenti ma non difende l'azienda dalla competizione per i nuovi e, quindi, c'è meno ragione di aspettarsi dinamiche di *winner-takes-all* (Hagiu & Wright, 2023 b).

Come naturale conseguenza delle dinamiche appena descritte, i *feedback loop* che combinano entrambi gli approcci tendono ad essere i migliori nel creare una posizione dominante e difendibile nel mercato. Si può sostenere, infatti, che Google search, presenti entrambe le caratteristiche e si è affermato come *leader* indiscusso di mercato. Oltre ad apprendere quali risultati vengono maggiormente *clicked* dalla maggior parte degli utenti in risposta a diverse *query*, Google è anche in grado di personalizzare i risultati proposti, in certa misura, sulla base della cronologia di ricerca precedente, della posizione o della lingua utilizzata. Analogamente, i *recommender engines* di Amazon, Instagram, Netflix, Spotify, TikTok e YouTube sfruttano entrambi i tipi di apprendimento. Essi correlano i dati provenienti dalle esperienze di molti utenti su vari *item* per prevedere la probabilità che un utente possa apprezzare un particolare elemento, in base alle proprie interazioni con elementi simili o in qualche modo correlati. Questo è il motivo per cui tali *recommendation services* traggono vantaggio da potenti *feedback loop.*

Quanto forti sono i vantaggi derivanti da un apprendimeto basato sugli UGD rispetto ai tradizionali *learning by doing* e *network effect?* Quali sono le diverse implicazioni sulla

concorrenza? E quali dinamiche economiche e di *welfare* entrano in gioco? Come si inseriscono i *foundation models* in questo quadro di riferimento ?

Per rispondere a queste domande i professori A. Hagiu e J. Wright propongono un semplice modello matematico di competizione basato sulla competizione di Bertrand[85], tra due aziende *infinitely lived,* che differiscono per la quantità di dati raccolti negli esercizi precedenti e la forma della funzione di apprendimento e la tipologia.

Il loro studio si riferisce in generale a tutti i sistemi *"data enabled learning"* ma si vuole, in questa tesi, interpretarne i risultati inserendoli nella catena del valore dell'AI. In particolare parlando di *foundation models,* essi possono essere collocati nella categoria *across-user learning* in quanto per sviluppare una determinata versione di modello utilizzano, oltre ai dati raccolti da fonti pubbliche o terze parti, i dati storici delle interazioni con gli utenti accumulati sino ad un preciso momento. Una volta allenato e rilasciato il modello, gli utenti utilizzano quello nel mentre che l'azienda continua a raccogliere ulteriori dati, elaborarli e lavorare per migliorare l'offerta. Il modello quindi, incorpora le modifiche date dai nuovi dati sono nella *release* successiva, in un periodo successivo. Il *within period learning,* invece, è la modalità di apprendimento in cui le imprese imparano immediatamente dai consumatori che utilizzano il prodotto nel periodo corrente, piuttosto che apprendere nel periodo successivo (come assunto nel modello base *across-user learning*). Introduce pertanto, una dinamica in cui l'apprendimento basato sui dati dei clienti del periodo corrente ha un impatto immediato sul valore del prodotto per quegli stessi clienti, influenzando le loro decisioni di acquisto e creando potenzialmente un ruolo per il *customer belief* e problemi di coordinamento: a parità di altre condizioni, i consumatori preferiscono acquistare dall'azienda da cui si aspettano che anche altri acquistino, dato che beneficiano del miglioramento risultante del prodotto durante il periodo di consumo. Questa caratteristica è più facilmente riscontrabile nei prodotti *cloud-based* oppure nelle aplicazioni di intelligenza artificiale che quindi utilizzano i dati raccolti per continuare costantemente il processo di *fine tuning* di un modello LLM.

[85] La competizione di Bertrand è un modello economico che analizza la concorrenza tra imprese con un'offerta omogenea. In questo modello, le aziende competono fissando i prezzi dei loro beni piuttosto che le quantità. Se due imprese fissano lo stesso prezzo, si dividono il mercato; tuttavia, se una delle due riduce il prezzo, cattura l'intero mercato innescando un processo di guerra di prezzo che annulla i profitti di entrambe nel lungo termine.

Come detto in precedenza, i costi per affinare un *foundation model* sono di gran lunga inferiori ai costi per costruirlo (così come anche la quantità di dati necessari), motivo per cui, in generale, è più semplice per aziende *deployer* permettersi un continuo *re-training* e *release* quasi istantanee. Infine, l'ultimo modello di apprendimento analizzato è il *within-user learning* che migliora il modello esclusivamente per quel dato utente ed è tipico, anche in questo caso, delle applicazioni a valle della catena del valore più che dei modelli generali.

Modello *across-user learning* per i *FM*

Prima di riassumere i risultati del modello e analizzare gli effetti competitivi si elencano brevemente di seguito le principali assunzioni e come esse possano rispecchiare il mercato dei *FM* oltre quanto già descritto nel precedente paragrafo, giustificando l'associazione. Il modello *across-user*, si pone l'obiettivo di replicare la dinamica competitiva tra due aziende (i), identificate come *Incumbent* (I) ed Entrante (E), che si sfidano nel mercato avendo libera scelta di prezzo. Entrambe le aziende operano con costi marginali di produzione equivalenti (c) e sono caratterizzate da diverse funzioni di apprendimento crescenti dipendenti dalla tecnologia adottata (f_i). Nel mercato dei *foundation model,* come descritto nel capitolo precedente, i prezzi possono discostarsi molto a seconda delle scelte strategiche aziendali, mentre i costi marginali di produzione sono abbastanza costanti in termini di costo energetico per singolo calcolo essendo basati su *hardware* sottostanti, costituito da GPUs, simili per tutti i modelli. Tuttavia può essere diverso se compariamo aziende con *cloud* proprietario (come Google, Microsoft o Amazon) e altre aziende che invece acquistano potenza computazionale in *cloud* da *provider* terzi. Il fatto di avere diverse funzioni di apprendimento è ovviamente rispecchiato nel mercato di riferimento dove ogni azienda, con i propri ingegneri o comunità *open source* lavora per migliorare la funzione. La possibilità di gestire funzioni asimmetriche è ottenuta semplificando il modello ed eliminando l'incertezza circa la preferenza degli *user*, la cui domanda è normalizzata a 1 in ogni periodo. Di conseguenza il numero di periodi in cui una azienda è stata scelta dai consumatori corrisponde al numero di clienti da cui il modello impara (N_i). Inoltre, non vi è distinzione tra consumatori abituali e nuovi, poiché entrambi contribuiscono in modo equivalente al processo e vengono considerati atomistici, il che implica che le loro decisioni non

influenzano direttamente le strategie di prezzo delle aziende, e soprattutto dipendono esclusivamnete dal *surplus* offerto nel periodo corrente[86]. In generale questo può ricondursi al mercato dei *foundation model* dove, ad oggi, il prezzo è una variabile importante ma comunque subordinata alle *performance, main driver,* e dove i principali *provider,* grazie anche alle varie strategie di *bundling* con altri servizi dello stesso ecosistema, hanno un forte potere nella determinazione del prezzo. Importante è anche l'assunzione che f_i sia una funzione ad S, più o meno ripida con un preciso *upper bound* ($\overline{N}_i$) oltre il quale, l'accumulo di ulteriori dati non incide sulle *performance* del modello, a rappresentare il valore marginale decrescente dei dati nel contesto dei *FM,* e un *lower bound* ($\underline{N_i}$) livello minimo di dati necessari ad una azienda per immettere il proprio prodotto nel mercato ed iniziare il processo di raccolta di *user-data* per il *training.*

Le tre fonti di asimmetria tra *I* ed *E* evidenziate dal modello di Hagiu e Wright (2023) e rispecchiate anche nella realtà dei *FM* sono: (i) la differenza nel valore *standalone* (s_i) del prodotto, ossia il valore base a cui viene aggiunto l'*extra value* dato dall'apprendimento[87], (ii) differenze nella funzione di apprendimento e/o nei valori limite ($f_I \neq f_E$ e/o $\overline{N}_I \neq \overline{N}_E$), e (iii) la posizione in cui si trovano le due aziende lungo la curva di apprendimento, in particolare la vicinanza al raggiungimento del valore limite ($\overline{N}_i$) o, al contrario, al valore minimo necessario per erogare il prodotto ($\underline{N_i}$).

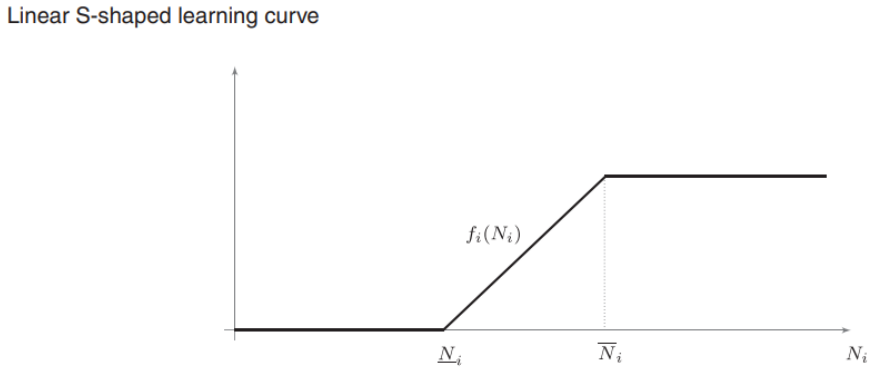


Figura 27: Forma ad S della curva di apprendimento nei sistemi data-enabled-learning
Source: (Hagiu & Wright, 2023 b)

[86] Come verrà spiegato nelle righe successive, un cliente sceglie l'azienda *E*, ossia *E* vince se: $s_E + f_E(N_E) \geq s_I + f_I(N_I)$

[87] Per cui il valore offerto da ognuna delle due aziende in un dato periodo è $V_i = s_i + f_i(N_i)$

Secondo questo modello, in un dato periodo, se entrambe le aziende si trovano lungo la curva di apprendimento, l'*Incumbent* gode di un vantaggio competitivo dato da un valore soglia $s_E - s_I = \Delta(N_I, N_E)$ che l'entrante deve superare per vincere. Questo valore, ovviamente, aumenta con N_I e decresce con N_E ricordando la teoria dell'"*increasing increasing dominance"* (IID) di Cabral e Riordan: più *I* vince, più vincerà con maggior facilità nei periodi successivi (Cabral & Riordan, 1994). Questo valore soglia rappresenta la differenza tra il valore attuale del surplus lordo generato dall'apprendimento dell'*incumbent* e dell'entrante, considerando i percorsi in cui ciascuna impresa vinca in ogni periodo a partire da quello corrente. Tuttavia, il modello evidenzia come il vantaggio competitivo di *I* sia limitato dalla funzione di apprendimento dell'entrante che potrebbe, infatti, godere di una migliore tecnologia ed essere più ripida ($f_E > f_I$) tanto da compensare lo svantaggio iniziale in termini di dati. Oppure *E* potrebbe avere un valore soglia $\overline{N_E}$ più alto, tanto da, con un fattore di sconto sufficientemente basso, aumentare il valore attuale del suo surplus futuro, permettendo all'entrante di ottenere un vantaggio competitivo indipendentemente da quanto ad oggi sia indietro rispetto all'*Incumbent* ($\Delta(N_I, N_E) \leq 0$in termini di dati. Questa dinamica è sata più volte osservata nel mondo dei *foundation model.* Basti pensare a DeepSeek, i modelli *open source* LLaMa, o anche alla stessa OpenAi, nata come piccola *startup* se confortata con i colossi digitali che erano presenti sul mercato.

Inoltre, il modello è utile a spiegare l'esistenza dell'*open source* o dei modelli gratuiti. E' prevista, infatti, la possibilità di offrire l'utilizzo del modello ad un prezzo negativo, economicamente interpretabile come forma di sussidio che l'azienda perdente offre ai consumatori in cambio dei loro dati. Rappresenta quindi il costo sostenuto dall'azienda perdente per rimanere a mercato che va a beneficio dei consumatori (più innovazione, prodotti gratuiti, aumento della competizione che limita i monopoli, etc ..).

