POLITECNICO DI TORINO

LAUREA MAGISTRALE in INGEGNERIA GESTIONALE

TESI DI LAUREA MAGISTRALE

Modelli di intelligenza artificiale generativa e assetti di mercato

Relatore

Prof. Carlo CAMBINI

Candidato

Giovanni TROTTA

APRILE 2025

Modelli di intelligenza artificiale generativa e assetti di mercato

Giovanni Trottola

Abstract

L'intelligenza artificiale generativa (IAG) sta ridefinendo profondamente i paradigmi economici globali, emergendo come una delle innovazioni tecnologiche più dirompenti e trasformative dell'ultimo decennio. Questo studio affronta le complesse dinamiche competitive all'interno del mercato dei Foundation Models, i modelli che costituiscono il cuore pulsante delle applicazioni di IAG. Attraverso un'approfondita analisi della catena del valore, integrando documentazioni tecniche dettagliate e valutazioni economiche specifiche, la ricerca mette in luce l'impatto significativo dei costi nei segmenti strategici del settore. Dall'indagine emerge una realtà contraddittoria: nonostante un numero elevato di modelli disponibili e la presenza diffusa di sviluppatori, pochi grandi player dominano in modo incontrastato le fasi cruciali della catena, generando importanti squilibri competitivi. Questi squilibri risultano ulteriormente accentuati dalle marcate disparità regionali, con Stati Uniti e Cina in una posizione nettamente privilegiata rispetto all'Europa. In tale contesto, le future scelte regolatorie, l'allocazione degli investimenti e la collaborazione tra istituzioni e imprese assumeranno un ruolo determinante, plasmando il futuro competitivo della IAG e garantendo uno sviluppo più equo e sostenibile dell'intero settore.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

1.1 L'impatto economico globale dell'intelligenza artificiale generativa

L'intelligenza artificiale generativa (IAG) è una branca dell'intelligenza artificiale che si distingue per la capacità di creare nuovi contenuti a partire da dati esistenti. L'IAG sta rapidamente ridefinendo l'economia globale, aumentando la produttività delle imprese che l'adottano, accelerando l'automazione e creando nuove opportunità di crescita economica. Nello studio *The Economic Potential of Generative AI: The Next Productivity Frontier, (2023)*[1] condotto da McKinsey, si stima che l'IAG possa aggiungere tra i 2,6 e i 4,4 trilioni di dollari all'economia globale ogni anno. Un valore economico paragonabile al PIL del Regno Unito del 2021. Questo impatto equivale a un aumento del 15-40% rispetto ai benefici già attribuiti alle tecnologie di IA esistenti, un incremento significativo che pone l'IAG al centro delle trasformazioni economiche globali. La tecnologia si distingue per la capacità di automatizzare processi complessi, ottimizzare operazioni aziendali e creare nuovi spazi di innovazione. I benefici principali si concentrano in quattro ambiti chiave: rapporti con i clienti, marketing e vendite, ingegneria del software e ricerca e sviluppo (R&D). Ad esempio, nel settore bancario, l'IAG potrebbe generare un valore aggiuntivo compreso tra 200 e 340 miliardi di dollari ogni anno, migliorando la gestione del rischio, aumentando l'efficienza operativa e potenziando la personalizzazione dei servizi al cliente. Parallelamente, nel settore retail e dei beni di consumo confezionati, l'IAG

[1] Il report utilizza il database O*Net del Bureau of Labor Statistics degli Stati Uniti, che suddivide circa 850 occupazioni in 2.100 attività lavorative dettagliate, analizzando ogni attività per valutare il livello di capacità necessario per svolgerla con successo, considerando 18 capacità rilevanti per l'automazione. Sono stati esaminati 63 casi d'uso dell'IAG in 16 funzioni aziendali.

ha il potenziale di incrementare il valore economico tra 400 e 660 miliardi di dollari grazie alla personalizzazione delle offerte, all'automazione delle operazioni e alla gestione ottimizzata della supply chain.

Questi numeri riflettono l'impatto diretto e indiretto della tecnologia, includendo sia il miglioramento dei processi esistenti che la creazione di nuovi mercati e opportunità di business. A livello globale, l'adozione dell'IAG potrebbe ridisegnare il panorama economico, consentendo alle organizzazioni di competere meglio in un mercato sempre più dinamico e di rispondere rapidamente alle mutevoli esigenze dei consumatori. Tuttavia, l'entità di questo impatto dipenderà dalla capacità delle imprese di investire strategicamente nella tecnologia e di integrare l'IAG nelle loro operazioni quotidiane. L'adozione diffusa dell'IAG non solo migliorerà la produttività, ma potrebbe anche stimolare la crescita economica sostenibile a lungo termine, rafforzando la resilienza delle organizzazioni e delle economie nazionali. Si stima infatti che l'impiego dell'IAG potrebbe migliorare l'efficacia delle attività di vendita, con un impatto stimato che varia tra il 3% e il 5% delle attuali spese globali per le vendite.

Sempre McKinsey, nel report *The state of AI in early 2024*[2], stima che nel 2024 il 6% delle organizzazioni utilizzava regolarmente soluzioni basate sull'IAG, quasi il doppio rispetto all'anno precedente, segnalando una chiara accelerazione nell'integrazione di questa tecnologia. In particolare, le organizzazioni che investono maggiormente nell'addestramento del personale e nell'integrazione della tecnologia nei loro processi aziendali tendono a ottenere un ritorno più elevato sugli investimenti. Questa tendenza si riflette anche nell'allocazione dei budget: molte organizzazioni dedicano oltre il 5% del loro budget digitale a soluzioni di IAG, con alcuni settori, come energia, tecnologia e finanza, che ne allocano fino al 20%.