Comprendere le logiche sottostanti gli incentivi è importante per l'analisi delle scelte strategiche e competitive delle aziende in un contesto dinamico di lungo periodo.

La disponibilità dell'azienda perdente a offrire un sussidio riflette l'aspettativa che essa ha di vincere in futuro ed è matematicamente rappresentato dal *present value* dei futuri profitti. Pertanto il sussidio di oggi sarà tanto maggiore quanto più *E* (o la perdente in generale) preveda la soglia $\Delta(N_I, N_E)$ diminuire. *E* sussidia i clienti quando, pur

perdendo ad oggi $[s_E - s_I \leq \Delta(N_I, N_E)]$, prevede di poter vincere in futuro se, invece, oggi vincesse $[(N_I, N_E + 1) \leq s_E - s_I]$.

Questo accade soprattutto con le nuove *startup* che, introducendo delle innovazioni di processo o prodotto nei nuovi modelli (rappresentate da una curva più ripida o un *upper bound* più elevato), e prevedendo la staturazione tecnologica dell' *incumbent* avversario, e, anticipano un delta soglia $s_E - s_I$ decrescente nel tempo scegliendo così di restare a mercato e sussidiare i clienti.

Per quanto riguarda la posizione delle aziende lungo la curva di apprendimento mostrata in figura 19 emergono due importanti considerazioni:

i. Quando entrambe le aziende si trovano nel primo tratto orizzontale della curva, aumentare il valore dei dati a disposizione, rafforza il vantaggio dell' *Incumbent* vincente in quanto la avvicina al *lower bound* $\underline{N_i}$, aumentando il valore attuale dei futuri profitti, che diventano, così, più vicini. Al contrario,in uno scenario che vede entrambe le aziende lungo la parte cescente della curva, migliorare la qualità delle informazioni estratte dai dati (ad esempio adottando una tecnologia più efficiente o ottimizzando il processo di *feedback*) riduce il vantaggio di *I,* rendendo il valore cumulato in precedenza relativamente meno importante.
ii. Considerazioni speculari possono essere fatte riguardo a possibili *shock* che forniscono alle aziende più dati, come l'acceso ad una nuova *data source*, e le avvicinano alla soglia $\underline{N_i}$, da cui possono iniziare ad offrire valore ai cienti. Nel tratto orizzontale la maggior vicinanza alla soglia favorisce l'*incumbent* ma lo svantaggia successivamente. Nello scenario in cui entrambe le aziende si trovano lungo il tratto crescente della curva di sviluppo, infatti, una riduzione del numero di periodi necessari al raggiungimento della soglia superiore diminuisce il numero di periodi in cui *I* può godere del suo vantaggio, aiutando *E* che ha "più da imparare".

Si dice spesso che i cosiddetti *data network effect* consolidino gli *incumbents* sia nei mercati di pura raccolta dei dati, sia in quelli adiacenti dove possono essere utilizzati. Se così fosse le più grandi piattaforme *online* — tra cui Alphabet, Meta, Apple, Amazon e Microsoft — dovrebbero avere un vantaggio significativo nei mercati in espansione dei servizi di intelligenza artificiale generativa. Eppure, i giganti della tecnologia non sono

stati finora in grado di sfruttare a pieno i loro vasti tesori di dati per superare *startup* come OpenAI e Midjourney o i nuovi modelli DeepSeek o i vari LLaMa (Manne & Auer, 2024). L' avere accesso a enormi quantità di dati, soprattutto storici, come nel caso delle *big tech* americane (le *Incumbent* nel nostro modello), non garantisce di per sé un vantaggio competitivo difendibile da altri *player* minori che potrebbero aver sviluppato tecnologie più all'avanguardia. Inoltre, la marginalità decrescente dei nuovi dati riduce nel tempo il vantaggio accumulato in favore di nuovi entranti. Con questo non si vuole affermare che non ci siano più barriere all'ingresso e che il valore del vantaggio cumulato dai grandi *player* sia nullo, ma piuttosto si vogliono evidenziare delle dinamiche economiche che possono essere di grande supporto ai *policymaker.* Infatti, la rapida evoluzione della tecnologia AI generativa mette in discussione diverse assunzioni fondamentali dei dibattiti odierni sulle politiche di concorrenza digitale soprattutto in riferimento ai dati e al loro "potere" nella competizione. I risultati economici suggeriscono come i vantaggi degli *incumbents* in termini di dati siano molto meno pronunciati di quanto comunemente si assuma lasciando ampio margine di ingresso a numerose *startup*. Di conseguenza, gli sforzi per impedire alle piattaforme del Web 2.0 di competere liberamente nei mercati dell'AI generativa potrebbero avere effetti controproducenti, eliminando una fonte importante di concorrenza (Manne & Auer, 2024). Da questo punto di vista la *deregulation* americana favorisce lo sviluppo dell'ecosistema AI (a discapito di potenziali rischi in merito a questioni di *privacy* e tutela dei dati personali), mentre l'Europa, con le sue direttive più rigide tende ad ostacolarlo, prestando più attenzione all'individuo, i suoi dati e all'etica dei processi (Preta, 2024).

Il modello solleva un'ulteriore questione importante. Seppur venga dimostrato che la soglia identificata $s_E - s_I = \Delta(N_I, N_E)$ e la libera competizione tra le aziende sia la soluzione socialmente efficiente che massimizza il *welfare* complessivo nel lungo termine, emerge comunque un disallineamento di interessi tra l'azienda vincente e i consumatori.

Come abbiamo detto l'azienda perdente *E* rimane nel mercato sussidiando i consumatori; al netto dell'eventuale sussidio che può offrire anche la vincitrice, il valore attuale di

questi sussidi rappresenta il loro *surplus* corrente[88]. Man mano che l'azienda vincente vince e acquisisce nuovi dati, diminuisce l'incentivo di *E* a sussidiare i consumatori, vedendo i suoi profitti più lontani e difficili da raggiungere; in maniera speculare, crescono i profitti dell' *Incumbent.* Pertanto il *surplus* dei consumatori e i rendimenti dell'azienda vincente si muovono in direzioni opposte rispetto ai principali fattori di vantaggio competitivo, dando luogo ad una inefficienza di mercato. Formalmente, si ha che:

i. Data *i* l'azienda vincente e *j* la perdente, il surplus dei consumatori aumenta con s_j e $f_j(N_j)$, e diminuisce debolmente con s_i e $f_i(N_i+k)$, per tutti $k \geq 0$.
ii. I profitti dell'azienda vincente *i* diminuiscono con s_j e $f_j(N_j)$, e aumentano con s_i e $f_i(N_i+k)$ per tutti $k \geq 0$.

All'aumentare del valore prodotto da *j* [aumento di s_j e/o $f_j(N_j)$] diminuisce la soglia necessaria a *j* per vincere, incentivandola a sussidiare i consumetori e rimanere a mercato. Contrario accade con l'aumento del valore offerto da *i* [aumento di s_i e $f_i(N_i+k);\ +k$ perché si suppone *i* stia vincendo](i), che ne riduce l'incentivo e quindi il *surplus* dei consumatori. a beneficio del proprio profitto aziendale (ii).

Con l'apprendimento progressivo dell*'incumbent* e il miglioramento delle *performance* del prodotto il valore immesso nel mercato cresce, con conseguente aumento del benessere sociale complessivo che però, non viene equiripartito tra consumatori e azienda vincente, favorendo quest'ultima. Viceversa, supportare l'apprendimento dell'azienda perdente la incentiva a rimanere a mercato sussidiando i consumatori e aumentandone il *surplus.*

In un modello più complesso di concorrenza è prevedibile che una parte dei benefici derivanti dall'apprendimento dell'impresa vincente venga trasferita ai consumatori. Tuttavia, l'osservazione secondo cui i consumatori possono subire un danno dall'apprendimento dell'impresa vincente, che disincentiva la perdente a sussidiare, è probabilmente applicabile in modo più ampio. Questa dinamica sottolinea l'importanza per i *policy maker* nel mercato dei *foundation model* di supportare le nuove aziende, assicurandone le condizioni ottimali per continuare a competere e ridurre quanto più

[88] Che equivale alla differenza tra il surplus offerto dall'azienda vincente e il suo profitto

possibile la distanza con gli *incumbent*. Possibili strumenti regolatori a tal fine riguardano i dati, la loro diffusione e utilizzo, in particolare l'introduzione di politiche di condivisione.

Richiedere all*'incumbent* di condividere alcuni dei suoi dati con l'entrante potrebbe essere un modo per aiutare quest'ultimo a raggiungere pari livello di apprendimento e, basandosi sulla precedente analisi del benessere, ciò sembrerebbe vantaggioso per i consumatori. Per quanto ciò possa essere vero nel caso in cui le aziende condividano *una tantum* i dati, una condivisione costante attenuerebbe gli incentivi di entrambe le aziende ad investire nell'estrazione di nuovi dati e, in particolare, potrebbe incoraggiare la svantaggiata ad adotare un comportamento di *free-riding* riducendo la competitività nel mercato e potenzialmente aumentando i prezzi. Più precisamente la condivisione incrementa il *surplus* dei consumatori quando l'azienda perdente è significativamente svantaggiata, mentre lo riduce quando le aziende sono più vicine. In situazioni di forte svantaggio, l'azienda perdente non offre sussidi, quindi l'effetto positivo di mantenere le aziende equilibrate nei periodi futuri prevale. Al contrario, quando le aziende sono simili, il *data sharing* eliminerebbe l'incentivo della perdente a sussidiare i consumatori, azione che naturlamente farebbe, determinando una perdita di *surplus* maggiore dei vantaggi ottenibili.

A livello di *policy* la differenza nell'economia dei modelli all'inizio e durante la fase di sviluppo sottolinea l'importanza di adeguare i regolamenti alla struttura del mercato evitando generalizzazioni che possono penalizzare i piccoli *player* e cercando di sostenerli in modo che continuino a competere. L'idea è quindi quella di, invece che forzare gli *incumbent* a licenziare o condividere i loro modelli o dati, incoraggiare la formazione di un robuso ecosistema di ricercatori accademici e aziende di diverso tipo che competano per lo sviluppo della prossima generazione di modelli e AI *tool* (Pierre Azoulay, 2024). In America, ad esempio numerose sono le iniziative a sostegno delle piccole *startup* o centri di ricerca universitari nel settore dell'AI che comprendono dedicate agenzie di finanziamento della ricerca che si impegnano a co-finanziare aziende nella fase iniziale di sviluppo premiando poi il raggiungimento di traguardi significativi (Pierre Azoulay, 2024). Questo approccio di sostegno flessibile ai nuovi entranti è sicuramente più supportato dalla teoria economica rispetto ad un sistema che prevede una regolamentazione rigida *ex-ante* che limita gli *incumbent* imponendo obblighi di accesso

e condivisione dei dati. I sostenitori di quest'ultimo interpretano le esternalità di rete, la presenza di pochi operatori (o meglio, di poche piattaforme di sviluppo e acquisto dei prodotti AI), le teorie economiche generali sulle economie di scala e scopo, che prevedono fallimenti di mercato e pratiche escludenti, come fattori critici che creano alte barriere all'ingresso e allo sviluppo riscontrabili in tutte le fasi della *value chain* dell'AI e che, quindi, rendono necessario un intervento rigido e preventivo su tutto il settore. Tuttavia, come dimostrato, imporre un obbligo di *sharing ex-ante* su infrastrutture come i *foundation model* potrebbe risultare un limite eccessivo al diritto di chi le sviluppa, di escludere dai frutti dei propri investimenti e, più in generale dal successo imprenditoriale, coloro che non hanno in alcun modo contribuito ad ottenerli (Preta, 2024). Ne discende che qualunque obbligo di accesso o condivisione dei dati debba essere valutato con particolare attenzione e prudenza, al fine di evitare di produrre effetti negativi sul mercato (Bourreau, Streel, & Graef, 2017). L'articolo del *DMA* che impone ai grandi motori di ricerca. *Gatekeeper,* di condividere i *search data* con i *competitors* offre un esempio di regolamento creato *ad hoc* che, seppur imponga la condivisione di dati, non mina gli investimenti in ricerca e sviluppo. In Europa il regolamento principale per quanto concerne l'accessibilità e la condivisione dei dati nel settore dell'intelligenza artificiale è il *EU AI Act*[89]*,* del 2023, che si pone l'obiettivo di rimuovere le barriere di accesso ai dati per il settore pubblico e privato, cercando di preservare gli incentivi alla *data production* e sviluppando un ambiente competitivo e corretto che benefici aziende e cittadini europei. L'Atto mira a prevenire lo sfruttamento abusivo di possibili squilibri contrattuali tra le parti riguardanti l'accesso e l'utilizzo dei dati raccolti e della proprietà intellettuale (Articolo 13(5)(b)). Il documento conferisce, inoltre, alle autorità pubbliche il diritto di accedere ai dati proprietari aziendali in situazioni critiche che possano costituire un concreto rischio per il pubblico e intende incentivare l'interoperabilità tra i servizi AI e *cloud* per ridurre gli *switching cost* per i clienti (Articoli 29 e 30) (Autorité de la concurrence, 2024).