1.2 Automazione e trasformazione della forza lavoro

Uno degli aspetti più rivoluzionari dell'IAG è la sua capacità di trasformare il lavoro umano, automatizzando attività che oggi occupano tra il 60% e il 70% del tempo lavorativo degli impiegati. Questo processo di automazione non solo riduce i costi operativi, ma permette di liberare risorse umane per concentrarsi su attività a maggiore valore aggiunto, come la strategia, l'innovazione e la gestione delle relazioni

[2]Il report basa l'analisi su una survey globale annuale condotta tra dirigenti e manager di aziende di diversi settori, con oltre 1.000 partecipanti provenienti da Paesi chiave e rappresentati sia da grandi multinazionali sia da piccole e medie imprese.

complesse. Rispetto alle precedenti tecnologie di automazione, l'IAG ha accelerato in modo significativo i tempi di adozione, con proiezioni che indicano che metà delle attività lavorative attuali potrebbe essere automatizzata tra il 2030 e il 2060, con un punto mediano stimato intorno al 2045. Questo progresso è reso possibile dalla capacità dell'IAG di comprendere e processare il linguaggio naturale, una competenza che estende il raggio d'azione della tecnologia a settori ad alta intensità cognitiva. La produttività del lavoro globale potrebbe crescere tra lo 0,1% e lo 0,6% annuo fino al 2040, con ulteriori incrementi possibili integrando l'IAG con altre tecnologie avanzate[3].

Tuttavia, la trasformazione non sarà priva di sfide. Una delle principali riguarda la necessità di investire in programmi di riqualificazione e aggiornamento delle competenze per i lavoratori, garantendo che possano adattarsi ai cambiamenti del mercato del lavoro. Inoltre, sarà fondamentale affrontare il rischio di disoccupazione tecnologica, specialmente nei settori più vulnerabili all'automazione, attraverso politiche di supporto che incentivino la transizione verso nuovi ruoli e attività. La collaborazione tra governi, imprese e istituzioni educative sarà cruciale per garantire una transizione equa e sostenibile, minimizzando gli impatti negativi sulla forza lavoro e massimizzando i benefici economici e sociali dell'automazione. La sfida principale sarà bilanciare l'adozione tecnologica con le esigenze sociali ed economiche, creando un ecosistema che favorisca l'innovazione senza sacrificare la coesione sociale.

1.3 I settori economici coinvolti

L'IAG ha trasformato profondamente numerosi settori industriali, portando con sé sia opportunità che sfide significative. Il settore del marketing e del design è tra i principali beneficiari di questa tecnologia, utilizzandola per creare contenuti visivi accattivanti e campagne pubblicitarie personalizzate, aumentando così l'efficacia e il coinvolgimento del pubblico. Questa sua capacità di generare immagini e testi complessi ha reso le strategie pubblicitarie più dinamiche e adattabili alle esigenze dei consumatori. Inoltre, permette di ridurre i costi di produzione dei contenuti e di accelerare il ciclo creativo, una risorsa preziosa per le aziende che cercano di mantenere un vantaggio competitivo in mercati saturi.

[3]The Economic Potential of Generative AI: The Next Productivity Frontier, (2023)

Il settore educativo sfrutta anche l'IAG per personalizzare i materiali didattici, migliorando l'accessibilità e la qualità dell'insegnamento. Gli algoritmi generativi possono creare contenuti educativi su misura per le esigenze di ciascun studente, favorendo l'apprendimento personalizzato e adattivo. Questa tecnologia, quindi, contribuisce a una democratizzazione dell'educazione, permettendo un accesso più equo a risorse di alta qualità. Tuttavia, l'implementazione di queste tecnologie richiede un attento monitoraggio per evitare che l'automazione sostituisca il ruolo cruciale degli educatori umani, sottolineato da studi sulla regolamentazione e l'uso responsabile.

In ambito scientifico e farmaceutico l'IAG ha dimostrato un potenziale straordinario nella scoperta e progettazione di nuovi materiali e molecole. Questo ha accelerato la ricerca e lo sviluppo, riducendo significativamente i tempi di sperimentazione. Ad esempio, i modelli generativi sono utilizzati per simulare e proporre strutture molecolari innovative, consentendo agli scienziati di testare virtualmente l'efficacia di nuovi composti. Tali applicazioni non solo migliorano la velocità della ricerca, ma possono portare a scoperte che altrimenti sarebbero state irraggiungibili con metodi tradizionali.

Inoltre, una ulteriori analisi condotta nel report *The state of AI in early 2024, McKinsey (2024)*, mostra che le modalità di adozione variano in base al settore. Se utilizzata per lavoro, il settore tecnologico risulta ovviamente al primo posto (39% degli intervistati), seguito dal settore dei servizi economici, legali e professionali (38%) e dal settore di media e telecomunicazioni (30%). Per uso personale, fuori dal lavoro, chi opera nel settore di energia e materiali ne fa utilizzo (26%), insieme a chi opera nelle industrie avanzate (26%). L'uso regolare per lavoro, nel complesso, non registra percentuali molto alte: in media tra tutti i settori il 12% degli intervistati ne fa uso. Si evidenzia in ogni caso un utilizzo "almeno una volta" con percentuali consistenti: infatti, il 51% degli intervistati nel campo della sanità ha fornito questa risposta, seguito dal 43% degli intervistati nei servizi finanziari, dal 38% nel commercio al dettaglio e dal 34% nel settore delle industrie avanzate.

In generale, mentre i settori più tecnologici come telecomunicazioni, tecnologia e servizi finanziari guidano l'adozione regolare dell'IAG, quelli tradizionali stanno ancora esplorando le sue possibilità, con alti livelli di sperimentazione ma basse percentuali di utilizzo consolidato. Questa tendenza suggerisce che, nonostante il forte interesse, molte industrie sono ancora nelle prime fasi di integrazione dell'IAG

nei loro processi operativi.

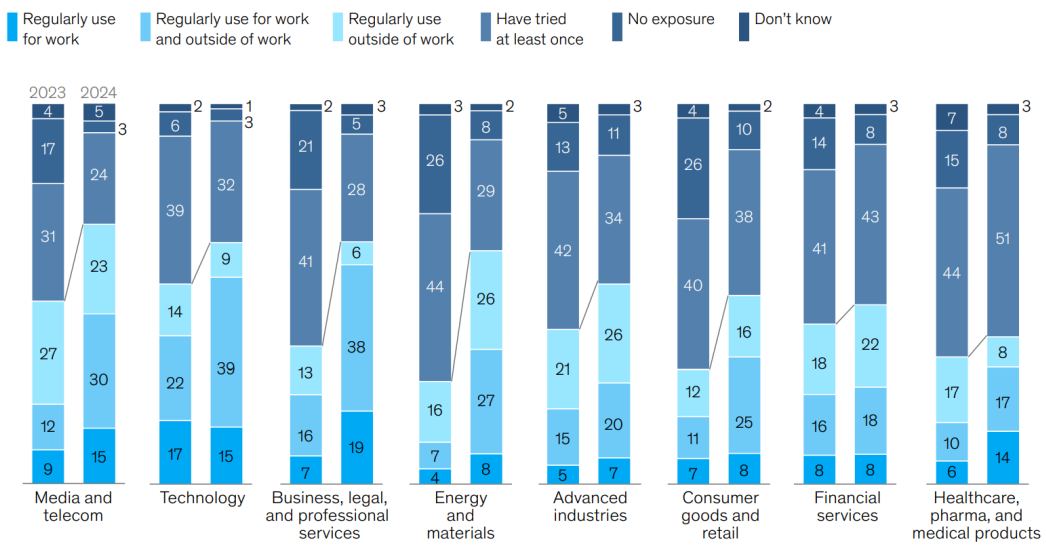

Figura 1.1: Esperienza personale con strumenti di IAG, per settore (% di intervistati, 2023-24).[4]

Analizzando l'area geografica degli intervistati, risulta che nel 2024 la maggior parte dei soggetti che facevano uso di IAG per esigenze lavorative erano pari al 19% in Cina, 15% nei mercati emergenti, 12% in Nord America, 11% in Europa e 3% nell'area asiatica-del Pacifico. Un utilizzo regolare combinato a lavoro e fuori dal lavoro era più consistente per l'area asiatica-pacifica (31%), seguita dalla Cina, dai mercati emergenti e dal Nord America (27%), e, non molto distante, dall'Europa (21%).

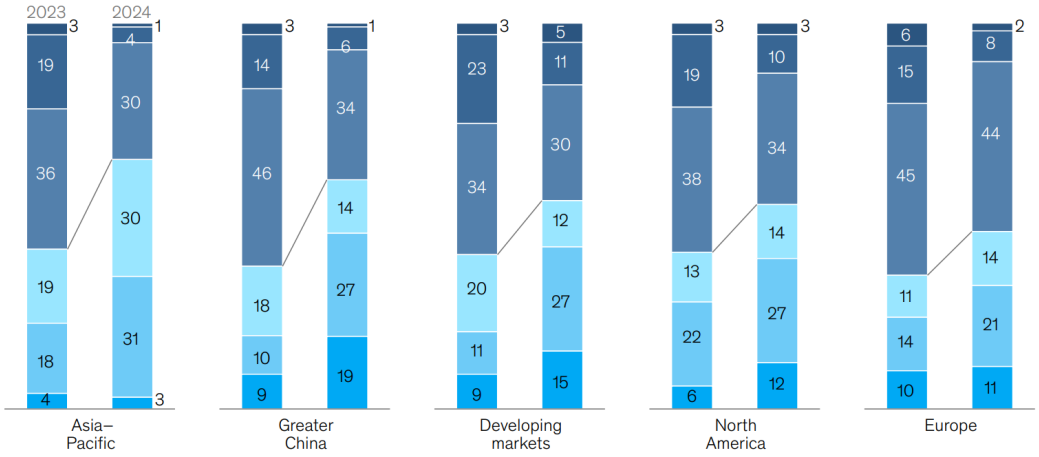

Figura 1.2: Esperienza personale con strumenti di IAG, per area geografica (% di intervistati, 2023-24).[3]

[4]The state of AI in early 2024: Gen AI adoption spikes and starts to generate value, McKinsey (2024).

L'IAG rappresenta quindi una delle innovazioni più trasformative dell'era moderna, con una crescente influenza in numerosi settori economici e sociali. L'analisi dettagliata dei modelli di IAG, dalle loro caratteristiche tecniche alle implicazioni economiche, mostra come questa tecnologia sia diventata una risorsa essenziale per il progresso e la competitività. Tuttavia, nonostante le loro potenzialità, permangono sfide significative legate ai costi di sviluppo, alla competizione nei diversi settori e alla regolamentazione.