Va inoltre sottolineato che, sebbene l'imposizione del *data sharing* sia una norma ampiamente discussa tra i *policymakers,* può in alcune circostanze trovarsi in contrasto con molte regolamentazioni già in essere sulla *governance* dei dati, che mirano a limitare

[89] https://artificialintelligenceact.eu/

il trasferimento di dati tra le imprese per una questione di *privacy*, come dimostra il Regolamento Generale sulla Protezione dei Dati dell'UE (GDPR)[90] (Pierre Azoulay, 2024).

Nel modello di Hagiu e Wright, volendo essere un *framework* generale per tutti i sistemi *data-enabled learning*, gli autori analizzano come politica di condivisione di informazioni esclusivamente le implicazioni del *data sharing;* tuttavia, nell'interpretazione proposta per il mercato dei *foundation model* il discorso necessita un'ulteriore generalizzazione. Nel mercato dei *FM,* infatti, la condivisione dei dettagli circa i dati utilizzati può rientrare nella definizione generale di "apertura" come abbiamo descritto nel capitolo 2 a proposito dell'*open source.* Comparabili alle politiche di *data sharing* ci sono una serie di misure che incentivano la complessiva trasparenza del modello, non solo riguardo ai dati, ma anche la loro elaborazione e preparazione, o riguardo i parametri, meta parametri e l'infrastruttura utilizzata. Alcune di queste proposte di regolamentazione degne di nota includono l'*EU AI Act*[61] in Europa e il *AI Foundation Model Transparency Act*[91] negli Stati Uniti (Bommasani, et al., 2024). Inoltre, i *FM* non vanno interpretati come prodotto *stand alone* ma inseriti nell'intera catena del valore dell'AI e lo studio delle dinamiche competitive non può prescindere dall'analisi della competizione a valle. A tal proposito i ricercatori Fasheng Xu, Xiaoyu, Wei Chen e Karen Xie propongono un modello di competizione dinamica a due stadi per studiare l'impatto del livello di apertura dei *foundation model* sulle decisioni di *deployment* di due *competitor* nel mercato a valle (*i.e.* la decisione circa il timing di ingresso a mercato, il livello di investimento in *fine tuning* e il prezzo offerto ai clienti finali) (Xu, Wang, Chen, & Xie, 2024). In questo contesto il livello di apertura del modello gioca un ruolo chiave creando spazio per un disallineamento di preferenze tra i due *deployer,* che crea attriti nella *value chain.* Con l'aumento del livello di apertura del modello sottostante il primo *deployer* è incentivato a ridurre strategicamente il proprio

[90] https://www.garanteprivacy.it/il-testo-del-regolamento

[91] https://www.congress.gov/bill/118th-congress/house-bill/6881/text

sforzo di *fine tuning*, con l'intento di dissuadere il *deployer* 2 da una *early adoption* [92] al fine di appropriarsi dei profitti del primo periodo in cui è monopolista. Questo comportamento strategico conduce ad una ridotta espansione del mercato, con conseguenze negative sui profitti del secondo *deployer* e del *developer* a monte, sul *surplus* dei consumatori e sul benessere sociale complessivo. Questa dinamica viene definita "*openess trap*": un *range* di livelli di apertura media in cui il social *welfare* risulta inferiore a quello che si avrebbe con un'apertura nulla. Questa trappola rappresenta una situazione paradossale in cui un incremento della trasparenza del modello può condurre a risultati subottimali per tutti gli attori coinvolti. Tutto ciò sfida la tradizionale convinzione secondo cui una maggiore apertura sia sempre un vantaggio per le aziende *deployer* di intelligenza artificiale, in quanto aiuta a ridurre i costi di *fine tuning*. Come immagine esplicativa si rimanda alla Figura 8.

Per evitare di cadere in questa "trappola" nel regolamentare il livello di apertura attraverso la divulgazione pubblica di dettagli tecnici, i *policymaker* devono essere consapevoli di questo rischio. La necessità di trasparenza, al di là delle mere implicazioni concorrenziali è importante nell'assicurare la sicurezza del modello e il rispetto delle leggi circa la sicurezza e la *privacy* dei dati, pertanto una possibile soluzione potrebbe essere la registrazione, ossia l'obbligo per le aziende di condividere le informazioni direttamente e in maniera esclusiva con l'ente regolatore e non pubblicamente (Xu, Wang, Chen, & Xie, 2024).

Quanto appena descritto vuole dimostrare e confermare due punti importanti di questa tesi:

i. La divulgazione di dettagli tecnici e la condivisione dei dati utilizzati, seppur in un contesto in cui la tecnologia di apprendimento si fonda su essi, che ne sono l'*input* primario, e può generare *feedback loop* e/o effetti di rete, non necessariamente aumenta il benessere dei consumatori, tanto meno il *welfare* complessivo.

[92] Un grande sforzo di *fine tuning* del primo *deployer* agevola il secondo grazie ad un effetto di *spillover* di conoscenza. Pertanto diminuire gli sforzi di *fine tuning* di 1 disincentiva 2 ad entrare subito a mercato ma lo spinge ad aspettare un momento in cui avrà accumulato sufficiente *spillover* per poter fare profitti. Allo stesso tempo la presenza di entrambi nel mercato ne aumenta l'attrattività e la domanda aggregata, avendo quindi un effetto benefico di espansione del mercato.

ii. Le classiche proposte di regolamentazione per l'AI (che richiamano quelle dei mercati digitali) che pongono al centro i dati come unica fonte di barriera all'ingresso da abbattere con la condivisione, non sono la panacea di tutte le questioni *AI-related.* Regolamentare il mercato dell'intelligenza artificiale ha le proprie sfide e i propri problemi di allineamento di incentivi unici per questo mercato (Guha, et al., 2023). I *policy maker* non dovrebbero pertanto affrettare la regolamentazione senza considerare la fattibilità tecnica, i *trade-off* economici ed eventuali ripercussioni non desiderate sull'intera filiera dell'AI.

Per continuare l'analisi delle dinamiche competitive sotto diversi scenari regolatori si considera ora l'introduzione di una rigida regolamentazione in materia di *privacy.* Una tale politica rende più difficile per le aziende ottenere dati, limitando la tipologia di dati utilizzabili o semplificando il processo di scelta per i consumatori di non condividerli. Nel modello descritto, ciò implica che le aziende possono utilizzare solo una frazione dei nuovi dati raccolti a seguito dell'implementazione della nuova *policy*, rallentando così il tasso di apprendimento[93], pur mantenendo invariati i valori soglia $\underline{N_i}$ e $\overline{N_i}$.

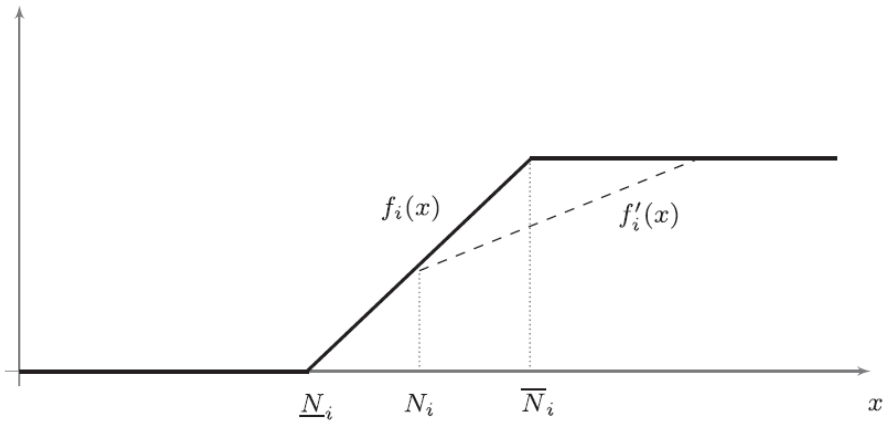


Figura 28: Effetti sulla curva di apprendimento di una politica di privacy più stringente. Source (Xu, Wang, Chen, & Xie, 2024)

[93] La nuova curva di apprendimento $f_i'(x)$ è, pertanto, meno ripida a seguito dell'implementazione della nuova politica

A causa della diversa posizione delle due aziende lungo la curva di apprendimento le conseguenze del cambiamento della funzione sono disallineate e tendoo a rafforzare il vantaggio competitivo dell'*incumbent* a discapito dei nuovi *competitor*. Quando *I,* già presente nel mercato, si trova più vicina alla soglia di apprendimento superiore e/o *E* ha una funzione f_E più ripida, rallentare il tasso di apprendimento ha un effetto relativamente più negativo sull'entrante. Ciò riflette il fattto che *E* ha più da guadagnare dai nuovi dati utente e, di conseguenza, più da perdere a seguito delle restrizioni. Contrariamente a quanto si potrebbe pensare, quindi, in presenza di *data policy* stringenti, l'*incumbent* (o più genericamente l'azienda vincente) beneficia di un peggioramento generale della tecnologia rappresentato dalla minor pendanza della curva. Come conseguenza il *customer surplus* diminuisce sia in termini di minor qualità sul mercato sia come minori sussidi offerti dall'azienda perdente. Sebbene questo risultato non modelli dierettamente i benefici complessivi per i consumatori derivanti da una maggiore *privacy*, esso vuole sottolineare come una politica di *privacy* più rigorosa, introdotta per la tutela della clientela, potrebbe avere effetti collaterali non intenzionali sulle dinamiche competitive, penalizzando principalmente le giovani entranti.

Le discussioni in tema di *privacy* sono più che mai attuali e controverse. In un contesto dove i dati, soprattutto i dati precisi e aggiornati, sono il principale *input* allo sviluppo tecnologico i *policymakers* si trovano ad affrontare un *trade-off* complesso: da un lato le discussioni etiche sulla proprietà dei dati come bene e il diritto dell'utente di deciderne l'utilizzo, dall'altro il progresso tecnologico (e competitivo) del mercato. Ad oggi, in Europa, il principale riferimento legale in merito è il GDPR (altri regolamenti simili nel resto del mondo) che decreta diversi principi guida di *data governance* e *privacy.* In particolare viene stabilito il "principio di limitazione dello scopo" con cui gli individui hanno il diritto di essere informati sugli scopi precisi del trattamento dei propri dati personali nel momento stesso della raccolta, avendo la possibilità di negarne il consenso. Se questi scopi dovessero cambiare, ad esempio a seguito di una acquisizione strategica *data driven* in cui i dati diverrebbero *input* di *training* per un modello AI*,* l'azienda dovrebbe richiedere nuovamente l'autorizzazione dell'utente per l'utilizzo dei dati ai nuovi fini, con tutte le difficoltà che ciò comporta (Usercentrics, 2023). L'attenzione alla *privacy* è centrale nelle procedure di *merger control* delle transazioni *data driven* già dalla nascita delle grandi piattaforme del Web2 come testimoniano le indagini della

commissione europea sulle acquisizioni di Google/DoubleClick o Facebook/Whatsapp (KARAÇAM, 2019).

Tralasciando la questione intrasferibilità dei dati dovuto alle regolamentazioni sul limite di scopo descritto, applicabile solo in Europa e difficilmente tracciabile, il modello economico descritto può essere utilizzato anche per analizzare le ripercussioni sociali e competitive delle acquisizioni strategiche; in particolare riferendosi alle *killer data acquisition* con cui una azienda acquisisce una rivale (o il suo *dataset*) con il solo scopo di evitarne la concorrenza o precluderne l'acquisto ad un *competitor*, preservando il proprio vantaggio competitivo.

Si considera uno scenario in cui entrambe le aziende I ed E abbiano l'opportunità di acquisire in maniera esclusiva una certa quantità $N_A > 0$ di dati per prevedere quale delle due sarà disposta a pagare di più e l'efficienza del risultato. Per ciascuna azienda ci sono potenzialmente due incentivi a favore dell'acquisizione: (i) utilizzare i dati per progredire lungo la curva di apprendimento e crescere i propri profitti e (ii) proteggere la propria posizione nel mercato e i profitti attuali negando dati preziosi al concorrente. Ad esempio, se l'*incumbent* si trova nei pressi *dell'upper bound* della curva, ma l'entrante dispone di una tecnologia di apprendimento superiore, l'incentivo del primo sarà proteggere i profitti, del secondo il miglioramento tecnologico[94]. Si crea, in questo contesto, un valore soglia $s_E - s_I$ di equilibrio sotto il quale E acquisterà sempre i dati, viceversa lo farà I aprendo la possibilità a grandi inefficienze. Si supponga, senza perdita di generalità, che I abbia già raggiunto la soglia $\overline{N}_I$; dalla condizione di equilibrio risulta che I continua a vincere i dati, anche se non ne trae più alcun beneficio tecnico ostacolando di conseguenza l'innovazione e lo sviluppo dell'azienda concorrente e i benefici pubblici che ne deriverebbero, da cui il nome *killer acquisition.*

Quanto appena descritto giustifica l'attenzione dei *policymaker* per le *data-partnership* o le *data-driven acquisition* tipiche del mercato dell'AI generativa e, più in generale, suggerisce un maggior controllo sulle dinamiche e la *ratio* che sottostanno le *M&A* e

[94] In particolare, se I acquisisce N_A, allora la soglia di E per vincere aumenta a $s_E - s_I = \Delta(N_I + N_A, N_E)$, viceversa diminuisce a $s_E - s_I = \Delta(N_I, N_E + N_A)$ ricordando che $\Delta(N_I + N_A, N_E) = \Delta(\overline{N}_I, N_E)$ ogni volta che $N_I + N_A \geq \overline{N}_I$ e $\Delta(N_I, N_E + N_A) = \Delta(N_I, \overline{N}_E)$ ogni volta che $N_E + N_A \geq \overline{N}_E$.

partnership nel settore (Carugati & Kar, 2024) (Maggiolino & Zoboli, 2024) (KARAÇAM, 2019). A livello *upstream*, le autorità garanti della concorrenza si occupano di evitare che grandi *incumbent* limitino l'accesso a *input* chiave (*i.e.* dati, potenza di calcolo, servizi cloud..) a possibili entranti. *Data partnership* esclusive, così come le *killer acquisition* possono quindi rafforzare queste preoccupazioni. Tuttavia va sottolineato che il reale rischio per la concorrenza è costituito dall'esclusività dei contratti in essere e non dalla transazione in sé (Carugati & Kar, 2024). Ricordando quanto descritto nel capitolo 2, i dati necessari all'allenamento dei *foundation models* vengono primariamente estratti da dataset pubblici o direttamente dal web e sono limitatamente costituiti da dataset proprietari. Inoltre, complice la, per definizione, generalità dei *FM,* è raro che siano indispensabili dataset unici e non replicabili. Così come nell'esempio dei sistemi di radiologia con supporto AI: una azienda *developer* che stringe accordi di utilizzo esclusivo dei loro *dataset*, non impedisce ad altre aziende di procurarsi dati simili da altre fonti. Seppur, quindi, il modello teorizzi il verificarsi di possibili acquisizioni *killer* in un mercato caratterizzato da un prodotto *data-enabled learning* e di come esse possano rafforzare la posizione dominante dell'*incumbent,* nel preciso mercato dei *FM* questa conclusione richiede un approfondimento maggiore *ad hoc* sul singolo caso che dimostri l'esclusività dell'acquisto.

A conclusione del presente capitolo si può quindi affermare che, contrariamente al pensiero comune circa la forza degli *user data* nell'ecosistema dell'intelligenza artificiale, i *feedback loop* e gli effetti di rete che essi creano potrebbero non essere i principali determinanti del vantaggio competitivo di una azienda, tanto meno essere il primario *driver* di competizione. Per tanto, i *policy maker* devono evitare la corsa alla regolamentazione *ex ante* del mercato e la proiezione di vecchie paure (nate dalla concentrazione del Web2), in questo nuovo contesto competitivo, che ha le sue regole e le sue dinamiche e deve essere studiato separatemente e senza pregiudizi.

Capitolo 4: *Policy Consideration*

Se le dinamiche rimanessero come sono ad oggi ci sarebbero pochi motivi di preoccupazione circa la competizione nel mercato, tuttavia i rapidi sviluppi rendono necessario un attento *monitoring* nel prevenire tendenze oligopolistiche. Nella regolamentazione del mercato dei *FM,* inserendosi nel più ampio spettro di prodotto digitale vengono in aiuto molte delle disposizioni già in essere per i servizi ausiliari, con alcune precisazioni necessarie.

La principale preoccupazione per le autorità è che si verifichi nuovamente quanto accaduto con le grandi piattaforme digitali del Web2 o i *search engine* agli inizi del secolo. Similmente al mercato dell'AI generativa, anche questi mercati videro inizialmente una accesa competizione per poi, anni dopo, vedersi consolidati nelle mani di pochi *player* a causa di tre principali forze: (i) significative economie di scala e scopo, (ii) effetti di rete (e *data feedback loop*) e (iii) l'inerzia al cambiamento dei consumatori. Tre dinamiche che si riscontrano anche nel mercato dei *foundation models,* da cui le preoccupazioni. Si possono evidenziare infatti similitudini e differenze la cui analisi è importante nello stabilire il rischio di concentrazione.

In primis la rapida crescita dei costi e dell'infrastruttura necessari per lo sviluppo dei modelli di frontiera restringe il numero di aziende con *asset* finanziari sufficienti; tuttavia, a controbilanciare questa forza si trovano la crescita complessiva del mercato per l'AI e l'ottimizzazione dei costi data dall'innovazione.

Inoltre, a differenza degli *search engine* che hanno un utilizzo ben definito, i *FM* possono essere implementati in un'infintà di applicazioni, varianti e mercati (che necessitano diversi livelli di *fine tuning*), rendendo non credibile l'idea che un solo *player* possa coprire l'intera domanda.

Terzo, il ruolo dei *network effect* è significativamente più ridotto se paragonato alle piattaforme digitali il cui valore cresce con la diffusione più che con le *performance, main driver,* invece, per i *FM.* Come ampiamente discusso nel capitolo precedente, l'esistenza di *data feedback loop* non garantisce necessariamente dominanza nel mercato e non è neanche sempre chiaro quanto dei dati e dei *feedback* raccolti dai *deployer* a valle risalga ai *developer* diventando nuovi dati per l'allenamento. *Training data* che,

principalmente, sono non proprietari e non unici, ossia che non derivano da accordi esclusivi, limitandone quindi il potere in termini di barriere all'ingresso.

Inoltre, seppur possano essere comparabili a sistemi operativi per le applicazioni che li implementano, i *FM* non godono dei tipici *network effect cross-side.* Gli user, infatti, non devono adottare un singolo modello in quanto, ad oggi, non sono implementati direttamente su *device* personali, ma possono utilizzare diverse applicazioni con diversi *FM* sottostanti. Teoricamente, i fornitori di modelli fondativi potrebbero evolversi verso una simile configurazione, offrendo dispositivi specifici progettati per ospitare esclusivamente applicazioni AI costruite sui propri modelli. In questo modo, potrebbero emergere effetti di rete tra utenti e sviluppatori di *app*. Tuttavia, appare uno scenario poco probabile. Non si registra, quindi, ad oggi un attaccamento naturale degli utenti finali al *brand* di *FM* dietro le applicazioni di loro interesse, quanto bensì alle *performance* e ai costi. (Hagiu & Wright, 2024; Korinek & Vipra, 2024).

Ultimo fattore, ma non per importanza, la presenza dell'*open source,* che gioca un ruolo fondamentale nel mantenere alto il livello di innovazione e di competizione tra le aziende *leaders.*

I fattori appena discussi suggeriscono una scarsa tendenza dei *FM* nel risultare in un mercato altamente concentrato. Non mancano, comunque, visioni contrastanti che sottolineano i rischi legati alla struttura del mercato e alle forti economie di scala e alti costi di *compute* (Jon Schmid, 2024).

Le principali preoccupazioni sono le medesime dei mercati digitali tradizionali: abuso di posizione dominante, *tying e bundling* di prodotti e servizi, aumento dei costi di *switching* con conseguente effetto di *lock in,* dinamiche escludenti, discriminatorie o di *self preferencing,* contratti esclusivi, acquisizioni ostili o *merge/partnership* che minano la competizione e danneggiano i consumatori, e infine la questione dei dati, della *privacy* e del consenso al loro utilizzo.

In merito a queste pratiche, numerose sono le disposizioni europee che coordinano e monitorano i mercati digitali assicurandosi di bilanciare gli incentivi all'innovazione e agli investimenti, con una corretta tutela dei consumatori.

La presente tesi vuole essere una dimostrazione di quanto una *over regulation* mini il progresso innovativo e possa incentivare la concentrazione di mercato, rappresentando

un fardello troppo ingombrante per piccoli entranti (Yun, 2024). L'*EU AI Act,* ad esempio, si distingue dai precedenti testi regolatori per la sue disposizioni in materia di "*General purpose AI*" (GPAI), che comprendono i grandi LLMs e *foundation model* di diverse tipologie. Tuttavia, in aggiunta ai rischi connessi all'*openess trap* dovuti ad un alto livello di trasparenza richiesto*,* i rigorosi requisiti di documentazione tecnica e di *risk management* potrebbero rappresentare un onere eccessivo per le piccole aziende o *startup* (Korinek & Vipra, 2024).
Si suggerisce, invece, l'adozione di politiche meno restrittive in termini di obblighi di condivisione e licenza, preferendo l'impegno verso la creazione di un ecosistema robusto e dinamico di ricercatori accademici e aziende che collaborino nello sviluppo di modelli e strumenti AI di nuova generazione.

Sebbene l'AI presenti grandi opportunità, è importante procedere con cautela nella regolamentazione *ex-ante*, evitando di compromettere la ricerca e la competitività del mercato. Dato anche l'ampio spettro di soluzioni diverse in termini di *business model*, risorse finanziarie e mercato di riferimento, si consiglia, ai *policymakers* di analizzare *ex-post* specifici casi uno ad uno piuttosto che adottare misure generali che potrebbero ostacolare lo sviluppo del settore.

Prima di affrettarsi a regolamentare in maniera restrittiva il mercato è necessario aumentare il livello di comprensione del settore e delle sue dinamiche complesse accertandosi e quantificando i reali rischi per la società e l'affidabilità dei sistemi AI. Infatti, la maggior parte dei rischi di mercato potrebbero essere già coperti dall'attuale regolamentazione che governa i settori complementari allo sviluppo dei *FM* (Guha, et al., 2023).