Capitolo 2

Una panoramica sui modelli di IAG

2.1 Foundation Models: Large e Small Language Models

L'intelligenza artificiale generativa (IAG) rappresenta un ambito in continua e rapida evoluzione che ha trasformato in maniera profonda il modo in cui le macchine interagiscono con il linguaggio, le immagini, l'audio e altri dati complessi. Questa trasformazione ha aperto nuove prospettive nella ricerca e nell'innovazione tecnologica, oltre a ridefinire significativamente le dinamiche economiche, sociali e competitive su scala globale. Le capacità dei modelli IAG di generare contenuti nuovi e originali, che spaziano dai testi alle immagini, sono alimentate dai *Foundation Models (FMs)*, ovvero modelli che fungono da "fondamenta" per applicazioni specializzate. Si tratta di sistemi di intelligenza artificiale addestrati su enormi quantità di dati, tra cui testo, immagini, video e altro. Le loro peculiarità li distinguono nettamente dai modelli tradizionali e ne spiegano il successo e l'adozione su larga scala. In particolare, le principali caratteristiche sono:

1. *Linguaggio naturale*: i FMs impiegano il linguaggio naturale come modalità primaria di interazione, consentendo la realizzazione di interfacce utente semplificate, accessibili anche ad utenti non specializzati, e favorendo in tal modo la loro ampia diffusione e adozione;

2. *Adattabilità*: indica la capacità dei FMs di essere adattati per compiti specifici attraverso tecniche come il fine-tuning. Questi modelli, una volta pre-

addestrati su dati generali, possono essere adattati a una vasta gamma di applicazioni[5], riducendo la necessità di sviluppare modelli da zero per ogni nuovo compito. Questa flessibilità rende i FMs strumenti versatili in vari domini;

3. *Scalabilità*: si riferisce alla capacità dei FMs di migliorare le proprie prestazioni all'aumentare delle risorse computazionali e della quantità di dati di addestramento. L'incremento della potenza computazionale e della dimensione del dataset di addestramento tende a migliorare progressivamente le capacità del modello, seguendo un andamento prevedibile[6]. Un fenomeno strettamente legato alla scalabilità, ma distinto, è rappresentato dalle *capacità emergenti*, ovvero abilità che si manifestano spontaneamente quando un modello supera una certa soglia di complessità. Queste capacità, che non erano state programmate o previste durante l'addestramento iniziale, si manifestano solo in modelli sufficientemente grandi[7]. Ad esempio, la capacità di risolvere problemi logici o di comprendere battute emerge solo dopo che il modello raggiunge un certo livello di parametri. Pur dipendendo dalla scalabilità per manifestarsi, le capacità emergenti non seguono un andamento prevedibile, ma rappresentano un sottoprodotto inatteso del processo di scaling.

I FMs possono basarsi su diverse architetture approfondite in appendice A. Tra queste, quelle che hanno visto la maggior diffusione sono i Transformers su cui si basano i modelli per l'elaborazione del linguaggio naturale, noti come Large Language Models (LLM) e Small Language Models (SLM).

Gli LLM sono modelli progettati per elaborare e generare linguaggio naturale che sfruttano le loro immense dimensioni per identificare pattern e sfumature contestuali estremamente complessi. Grazie alla loro capacità di "apprendimento non supervisionato", ossia la capacità di imparare autonomamente dai dati senza richiedere etichettature esplicite, sono in genere pre-addestrati su enormi set di dati non strutturati, come pagine web, libri e articoli, per poi venire ottimizzati per compiti specifici tramite le tecniche di fine-tuning, rendendoli adattabili a varie applicazioni.

Un elemento fondamentale di questi modelli è rappresentato dai parametri, conosciuti anche come "pesi", che determinano come il modello processa i dati e genera le

[5]Bommasani, R. et al. (2021) On the Opportunities and Risks of Foundation Models. https://arxiv.org/abs/2108.07258.

[6]Sastry, G. et al. (2024) Computing power and the governance of artificial intelligence. https://arxiv.org/abs/2402.08797.

[7]Wei, J. et al. (2022) Emergent abilities of large language models. https://arxiv.org/abs/2206.07682.

risposte. I parametri sono numeri che il modello "impara" durante l'addestramento e che gli permettono di riconoscere relazioni nei dati di input. Maggiore è il numero di parametri, più il modello è in grado di catturare dettagli complessi e di conseguenza avere prestazioni migliori. Il numero di parametri è importante perché consente al modello di generalizzare meglio, ovvero di affrontare compiti diversi senza dover essere specificamente programmato per ciascuno di essi.

Recentemente, gli LLM hanno esteso le loro capacità oltre il linguaggio naturale, evolvendosi in modelli multimodali. Questi sono i Multimodal Large Language Models (MLLM), definiti come modelli basati su LLM con la capacità di ricevere, elaborare e produrre output su informazioni multimodali. Questi modelli combinano input di diverse tipologie, come testo, immagini, audio e video, superando i limiti intrinseci dei modelli tradizionali unicamente testuali o visivi. La timeline dei MLLM (figura 2.1) mostra come dopo il rilascio di DALL-E di OpenAI nel 2021, un modello capace di generare immagini tramite input testuale, lo sviluppo degli MLLM abbiano subito un'accelerazione significativa a partire dal 2022.

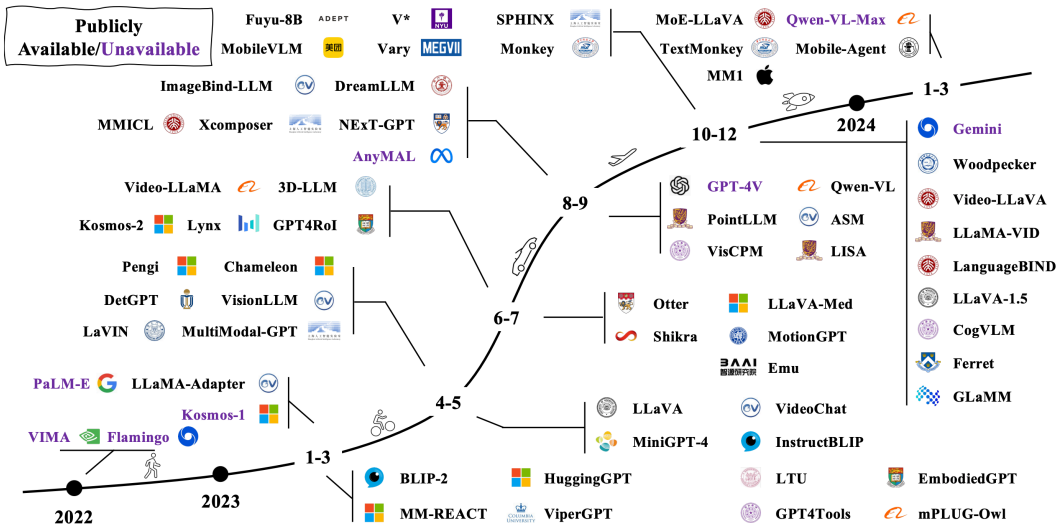

Figura 2.1: Timeline Multimodal Large Language Models.[8]

Questa crescita esponenziale evidenzia come i MLLM siano diventati sempre più sofisticati, adattandosi a una vasta gamma di applicazioni pratiche permettendo di rivoluzionare completamente numerosi settori. Nonostante i LLM e i MLLM rappresentino una pietra miliare nel campo dell'intelligenza artificiale, la loro grandezza e complessità pongono sfide significative. L'addestramento e l'utilizzo di modelli così imponenti richiedono ingenti risorse computazionali, elevate quantità di energia e

[8]Yin, S. et al. (2024) 'A survey on multimodal large language models,' National Science Review, 11(12). https://arxiv.org/abs/2206.07682.

infrastrutture avanzate che non sono sempre accessibili a tutte le organizzazioni o applicabili in ogni scenario. È in questo contesto che emergono gli Small Language Models (SLM) come una soluzione complementare ai modelli citati. La differenza tra i modelli di grandi e di piccole dimensioni risiede nella numerosità dei parametri. Ad esempio, uno degli LLM più avanzati è GPT-4o di OpenAI, un modello progettato per compiti generici e complessi che si adatta a molteplici contesti. Il numero stimato di parametri di questo modello risulta essere dell'ordine dei triliardi, un numero estremamente superiore se confrontato con un SLM come DistilBERT, il cui numero di parametri si aggira attorno ai 66 milioni. Questa riduzione delle dimensioni comporta vantaggi in termini di efficienza computazionale e accessibilità, rendendo gli SLM particolarmente adatti per applicazioni su dispositivi con risorse limitate, come smartphone o dispositivi IoT.

Esattamente come gli LLM, anche gli SLM fungono da base per modelli multimodali, ovvero i Multimodal Small Language Models (MSLM). Questi modelli stanno rivoluzionando il mondo dell'intelligenza artificiale, in quanto la loro scala più piccola non è una limitazione perché, sebbene anche gli LLM possano essere adattati a compiti specifici attraverso tecniche di fine-tuning, negli SLM questo processo, grazie alla loro dimensione più contenuta, richiede meno dati e meno risorse computazionali rendendoli più adatti per attività specialistiche. In questi contesti, le prestazioni dei modelli di piccole dimensioni sono paragonabili a quelle dei modelli di grandi dimensioni rendendo gli SLM e MSLM particolarmente vantaggiosi per organizzazioni o applicazioni che operano con risorse limitate, in quanto, a parità di performance, presentano un costo di implementazione e manutenzione significativamente inferiore. Di seguito una tabella comparativa tra le principali caratteristiche dei modelli.

Tabella 2.1: Confronto tra LLM e SLM.

Caratteristica	**LLM**	**SLM**
Elevato numero di parametri	✗	✓
Elevata efficienza computazionale	✗	✓
Adatto a dispositivi a risorse limitate	✗	✓
Alte prestazioni su compiti generalisti	✓	✗
Fine-tuning per compiti specifici	✓	✓
Alta velocità di inferenza	✗	✓
Ideale per applicazioni complesse	✓	✗
Ideale per applicazioni mirate	✗	✓

Tipicamente, l'utilizzo di questi modelli avviene congiuntamente in modo da sfruttare i punti di forza di entrambi. Questo approccio consente una sinergia collaborativa, in cui ogni modello contribuisce con le sue capacità uniche. SLM e MSLM possono fornire approfondimenti specializzati o eseguire attività mirate, mentre gli LLM e MLLM possono aggiungere ampiezza e profondità di conoscenza.

Dopo aver delineato le fondamenta degli LLM e degli SLM, è importante analizzare il contesto in cui questi modelli vengono sviluppati e distribuiti, distinguendo tra modelli open source e closed source. Questa distinzione rappresenta un aspetto cruciale, poiché il grado di accessibilità e trasparenza di un modello influenza non solo il modo in cui viene utilizzato, ma anche il suo impatto sull'innovazione, sulla concorrenza e sulla democratizzazione dell'intelligenza artificiale. Comprendere le caratteristiche e le implicazioni di queste due categorie di modelli è essenziale per inquadrare il ruolo dei modelli generali e specializzati, in particolare perché molti dei modelli specializzati (strumenti fondamentali in settori come sanità e finanza) tendono a essere open source.