Inoltre, è importante creare una nuova tassonomia *ad hoc* per il settore o rivedere le definizioni precedenti. Ciò è particolarmente utile in merito alle *partnership,* dato il loro ruolo centrale nelle dinamiche competitive. È necessario soprattutto definire la differenza in termini di *policy* con *merge/acquisition* e il loro collocamento (in Europa) nelle procedure di controllo "EU Merge Control" (Carugati, 2024b).

Un esempio in tema è l'indagine condotta dal CMA sulla *partnership* tra Microsoft e OpenAi, i cui risultati citano comunque la parola "*merge"* e sono stati ottenuti con uno *framework* di analisi inizialmente crato per le operazioni *M&A*. La decisione finale

pubblicata lo scorso 5 marzo 2025, dichiara, infatti, che "*Microsoft's partnership with OpenAI does not qualify for investigation under the merger provisions of the Enterprise Act 2002*", ossia che la *partnership* non costituisce una "*relevant merge situation*" [95].

Vi sono, tuttavia, altri rischi sollevati dall'intelligenza artificiale che risultano nuovi nello scenario regolatorio; come, ad esempio, il ruolo degli algoritmi di prezzo nel facilitare la collusione tacita tra imprese (anche detta collusione algoritmica). Sebbene la capacità di colludere (anche in modo tacito) in contesti dinamici non sia una novità, il fissaggio dei prezzi tramite algoritmi solleva nuove questioni relative alla rilevazione e applicazione delle norme e alla responsabilità. È possibile considerare l'uso di tali algoritmi come una pratica facilitante? Se i concorrenti aumentano i prezzi utilizzando algoritmi forniti da un comune fornitore, ciò costituisce una forma di cartello *hub-and-spoke*? In tal caso, il fornitore dell'algoritmo è responsabile di eventuali accuse di collusione? I fornitori di questi algoritmi di *pricing* dovrebbero essere soggetti a regolamentazioni specifiche circa i dati degli utenti? È lecito che un'azienda condivida ai propri concorrenti all'utilizzo di un particolare algoritmo dato che ciò potrebbe essere utilizzato per coordinarsi con il medesimo software?" La risposta a queste domande esula dalle finalità della presente tesi e si rimanda alla crescente letteratura sul tema (Gans, 2024) (Hanspach & Galli, 2024).

Un rischio più generale riguarda l'affidabilità legale dei sistemi AI, le cui pratiche scorrette e/o anticompetitive potrebbero essere più difficili da individuare. Inoltre, seppur un applicativo AI possa essere progammato per rispettare le leggi sulla concorrenza, potrebbe comunque trovare modi alternativi per aggirare le *policy* nel tentativo di massimizzare i profitti dell'azienda.

L'opacità intrinseca di molti modelli di *deep learning* rappresenta una ulteriore sfida per la supervisione regolamentare. La loro natura "*black-box*" rende, infatti, difficile per le autorità stabilire se un'intelligenza artificiale persegua obiettivi anti-competitivi o se stia correttamente rispondendo in maniera strategica alle dinamiche di mercato. Questo scenario solleva importanti interrogativi riguardo all'intersezione tra AI e il diritto della concorrenza. "È possibile sviluppare simulazioni o altri metodi analitici per testare se i modelli di AI producono risultati anti-competitivi? È possibile allenare i modelli AI

[95] https://www.gov.uk/cma-cases/microsoft-slash-openai-partnership-merger-inquiry

affinchè aderiscano sia alla lettera che allo spirito della legge sulla concorrenza?" (Hagiu & Wright, 2024)

A tutte queste domande si aggiungono quelle sui dati, la loro proprietà, il diritto alla loro divulgazione e quelle riguardanti i diritti di *privacy* e *copyright.* I regolatori potrebbero, ad esempio, dover considerare alcuni dataset come strutture essenziali per lo sviluppo dei *FM*. Potrebbero anche essere necessarie nuove modalità di compensazione dei *creators* per l'utilizzo dei loro contenuti come *training data.*

Conclusioni

La presente tesi aveva l'obiettivo di esplorare le dinamiche competitive nel mercato dei *foundation model*, evidenziandone sfide e opportunità con l'intento di stabilire se le forti economie di scala e gli ingenti investimenti *hardware* necessari costituiscano una barriera all'ingresso così alta da rendere la minaccia di oligopolio o monopolio sempre più reale.

Il mercato circoscritto dei *FM* è, oggi, altamente dinamico e competitivo, con numerosi *player* globali che si sfidano in termini di prezzo e soprattutto di *performance*, favorendo ricerca e innovazione. Oltre alle sette grandi *big tech* vi sono, infatti, numerose *startup* e progetti *open source* che, supportati da finanziamenti pubblici e privati, competono attivamente e a, quasi, pari livello.

Nonostante ciò, persistono le preoccupazioni per la concentrazione nei mercati a monte dei *FM*, in particolare riguardo agli *hyperscalers* e la possibilità che sfruttino il loro potere di mercato nel settore *cloud* e infrastruttura *web* per rafforzare la propria presenza in altri livelli della *value chain* incorrendo in pratiche anti-competitive come *tying*, *bundling*, pratiche discriminatorie o di *self preferencing* sui loro *marketplace*. Altrettante sono le attenzioni per Nvidia di cui si teme un *leveraging* della sua posizione dominante dal settore semiconduttori e GPUs agli altri *layer* dello *stack* tecnologico.

Per quanto riguarda i dati, la generalità dei *foundation models*¸ implica che, nella fase di *training*, vengano utilizzati prevalentemente dati pubblici, ottenibili dal *web* o dati proprietari, non unici, rendendo l'accesso ai dati una barriera all'ingresso non insormontabile e, soprattutto, rendendo superflua l'introduzione di nuove politiche di *data sharing*, che, al posto di facilitare l'ingresso di nuovi entranti e stimolare il mercato, potrebbero risultare in una diminuzione di *social welfare* conseguente a dinamiche di *free riding*. Queste ultime, infatti, rappresenterebbero un disincentivo per le grandi piattaforme a investire nella raccolta e nel trattamento di questi dati, compromettendo una fonte importante per l'innovazione.

Un'altra possibile barriera indagata in questa tesi è costituita dai *data feedback loop* con cui un modello, raccogliendo *input* e dati dai propri *user* riesce a migliorarsi ed attrarre, di conseguenza nuovi utenti. Si dimostra come, tuttavia, questi effetti sono limitati e di per sé non costituiscono a priori una fonte di vantaggio competitivo a lungo termine. Anche gli effetti di rete tradizionali, seppur più stabili e forti non sono, in questo mercato

un principale ingrediente del potere di mercato, a differenza del ruolo occupato nelle piattaforme digitali.

In questo contesto le classiche *policy* introdotte con il DMA, circa la regolamentazione dei colossi digitali, che considera la trasparenza il principio cardine della competizione e della sicurezza nel mercato, non sono applicabili attendendovi i medesimi risultati. Tra i *foundation model*, infatti, una richiesta troppo elevata di divulgazione tecnica riguardo architettura e dati utilizzati indurrebbe le aziende verso minori investimenti, con un conseguente rallentamento dell'innovazione e una riduzione del benessere sociale, priorità dei *policymakers*.

È essenziale, quindi, che questi ultimi e le autorità garanti non si affrettino a regolamentare il settore per paura di una futura concentrazione ma si prendano il giusto tempo per analizzare e comprendere le dinamiche e le forze che danno forma al mercato considerando la *big picture* dell'intera *value chain*, in modo da non frenare lo sviluppo scientifico e gravare in maniera eccessiva sulle piccole *startup*. E' necessario, infatti, che le attenzioni vengano canalizzate nell' incentivare la creazione di un ecosistema di università, ricercatori e imprese che attivamente collaborino per la creazione della prossima generazione di *foundation model*.

Il settore è ancora molto giovane e altamente dinamico per cui si ritnengono necessarie ulteriori ricerche e analisi interdisciplinari: economiche, legali e tecniche, che monitorino il mercato e le sue dinamiche e che possano, in futuro, fornire risposte più certe.

Riferimenti

Acemoglu, D. (2024). *The Simple Macroeconomics of AI.* National Bureau of Economic Research. doi:10.3386/w32487

Alcaraz, M. (2024, Maggio 23). *« Méga accord » entre OpenAI et News Corp*. Retrieved from LesEchos: https://www.lesechos.fr/tech-medias/intelligence-artificielle/mega-accord-entre-openai-et-news-corp-2096599

Alec Radford, J. W. (2019). *Language Models are Unsupervised Multitask Learners.* Retrieved from openai.com: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Altchek, A. (2024, Novembre 21). *This chart shows how crazy fast the value of Elon Musk's xAI has risen in 16 months.* Retrieved from Business Insider: https://www.businessinsider.com/elon-musk-xai-startup-valuation-history-chart-2024-11

Amer, M. (n.d.). *Evaluating Outputs.* Cohere.com/llmu. Retrieved from https://cohere.com/llmu/evaluating-llm-outputs

Andreas Liesenfeld, A. L. (2023). *Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators.* Retrieved from arXiv:2307.05532

Anil, R. (2023). *PaLM 2 Technical Report.* doi:https://doi.org/10.48550/arXiv.2305.10403

Anthropic. (2024). *Claude 3.5 Sonnet Model Card Addendum.* Retrieved from https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf

Anton Korinek, J. V. (2023). *Market concentration implications of foundation models: the invisible hand of ChatGPT.* Brookings, Regulation and Markets. Retrieved from https://www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt/

Ashish Vaswani, N. S. (2017). *Attention Is All You Need.* doi: arXiv:1706.03762

Autoridade da concorrencia. (2023). *Competition and generative artificial intelligence.*

Autorité de la concurrence. (2024). *on the competitive functioning of the generative artificial intelligence sector.* Retrieved from https://www.autoritedelaconcurrence.fr/en/opinion/competitive-functioning-generative-artificial-intelligence-sector

Bajari, P. V. (2019). *he Impact of Big Data on Firm Performance: An Empirical Investigation.* AEA Papers and Proceedings 109: . doi:DOI: 10.1257/pandp.20191000

Bass, D. (2023). *OpenAI Needs Billions to Keep ChatGPT Running. Enter Microsoft.* Bloomberg. Retrieved from https://www.bloomberg.com/news/articles/2023-01-26/microsoft-openai-investment-will-help-keep-chatgpt-online

Beatman, A. (2023). *Introducing Azure OpenAI Service On Your Data in Public Preview.* Retrieved from https://techcommunity.microsoft.com/blog/azure-ai-services-blog/introducing-azure-openai-service-on-your-data-in-public-preview/3847000

Biagio, S. (2024). Intelligenza artificiale, Amazon punta altri 4 miliardi su Anthropic. *ilsole24ore*. Retrieved from https://www.ilsole24ore.com/art/intelligenza-artificiale-amazon-punta-altri-4-miliardi-anthropic-AGZDE2KB

Black, D. (2020). *Microsoft's Massive AI Supercomputer on Azure: 285k CPU Cores, 10k GPUs.* Retrieved from https://www.hpcwire.com/2020/05/20/microsofts-ai-supercomputer-on-azure-combinations-of-perceptual-domains/

Bommasani, R., Kapoor, S., Klyman, K., & Longpre, S. (2023). *Considerations for governing open foundation models.* Stanford University: Human-Centrered Artificial Intelligence HAI. Retrieved from https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Zhang, B. X., & Liang, P. (2024). *The Foundation Model Transparency Index v1.1.* Stanford Center for Research on Foundation Models (CRFM) and Stanford Institute for Human-Centered Artificial Intelligence (HAI). Retrieved from https://crfm.stanford.edu/fmti/May-2024/index.html