2.2 Modelli closed source e open source

In base alle modalità di distribuzione, i FMs si suddividono in modelli open source e in modelli closed source, noti anche come "modelli proprietari". I modelli closed vengono sviluppati internamente alle aziende, le quali mantengono il controllo totale

sui dati di addestramento, l'architettura e i parametri del modello. Le modalità di distribuzione e di monetizzazione variano in base alla natura del modello. Per i modelli closed, questi possono essere:

1. Integrati per migliorare le prestazioni di prodotti già esistenti ed aumentarne il valore. Ad esempio, Microsoft ha stretto partnership strategiche con OpenAI, consentendo l'integrazione dei suoi modelli proprietari all'interno dei suoi servizi. Questo tipo di accordo permette alle aziende di offrire soluzioni di IA avanzate attraverso servizi cloud dedicati, facilitando l'accesso alle aziende clienti che desiderano integrare queste tecnologie nei propri sistemi senza dover sviluppare internamente modelli di IA;

2. Impiegati per sviluppare nuovi prodotto o servizi, spesso venendo monetizzati tramite abbonamenti o modelli freemium. Un esempio significativo di questa strategia è rappresentato da ChatGPT Plus, una versione premium dei modelli di OpenAI. Pur essendo un modello closed source, ChatGPT Plus è accessibile al pubblico tramite un modello di abbonamento mensile, che garantisce agli utenti vantaggi esclusivi, come tempi di risposta più rapidi, accesso prioritario anche nei momenti di alta richiesta e funzionalità avanzate rispetto alla versione gratuita. Questa modalità di monetizzazione è particolarmente efficace per massimizzare l'accesso diffuso a tecnologie avanzate, senza rivelare o distribuire il modello sottostante;

3. Distribuiti tramite API, ovvero una modalità di accesso in cui i modelli vengono forniti a terzi per essere integrati nei propri servizi. Ad esempio, è possibile integrare i modelli di OpenAI e quelli di altri sviluppatori sfruttando le API a disposizione nella piattaforma Azure di Microsoft o di altri CSP (Cloud Service Providers). La monetizzazione avviene con un modello a consumo, in cui gli utenti pagano in base al numero di richieste API o al volume di dati elaborati. Ad esempio, le tariffe variano in base al tipo di analisi richiesta e al numero di caratteri processati. Questa modalità permette agli sviluppatori di offreire una soluzione scalabile e flessibile che può essere integrata in applicazioni aziendali, piattaforme di customer service e molto altro.

I modelli open source, invece, sono sviluppati con l'intento di essere condivisi liberamente, permettendo a chiunque di accedere al codice sorgente, all'architettura del modello, ai pesi e talvolta ai dati di addestramento utilizzati. Il modello di busi-

ness dei modelli open source spesso si basa su un ecosistema di supporto, che può includere servizi di consulenza, infrastrutture di calcolo, e supporto tecnico per le aziende che utilizzano questi modelli. Per esempio, piattaforme come Hugging Face non solo forniscono accesso gratuito ai modelli ma anche servizi di supporto e infrastrutture cloud per il loro utilizzo e addestramento, generando introiti attraverso servizi premium. Questo modello ibrido, che combina l'accessibilità del codice aperto con servizi aggiuntivi a pagamento, permette alle aziende di generare profitto senza compromettere l'apertura del modello stesso.

I due modelli presentano quindi sostanziali differenze nelle modalità di sviluppo e di distribuzione che hanno impatti non solo dal punto di vista tecnologico dei modelli, ma hanno anche fondamentali conseguenze su diverse dimensioni quali la sicurezza dei modelli, il potenziale di innovazione e la trasparenza.

2.2.1 Sicurezza

Il controllo centralizzato degli sviluppatori nei modelli closed source permette di applicare delle restrizioni che riducono il rischio di manomissione dei dati e minimizza la possibilità di utilizzo improprio del modello. Al contrario, nei modelli open source, una volta resi pubblici i pesi del modello, gli sviluppatori perdono quasi completamente il controllo sull'uso del modello a valle. Seppur quindi ci siano dei benefici nel permettere di visionare e modificare il modello, l'applicazione di limitazioni per evitare che gli utenti sfruttino il modello o i dati a scopi malevoli risulta difficilmente applicabile. Inoltre, il rilascio dei pesi costituisce un'azione irreversibile in quanto, nonostante lo sviluppatore può interrompere l'accesso al modello, non può né revocare le copie dei pesi ormai create e né tantomeno impedirne la ridistribuzione peer-to-peer.

Tuttavia, il rilascio dei pesi permette di distribuire i modelli open source su hardware locale e di eseguire l'inferenza. Ciò implica che gli utenti non sono costretti a condividere i propri dati con gli sviluppatori, il che è particolarmente importante quando si utilizza il modello in settori in cui si trattano dati sensibili. D'altro canto, questa possibilità riduce il controllo e il monitoraggio degli usi del modello da parte degli utenti a valle. Nonostante alcuni sviluppatori di modelli closed forniscano meccanismi che consentono agli utenti di rifiutare esplicitamente la raccolta dei dati, le procedure di archiviazione, condivisione e utilizzo dei dati degli sviluppatori non sono sempre trasparenti.

2.2.2 Innovazione

Una ulteriore caratteristica dei modelli open source è l'elevato livello di personalizzazione nelle applicazioni a valle tramite diverse tecniche di fine-tuning applicabili da parte degli utenti. Sebbene anche alcuni modelli closed source permettano di applicare metodi di adattamento del modello, questi tendono ad essere più restrittivi e costosi. La personalizzazione, l'acceso più ampio e l'inferenza locale permette agli sviluppatori di addestrare i modelli su dati proprietari favorendo una personalizzazione più aggressiva, che non ha eguali nei modelli closed source a causa delle limitazioni imposte dallo sviluppatore, e permette quindi di supportare l'innovazione in una vasta gamma di applicazioni.