Bommasani, R., Klyman, K., Longpre, S., Xiong, B., Kapoor, S., Maslej, N., . . . Liang, P. (2024). Foundation Model Transparency Reports. doi:https://doi.org/10.48550/arXiv.2402.16268

Bourreau, M., Streel, A. d., & Graef, I. (2017). Big Data and Competition Policy: Market power, personalised pricing and advertising. *CERRE*, 35-37. Retrieved from https://cerre.eu/publications/big-data-and-competition-policy/

Cabral , L., & Riordan, M. (1994, Settembre 5). The Learning Curve, Market Dominance, and Predatory Pricing. *The Econometric Society, 62*(5), 1115-1140. doi:https://doi.org/10.2307/2951509

Carugati, C. (2023). *COMPETITION IN GENERATIVE ARTICIFIAL INTELLIGENCE FOUNDATION MODELS.* Bruegel. Retrieved from https://www.jstor.org/stable/resrep52128

Carugati, C. (2023b). *The competitive relationship between cloud compting e genereative AI.* Working Paper 19/2023, Bruegel. Retrieved from https://www.bruegel.org/working-paper/competitive-relationship-between-cloud-computing-and-generative-ai

Carugati, C. (2024b). *How Should Europe Revamp Merger Policy for Non-Notifiable Deals?* Retrieved from https://www.digital-competition.com/comment/how-should-europe-revamp-merger-policy-for-non-notifiable-deals%3F

Carugati, C., & Kar, N. (2024). *Assessing competitive dynamics of AI partnership.* doi:https://dx.doi.org/10.2139/ssrn.5025769

Carugati, C., & Kar, N. (2024). *Assessing the Competitive Dynamics of AI Partnership.* doi:https://dx.doi.org/10.2139/ssrn.5025769

Character.AI. (2024). *Our Next Phase of Growth*. Retrieved from https://blog.character.ai/our-next-phase-of-growth/

CMA . (2024). *AI Foundation Models: Technical update report.* Competition & Market Authority. Retrieved from https://www.gov.uk/government/publications/ai-foundation-models-update-paper

CMA. (2025). *Cloud services market investigation.* Retrieved from https://www.gov.uk/cma-cases/cloud-services-market-investigation#full-publication-update-history

Collinas, B. (n.d.). *Death By API: Reddit Joins Twitter In Pricing Out Apps*. Retrieved from Forbes: https://www.forbes.com/sites/barrycollins/2023/06/01/death-by-api-reddit-joins-twitter-in-pricing-out-apps/

Competition & Market Authority. (2023). *AI Foundation Models: short version.* CMA.

Cottier, B. R. (2024). *'The rising costs of training frontier AI models'.* Retrieved from arXiv. https://arxiv.org/abs/2405.21015

CRFM. (2024). *On the Opportunities and Risks of Foundation Models.* Stanford: Center for Research on Foundation Models, Stanford Institute for Human-Centered Artificial Intelligence (HAI).

Currier, J. (2020). *What Makes Data Valuable: The Truth About Data Network Effects*. Retrieved from nfx.com: https://www.nfx.com/post/truth-about-data-network-effects

Dan Hendrycks, C. B. (2021). *Measuring Massive Multitask Language Understanding.* International Conference on Learning Representations. Retrieved from https://arxiv.org/abs/2009.03300

David Gray Widder, S. W. (2023). *Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI.* Retrieved from https://ssrn.com/abstract=4543807

DeepSeek-AI, A. L. (2024). *DeepSeek-V3 Technical Report.* arXiv:2412.19437.

Dylan Patel, A. A. (2023). *Google "We Have No Moat, And Neither Does OpenAI" Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI.* Retrieved from https://semianalysis.com/2023/05/04/google-we-have-no-moat-and-neither/

ecosystem graphs. (n.d.). Retrieved from https://crfm.stanford.edu/ecosystem-graphs/index.html?mode=table

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.* doi:https://doi.org/10.48550/arXiv.2303.10130

Epoch Ai. (2025). *Machine Learning Trends*. Retrieved from epoch.ai: https://epoch.ai/trends

Fazlioglu, M. (2023). *Privacy and Consumer Trust.* IAPP.org. Retrieved from https://iapp.org/resources/article/privacy-and-consumer-trust-summary/

Fernandez, J. (2023). *Generative AI Market Report 2023–2030.* Retrieved from https://iot-analytics.com/product/generative-ai-market-report-2023-2030/

Ferrandis, C. M. (2022). Open Sourcing AI: Intellectual Property at the Service of Platform Leadership. *JIPITEC 13 (3)*. Retrieved from https://www.jipitec.eu/issues/jipitec-13-3-2022/5557

Fletcher, R. (2024). *How many news websites block AI crawlers?* . Retrieved from Reuters Institute for the Study of Journalism, Oxford University.

Gans, J. S. (2024). *COPYRIGHT POLICY OPTIONS FOR GENERATIVE ARTIFICIAL INTELLIGENCE.* NATIONAL BUREAU OF ECONOMIC RESEARCH. Retrieved from http://www.nber.org/papers/w32106

Gillespie, N., Lockey, S., Curtis, C., & Pool, J. a. (2023). *Trust in Artificial Intelligence: A global study.* Brisbane, Australia; New York, United States: The University of Queensland; KPMG Australia. Retrieved from https://doi.org/10.14264/00d3c94

Global Market Insight. (2024). *AI Server Market Size - By Servers, By Hardware, By End User, Forecast 2024 - 2032.* Retrieved from https://www.gminsights.com/industry-analysis/ai-server-market

Global Mrket Insight. (2024). *AI Server Market Size - By Servers, By Hardware, By End User, Forecast 2024 - 2032.* Retrieved from https://www.gminsights.com/industry-analysis/ai-server-market

Goldman Sachs. (2025). *China's advances could boost AI's impact on global GDP.* Retrieved from https://www.goldmansachs.com/insights/articles/chinas-advances-could-boost-ai-impact-on-global-gdp

Google. (2024). *Gemma: Introducing new state-of-the-art open models.* Retrieved from https://blog.google.

Grand View Research. (2023). *GVR Report coverPersonal Development Market Size, Share & Trends Report By Instrument (Books, e-Platforms), By Focus Area (Mental Health, Physical Health), By Region, And Segment Forecasts, 2023 - 2030.* Retrieved from https://www.grandviewresearch.com/industry-analysis/personal-development-market

Guha, N., Lawrence, C., Gailmard, L. A., Rodolfa, K., Surani, F., Bommasani, R., . . . Percy Liang, e. a. (2023). AI Regulation Has Its Own Alignment Problem: The Technical and

Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing. *George Washington Law Review, Forthcoming*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4634443

Hacohen, U. Y. (2023). User-Generated Data Network Effects and Market Competition Dynamics. *Fordham Intellectual Property, Media & Entertainment Law Journal, 34*(1). Retrieved from https://ir.lawnet.fordham.edu/iplj/vol34/iss1/1

Hagey, K. a. (2024, February 8). am Altman Seeks Trillions of Dollars to Reshape Business of Chips and Ai. *The Wall Street Journal*. Retrieved from https://perma.cc/C6RE-WLPH

Hagiu, A., & Wright, J. (2024). *Artificial intelligence and competition policy.* International Journal of Industrial Organization. doi:https://doi.org/10.1016/j.ijindorg.2025.103134

Hagiu, A., & Wright, J. (2020). *When Data Creates Competitive Advantage.* Retrieved from https://hbr.org/2020/01/when-data-creates-competitive-advantage

Hagiu, A., & Wright, J. (2023 a). *To Get Better Customer Data,Build FeedbackLoops into your product.* Harvard Business Review . Retrieved from https://hbr.org/2023/07/to-get-better-customer-data-build-feedback-loops-into-your-products

Hagiu, A., & Wright, J. (2023 b). Data-enabled learning, network effects, and competitive advantage. *RAND Journal of Economics, 54*(4), 638-667. doi:http://dx.doi.org/10.1111/1756-2171.12453

Hagiu, A., & Wright, J. (2024). *Artificial intelligence and competition policy.* doi:https://doi.org/10.1016/j.ijindorg.2025.103134

Hanspach, P., & Galli, N. (2024). *Collusion by Pricing Algorithms in Competition Law and Economics.* Robert Schuman Centre for Advanced Studies Research Paper 2024-06. Retrieved from https://ssrn.com/abstract=4732527

Heim, L. (2021). *'Transformative AI and Compute - EA Forum.* Retrieved from https://forum.effectivealtruism.org/s/4yLbeJ33fYrwnfDev

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., . . . Damoc, B. (2022). Training Compute-Optimal Large Language Models. doi:https://doi.org/10.48550/arXiv.2203.15556

HuggingFace. (2022). *BigScience RAIL License v1.0.* Retrieved from https://huggingface.co/spaces/bigscience/license

IBM. (n.d.). Retrieved from https://www.ibm.com/it-it/topics/neural-networks

IBM. (2023). *IBM Unveils the Watsonx Platform to Power Next-Generation Foundation Models for Business.* Retrieved from https://newsroom.ibm.com/2023-05-09-IBM-Unveils-the-Watsonx-Platform-to-Power-Next-Generation-Foundation-Models-for-Business#:~:text=IBM%20watsonx.ai%3A%20A%20next%20generation%20enterprise%20studio%2C%20expected,models%20through%20an%20open%20and%20intui

Jared Kaplan, S. M. (2020). *Scaling Laws for Neural Language Models.* doi:https://doi.org/10.48550/arXiv.2001.08361

Jon Schmid, T. S. (2024). *Evaluating Natural Monopoly Conditions in the AI Foundation Model Market.* RAND Corporation, Santa Monica, California. Retrieved from https://www.rand.org/pubs/research_reports/RRA3415-1.html

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Benjamin Chess, Child, R., . . . Jeffrey Wu, a. D. (2020). *Scaling Laws for Neural Language Models.* Retrieved from https://doi.org/10.48550/arXiv.2001.08361

KARAÇAM, D. A. (2019). *Privacy and Monopoly Concerns in Data-Driven Transaction.* IOS Press. doi:doi:10.3233/FAIA190316

Klein, T. J. (2022). *How Important Are User-Generated Data for Search Result Quality? Experimental Evidence.* doi:https://dx.doi.org/10.2139/ssrn.4229292

Knight, W. (2024). *Amazon's Cloud Boss Likens Generative AI Hype to the Dotcom Bubble*. Retrieved from Wired.com: https://www.wired.com/story/amazons-cloud-boss-selipsky-generative-ai-hype/

Konrad, A. (2025, Gennaio 8). *Anthropic's Pending 60Billion Valuation will make all seven co-founders billionaires.* Retrieved from Forbes: https://www.forbes.com/sites/alexkonrad/2025/01/08/anthropic-60-billion-valuation-will-make-all-seven-cofounders-billionaires/

Korinek, A., & Vipra, J. (2024). *CONCENTRATING INTELLIGENCE: SCALING AND MARKET STRUCTURE IN ARTIFICIAL INTELLIGENCE.* NATIONAL BUREAU OF ECONOMIC RESEARCH, Cambridge,. Retrieved from http://www.nber.org/papers/w33139

Lewis Tunstall, ∗. E. (2023). *ZEPHYR: DIRECT DISTILLATION OF LM ALIGNMENT.* https://huggingface.co/HuggingFaceH4. Retrieved from https://arxiv.org/pdf/2310.16944

Licata, P. (2024). *Cloud, nel 2025 boom dell'Infrastructure as a service: +25% di investimenti.* Retrieved from https://www.corrierecomunicazioni.it/digital-economy/cloud/cloud-nel-2025-boom-dellinfrastructure-as-a-service-spesa-a-25/

LMArena. (n.d.). Retrieved from https://lmarena.ai/

MacroPolo.org. (2023). Retrieved from https://macropolo.org/interactive/digital-projects/the-global-ai-talent-tracker/

Maggiolino, M., & Zoboli, L. (2024). *Preserving Competition in Generative AI: Addressing the Merger Conundrum.* Concurrences – "AI & Competition Policy". Retrieved from https://ssrn.com/abstract=4823561

Manne, G. A., & Auer, D. (2024, febbraio). FROM DATA MYTHS TO DATA REALITY: WHAT GENERATIVE AI CAN TELL US ABOUT COMPETITION POLICY (AND VICE VERSA). *CPI Antitrust Chronicle February 2024*. Retrieved from Competitionpolicyinternational.com: https://laweconcenter.org/wp-content/uploads/2024/02/4-FROM-DATA-MYTHS-TO-DATA-REALITY-WHAT-GENERATIVE-AI-CAN-TELL-US-ABOUT-COMPETITION-POLICY-AND-VICE-VERSA-Geoffrey-A-Manne-Dirk-Auer-1.pdf

Maurice E. Stucke, A. P. (2016). *Big Data and Competition policy.* Oxford University Press. doi:http://dx.doi.org/10.1093/law:ocl/9780198788133.001.0001

Meta. (2024, Aprile 18). *Introducing Meta Llama 3: The most capable openly available LLM to date*. Retrieved from https://ai.meta.com/blog/meta-llama-3/

Mojan Javaheripi, S. R. (2023). *Phi-2: The surprising power of small language models.* Microsoft. Retrieved from https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

Morris, C. (2024). *Here are the companies OpenAI has made deals with to train ChatGPT.* Retrieved from https://www.fastcompany.com/91130785/companies-reddit-news-corp-deals-openai-train-chatgpt-partnerships

Nadella, S. (2024). *Mustafa Suleyman, DeepMind and Inflection Co-founder, joins Microsoft to lead Copilot.* Retrieved from https://blogs.microsoft.com/blog/2024/03/19/mustafa-suleyman-deepmind-and-inflection-co-founder-joins-microsoft-to-lead-copilot/

Nestor Maslej, L. F. (2024). *The AI Index 2024 Annual Report.* Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University,.

Nirmal Kumar Juluru, S. L. (2024). *Simplify Custom Generative AI Development with NVIDIA NeMo Microservices.* Retrieved from https://developer.nvidia.com/blog/simplify-custom-generative-ai-development-with-nvidia-nemo-microservices/

O'Brien, M. (2023). *ChatGPT Chief Says Artificial Intelligence Should Be Regulated by a US or Global Agency.* Alton Telegraph. Retrieved from https://perma.cc/ST6E-H3CL

OpenAI. (2023, Marzo 27). GPT-4 Technical Report. Retrieved from https://cdn.openai.com/papers/gpt-4.pdf

OpenAi. (2024). *How ChatGPT and our foundation models are developed*. Retrieved from https://help.openai.com/: https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed

Pablo Villalobos, A. H. (2024). *Will we run out of data? Limits of LLM scaling based on human-generated data.* Retrieved from https://arxiv.org/abs/2211.04325

Pablo Villalobos, J. S. (2022). *Will We Run out of Data? An Analysis of the Limits of Scaling Datasets in Machine Learning.* doi:http://arxiv.org/abs/2211.04325

Patel, A. (2024). *NVIDIA Releases Open Synthetic Data Generation Pipeline for Training Large Language Models. June.* Nvidia. Retrieved from https://perma.cc/PD97-ZJNT

Petit, N. a. (2021). Innovating Big Tech Firms and Competition Policy: Favoring Dynamic Over Static Competition. doi:https://dx.doi.org/10.2139/ssrn.3229180

Picchi, A. (2025, Genanio 28). *What is DeepSeek, and why is it causing Nvidia and other stocks to slump?* Retrieved from CBS News: https://www.cbsnews.com/news/what-is-deepseek-ai-china-stock-nvidia-nvda-asml/

Pierre Azoulay, J. L. (2024). *OLD MOATS FOR NEW MODELS: OPENNESS, CONTROL, AND COMPETITION IN GENERATIVE AI.* NATIONAL BUREAU OF ECONOMIC RESEARCH. doi:10.3386/w32474

Potts, J. (2012). The Innovation Commons. doi:http://dx.doi.org/10.2139/ssrn.2706856

Preta, A. (2024). L'economia dei dati e l'intelligenza artificiale tra politica economica, concorrenza e regolazione dei mercati. Retrieved from https://www.uer.it/jeanmonnetchair/digitrai/wp-content/uploads/2020/06/11-Preta-169-184.pdf

Qi, L. (2024). Stock Data Analysis of Competing Companies in Competitive Market: The case of NVIDIA Corporation. *94*, pp. 493-503. doi:http://dx.doi.org/10.54097/vnv0ec57

Regulatio EU. (2022). *on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).* Retrieved from https://eur-lex.europa.eu/eli/reg/2022/1925/oj/eng

Rohan Taori, I. G. (2023). *Alpaca: A Strong, Replicable Instruction-Following Model.* Stanford University. Retrieved from https://crfm.stanford.edu/2023/03/13/alpaca.html

Rohlfs, J. (1974). *A Theory of Interdependent Demand for a Communications Service.* Retrieved from https://www.jstor.org/stable/3003090

Ruihan Huang, A. C. (2023). *Methodology for Global AI Talent Tracker 2.0.* MacroPolo.org. Retrieved from https://macropolo.org/interactive/digital-projects/the-global-ai-talent-tracker/methodology-for-global-ai-talent-tracker-2/

Sachs, G. (2024). *GenAi: Too much spend, too little benefit.* Retrieved from https://www.goldmansachs.com/insights/top-of-mind/gen-ai-too-much-spend-too-little-benefit

Samuelson, P. (2023). *Generative AI meets copyright.* doi:https://doi.org/10.1126/science.adi0656

Sayash Kapoor, R. B. (2024). *On the Societal Impact of Open Foundation Models.* Retrieved from arXiv:2403.07918

Schaefer, M., & Sapi, G. (2023). *Complementarities in learning from data: Insights from general search.* doi:https://doi.org/10.1016/j.infoecopol.2023.101063

Seger, D. M. (2023). *'Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives.* Centre for the Governance of AI. doi:https://dx.doi.org/10.2139/ssrn.4596436

Servidio, G. (2025, Gennaio 30). *Il nuovo modello AI di Alibaba potrebbe battere DeepSeek: cos'è Qwen2.5-Max, come funziona e i limiti.* Retrieved from Geopop.it: https://www.geopop.it/il-nuovo-modello-ai-di-alibaba-potrebbe-battere-deepseek-cose-qwen2-5-max-come-funziona-e-i-limiti/

Sevilla, J., & Roland, E. (2024). *Training Compute of Frontier AI Models Grows by 4-5x per Year.* Epoch.ai. Retrieved marzo 29, 2025, from https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

Shaw, F. X. (2023). *Microsoft Build Brings AI Tools to the Forefront For Developers.* Official Microsoft Blog, 23 May. Retrieved from https://blogs.microsoft.com/blog/2023/05/23/microsoft-build-brings-ai-tools-to-the-forefront-for-developers/

Sheen S. Levine, D. J. (2023). *Come gli eetti di rete rendono l'IA più intelligente.* Harvard Business Review - Italia. Retrieved from https://www.hbritalia.it/marzo-2023/2023/03/20/news/come-gli-effetti-di-rete-rendono-lia-piu-intelligente-15467/

Sheen S. Levine, D. J. (2023). *Come gli effetti di rete rendono l'IA più intelligente.* Retrieved from Harvard Business Review Italia: https://www.hbritalia.it/marzo-2023/2023/03/20/news/come-gli-effetti-di-rete-rendono-lia-piu-intelligente-15467/

Singla, S. (2023). *Regulatory Costs and Market Power.* . Social Science Research Network.

Stanford University. (2024). *Artificial Intelligence Index Report.* Retrieved from https://aiindex.stanford.edu/report/

Stefano, A. D. (2025, febbraio 11). *Quanto vale OpenAI, il colosso dell'Intelligenza artificiale che Musk vuole acquisire con quasi 100 miliardi di dollari?* Retrieved from startupitalia.eu: https://startupitalia.eu/tech/quanto-vale-openai-offerta-musk/

Suleyman, M. (2023). *The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma.* First edition. New York: Crown. Retrieved from https://the-coming-wave.com/

Teece, D. J. (1986). *Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy.* Research Policy Vol.15, Issue 6. doi:https://doi.org/10.1016/0048-7333(86)90027-2

Thibault Schrepel, &. A. (2024). *Competition Between AI Foundation Models: Dynamics and Policy Recommendations.* MIT Connection Science. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4493900

Thibault Schrepel, J. P. (2024). Measuring the Openness of AI Foundation Models. *Sciences Po Digital, Governance and Sovereignty Chair*. Retrieved from https://ssrn.com/abstract=4827358

Thompson, A. D. (2024). *What's in GPT-5? A Comprehensive Analysis of Datasets Likely Used to Train GPT-5.* Retrieved from https://lifearchitect.substack.com/p/the-memo-special-edition-whats-in

Tianchuan Du, K.-h. C. (2024). *Improving Taxonomy-based Categorization with.* Retrieved from https://cs.stanford.edu/people/paulliu/files/bigdata-2021.pdf

Timo Kaufmann, P. W. (2024). *A Survey of Reinforcement Learning from Human Feedback.* doi:https://doi.org/10.48550/arXiv.2312.14925

Università di Palermo. (-). *Tesi di dottorato.* Retrieved from https://iris.unipa.it/bitstream/10447/94918/2/Tesi_Dottorato%20MLC.pdf

Usercentrics. (2023). *Intelligenza artificiale (IA), dati personali e consenso*. Retrieved from Usercentrics.com: https://usercentrics.com/it/knowledge-hub/intelligenza-artificiale-ia-e-consenso/

Vanberg, A. D. (2023). *Coordinating digital regulation in the UK: is the digital regulation cooperation forum (DRCF) up to the task?* nternational Review of Law, Computers & Technology,. doi: https://doi.org/10.1080/13600869.2023.2192566

Varian, H. (2018). *ARTIFICIAL INTELLIGENCE, ECONOMICS, AND INDUSTRIAL ORGANIZATION.* NATIONAL BUREAU OF ECONOMIC RESEARCH, Cambridge. Retrieved from http://www.nber.org/papers/w24839

Venkatesh Balavadhani Parthasarathy, A. Z. (2024). The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. Retrieved from https://arxiv.org/pdf/2408.13296

Wei-Lin Chiang, L. Z. (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference.* Retrieved from https://arxiv.org/pdf/2403.04132v1

Wiggers, K. (2023, febbraio 22). *OpenAI's Foundry Will Let Customers Buy Dedicated Compute to*. Retrieved from TechCrunch (blog): https://techcrunch.com/2023/02/21/openai-foundry-will-let-customers-buydedicated-capacity-to-run-its-si-models/

Wouter van Wengen, R. R. (2024). *EU AI Act's Opt-Out Trend May Limit Data Use for Training AI Models.* Amsterdam: Greenberg Traurig. Retrieved from https://www.gtlaw.com/en/insights/2024/7/eu-ai-acts-opt-out-trend-may-limit-data-use-for-training-ai-models

Xiao Bi, D. C. (2024). *DeepSeek LLM: Scaling Open-Source Language Models with Longtermism.* arXiv:2401.02954.

Xinyang Geng, A. G. (2023). *Koala: A Dialogue Model for Academic Research.* Retrieved from https://bair.berkeley.edu/blog/2023/04/03/koala/

Xu, F., Wang, X., Chen, W., & Xie, K. (2024). *The Economics of AI Foundation Models: Openness, Competition and Governance.* doi:https://dx.doi.org/10.2139/ssrn.4999355

Yang, J. Z. (2023). Mechanism of innovation and standardization driving company competitiveness in the digital economy. *Journal of Business Economics and Management,*, 24(1), 54–73. doi:https://doi.org/10.3846/jbem.2023.17192

Yiheng Liu, H. H. (2025). *Understanding LLMs: A comprehensive overview from training to inference,.* doi:https://doi.org/10.1016/j.neucom.2024.129190.