Le piattaforme open source consentono quindi di promuovere un ambiente collaborativo dove sviluppatori e ricercatori possono contribuire al miglioramento continuo della tecnologia. Tuttavia, nonostante il grado di personalizzazione applicabile dagli utenti a valle sia più elevato rispetto ai modelli closed, questi non possono avere acceso ai feedback degli utilizzatori del modello, che rappresentano una fonte di dati estremamente importante per il miglioramento continuo dello stesso.

2.2.3 Trasparenza

Per analizzare le differenze di trasparenza tra le due categorie di modelli è possibile far affidamento al *Foundation Model Transparency Index, Bommasani, R. et al. (2024)*. Questo indicatore permette di concettualizzare il grado di trasparenza dei FMs lungo i tre domini della catena di approvvigionamento, ovvero:

1. le risorse coinvolte nello sviluppo a monte del modello;
2. il modello stesso;
3. l'uso del modello a valle.

Questi domini sono successivamente scomposti in 28 sottodomini a cui vengono assegnate 100 variabili binarie. Di seguito (figura 2.2) i risultati ottenuti dalla ricerca su 14 aziende di cui 6 open developers (Adept, BigCode/Hugging Face/ServiceNow, Meta, Microsoft, Mistral, Stability AI) e 8 closed developers (AI21 Labs, Aleph Alpha, Amazon, Anthropic, Google, IBM, OpenAI, Writer).

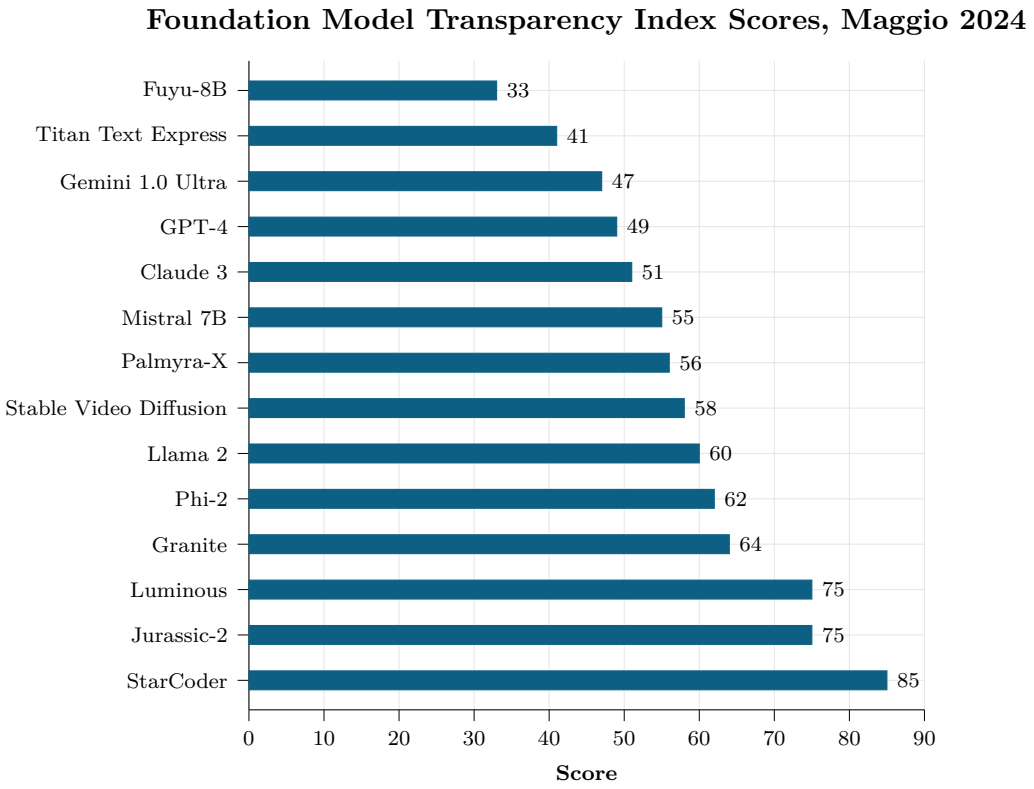

Figura 2.2: Foundation Model Transparency Index Scores, Maggio 2024.[9]

I modelli open source in genere ottengono punteggi più alti in termini di trasparenza rispetto a quelli closed source, con una differenza mediana di 5,5 punti tra le due categorie. Tuttavia, la condivisione pubblica dei pesi di un modello, pur essendo correlata a una trasparenza complessiva maggiore, non implica necessariamente una chiarezza superiore su aspetti specifici quali i dati, le risorse computazionali impiegate e le modalità di utilizzo del modello a valle. Infatti, la differenza dei punteggi è principalmente attribuibile alla trasparenza a monte dei modelli, in quanto i modelli open ottengono un punteggio pari o superiori in 18 sui 23 sottodomini. In particolare, i sottodomini con le differenze più ampie sono: lavoro sui dati, dati e utilizzo del modello. Questo risultato è coerente con le differenze intrinseche nei modelli.

Differente è il caso della trasparenza a valle, in quanto, nonostante per gli sviluppatori open source sia più difficile monitorare l'utilizzo dei modelli, questi ottengo un punteggio che è pari a quello dei modelli closed che possono attuare pratiche di monitoraggio e controllo nettamente superiori ai primi.

I modelli closed source ottengono risultati migliori rispetto a quelli open source in ambiti specifici legati alle politiche di utilizzo del modello. In particolare,

[9]Bommasani, R. et al. (2024) The Foundation Model Transparency Index V1.1: Maggio 2024. https://arxiv.org/abs/2407.12929.

forniscono maggiori dettagli sull'applicazione delle proprie politiche in relazione al comportamento degli utenti. Anche nelle aree di gestione dei rischi e delle mitigazioni, i modelli closed source registrano punteggi più alti, grazie a una maggiore propensione nel descrivere e dimostrare le misure di mitigazione dei rischi adottate.