Yun, J. M. (2024). *The Folly of AI Regulation.* Antonin Scalia Law School. George Mason University Law & Economics Research Paper Series. doi:https://dx.doi.org/10.2139/ssrn.4968887

Sitografia

https://blog.osservatori.net/it_it/storia-intelligenza-artificiale

https://www.lum.it/machine-learning/#:~:text=Il%20machine%20learning%20nasce%20nel,intelligenza%20artificiale%20Arthur%20Lee%20Samuel

https://crfm.stanford.edu/ecosystem-graphs/index.html?mode=table

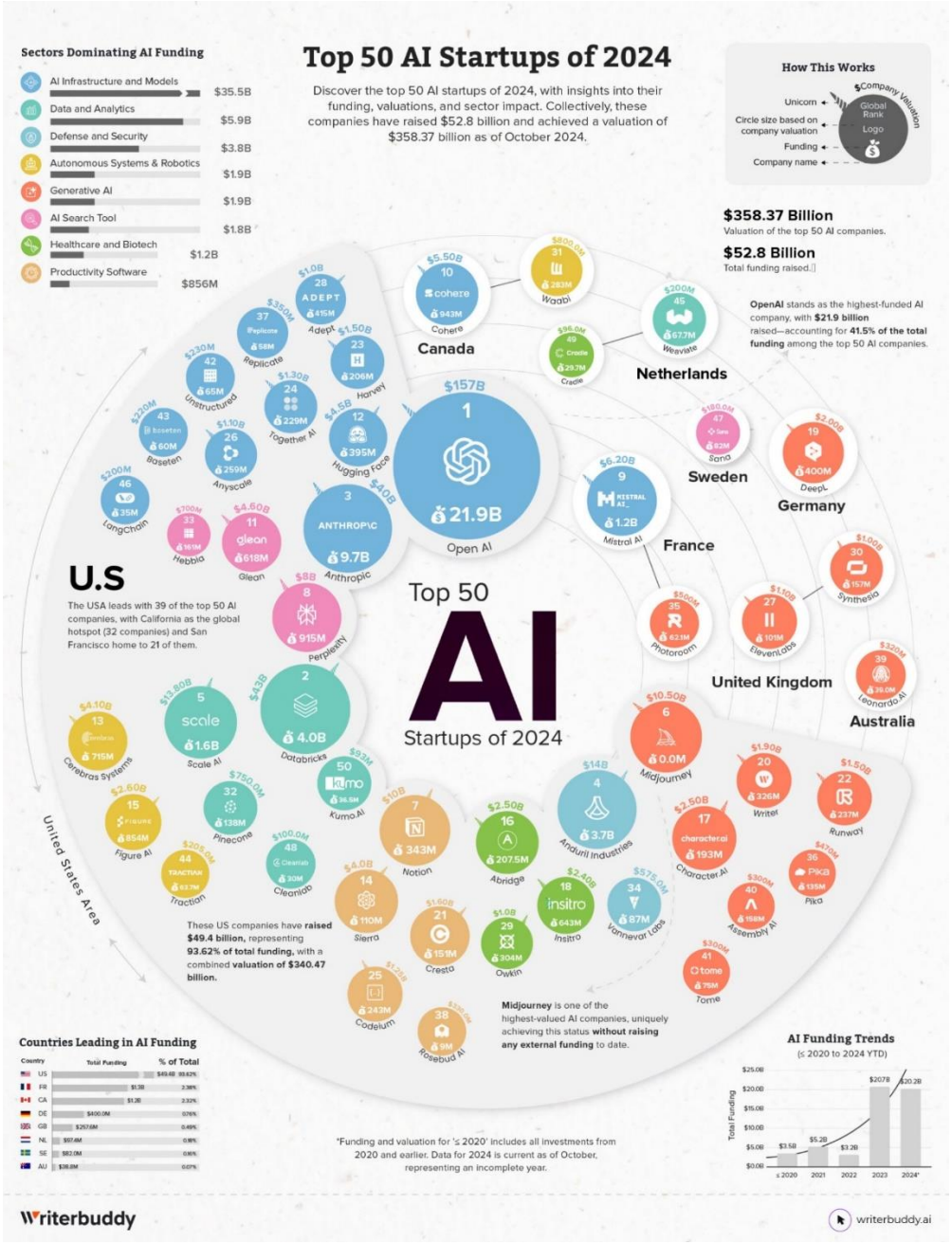
Gemma: Google introduces new state-of-the-art open models

https://press.un.org/en/2023/sgsm21880.doc.htm

Personal GPT

ANNEX

Annex A



L'infografica precedente, utile per avere una *overview* complessiva del mercato riassume tutti gli investimenti in *start-up* Ai del 2024 e ne indica la *market valuation* a ottobre 2024.

Source: Top 50 AI Companies of 2024: Funding, Valuation & Trends

Annex B

Company	Investments from large technology companies
OpenAI	Microsoft, Nvidia
Anthropic	Alphabet, Amazon, Salesforce, Zoom
Scale AI	Amazon, Meta, Nvidia, Intel, AMD
Perplexity	Nvidia, Amazon
Inflection AI	Microsoft, Nvidia
Hugging Face	Alphabet, AMD, Amazon, IBM, Intel, NVIDIA, Qualcomm, Salesforce
Mistral	Microsoft, Nvidia, Salesforce, Samsung, IBM
Baichuan	Alibaba, Tencent, Xiaomi
Moonshot	Alibaba, Tencent

Tabella 4: Investimenti nei leading AI labs delle Big Tech. Source: (Korinek & Vipra, 2024)

Annex C

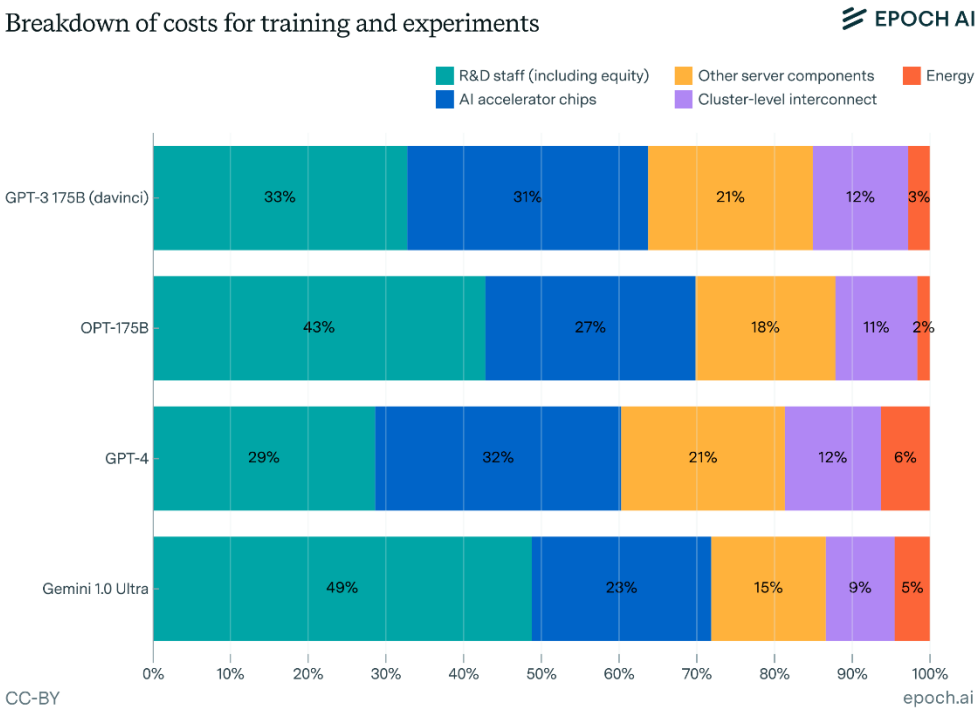


Figura 29: Breakdown dei costi di sviluppo per i modelli selezionati. Source: EpochAI

I costi dell'hardware sono ammortizzati in base al numero totale di ore di utilizzo dei *chip* impiegati negli esperimenti iniziali e nell'addestramento. I costi del personale di ricerca e sviluppo coprono l'intera durata del processo, dalle prime prove, alla pubblicazione.

Annex D

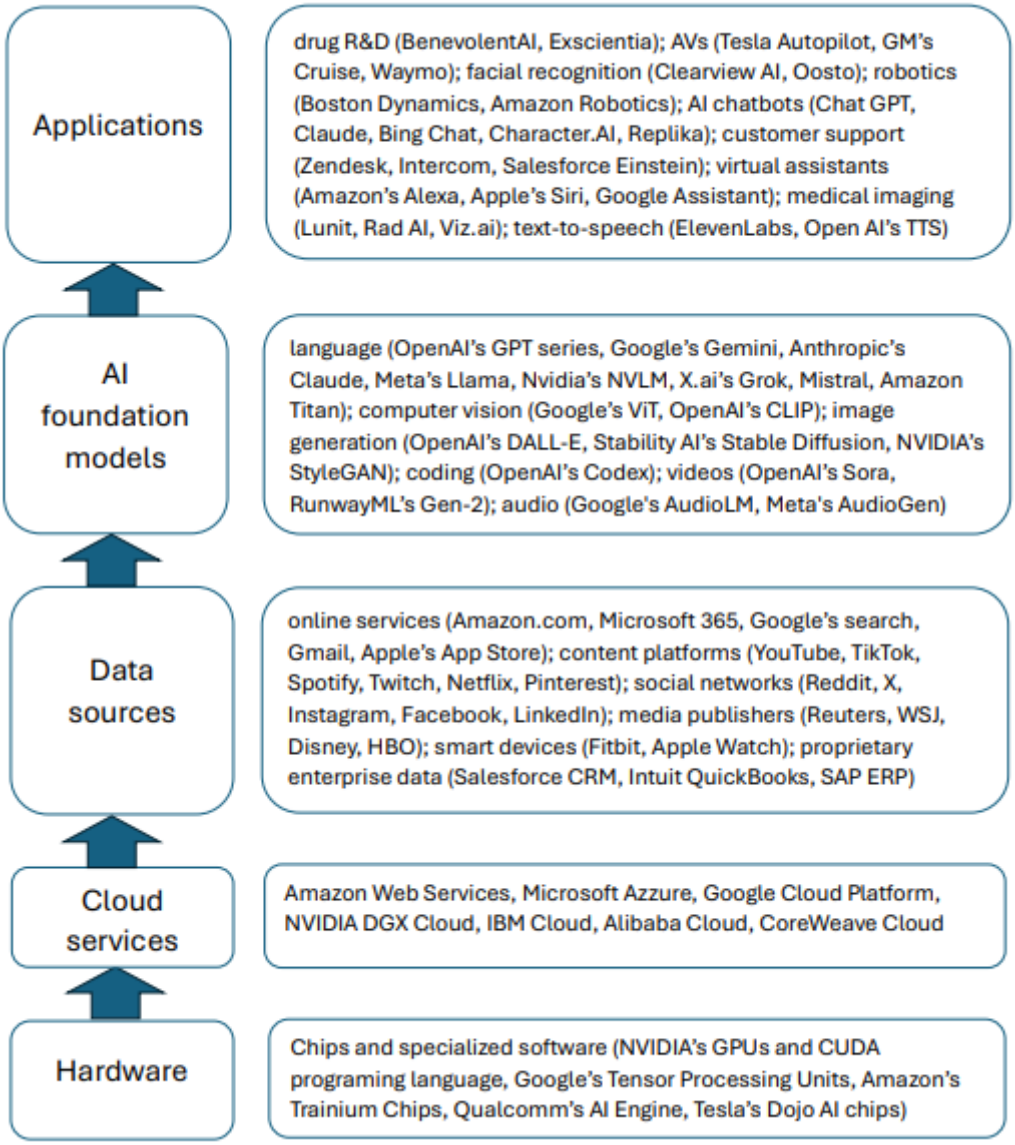


Figura 30: Esempi di aziende nella value chain del'AI