Le dinamiche di apertura e chiusura nello sviluppo e nella distribuzione dei modelli influiscono significativamente sulla loro applicazione e accessibilità. Come evidenziato, la natura open source favorisce la collaborazione, l'innovazione e l'adozione diffusa, rendendo questi modelli fondamentali in settori diversificati, soprattutto quando si tratta di applicazioni specializzate. Al contrario, i modelli closed source offrono maggiore controllo e protezione, caratteristiche particolarmente utili per contesti in cui la sicurezza e la gestione dei dati sono prioritarie. Questa distinzione getta le basi per comprendere come i FMs possano essere adattati a scopi specifici attraverso tecniche di specializzazione. Infatti, molte delle applicazioni più avanzate e mirate nascono proprio dalla capacità dei modelli open source di supportare l'innovazione in contesti settoriali.

2.3 Modelli generali e specializzati

Tramite questa sezione si vogliono approfondire le principali caratteristiche e utilizzi dei modelli generali e specializzati. Partendo da una definizione dei modelli generali, questi sono sistemi con una vasta gamma di possibili applicazioni, sia intenzionali che non previste dai loro sviluppatori. Possono essere utilizzati per diverse applicazioni in diversi settori senza ricevere ottimizzazioni specifiche. Di fatto, questi modelli sono quelli precedentemente definiti come FMs in quanto si caratterizzano per il loro utilizzo diffuso come modelli pre-addestrati per altri sistemi più specializzati. I modelli generali sono prevalentemente LLM e MLLM e vengono sempre più utilizzati per la costruzione di applicazioni avanzate nella medicina, finanza, chimica, istruzione, programmazione e in molti altri settori. In tabella 2.2 alcuni dei principali modelli generali nel mercato:

Tabella 2.2: Alcuni modelli generali del mercato.

Modello	**Sviluppatore**	**Parametri**
GPT-4o	OpenAI	-
Claude 3.7 Sonnet	Anthropic	-
Gemini 2.0 Pro	Google DeepMind	-
Copilot	Microsoft	-
Llama 3.3	Meta AI	70 miliardi
Mistral Large 2	Mistral	123 miliardi
Falcon Mamba 7B	Falcon	7,27 miliardi
DeepSeek-V3	DeepSeek	671 miliardi

I modelli specializzati, invece, vengono addestrati con set di dati più piccoli in modo da adattarli ad un determinato contesto o settore. Ad esempio, un dataset contenente documenti legali potrebbe essere utilizzato per migliorare la capacità di un modello di fornire consulenze legali o generare documenti legali. Questi modelli vengono spesso implementati dalle aziende operanti in settori in cui si trattano dati sensibili così da avere un maggiore controllo sui dati ed evitare di fare affidamento a modelli generali forniti da terzi e addestrati su dati sconosciuti. Inoltre, l'adozione di modelli specializzati, essendo addestrati su set di dati più piccoli e specializzati offrono prestazioni migliori rispetto ai generali in termini di:

1. Precisione dell'output;
2. Pertinenza al contesto dei contenuti generati;
3. Maggior efficienza richiedendo meno risorse computazionali, tempi di elaborazione più rapidi e un minore consumo energetico.

Grazie alle loro caratteristiche, i modelli specializzati vengono implementanti in diversi settori, che spaziano dal settore finanziario al sanitario, di seguito vengono illustrati i settori col maggior tasso di adozione di modelli specializzati e alcune delle loro possibili applicazioni.

1. *Sanitario*: i modelli di IAG specializzati stanno rivoluzionando il settore sanitario grazie alla capacità di analizzare enormi volumi di dati clinici, immagini mediche e informazioni storiche sui pazienti per migliorare la diagnosi e la gestione delle cure. L'uso di modelli avanzati permette di individuare pattern

complessi che facilitano la diagnosi precoce di malattie come il cancro, le patologie cardiovascolari e i disturbi neurologici. L'analisi di dati genetici e clinici consente di sviluppare trattamenti personalizzati, migliorando l'efficacia delle cure e riducendo gli effetti collaterali. I modelli sono in grado di esaminare dati provenienti da cartelle cliniche elettroniche per supportare i medici nelle decisioni cliniche, aumentando la velocità e la precisione delle diagnosi. Inoltre, l'IA viene utilizzata per analizzare enormi dataset biologici e chimici, accelerando lo sviluppo di nuovi farmaci e riducendo i tempi di sperimentazione. L'analisi predittiva basata su dati epidemiologici consente di identificare trend di salute pubblica e potenziali focolai, migliorando la risposta a livello di sistema sanitario. Queste applicazioni trasformano la pratica medica, passando da un approccio reattivo a uno preventivo e personalizzato, creando un ecosistema più efficiente;

2. *Istruzione*: i sistemi di tutoraggio intelligente rappresentano una delle applicazioni più diffuse dell'IA nel settore dell'istruzione. Sono in grado di raccogliere dati dettagliati a livello individuale per valutare i progressi e offrire feedback personalizzato. Questi sistemi facilitano l'apprendimento degli studenti tramite dei processi che ne analizzano il comportamento permettendo di rilevare e correggere delle incomprensioni che si verificano durante il processo educativo. Ad esempio, Cognitive Tutor di Carnegie Learning viene utilizzato nelle scuole superiori statunitensi per l'insegnamento della matematica, monitorando le interazioni degli studenti per evidenziare il livello di padronanza delle competenze. L'IA supporta anche la creazione di programmi di apprendimento personalizzati basati sulle prestazioni, identificando le aree di debolezza per adattare l'insegnamento alle esigenze specifiche. Modelli educativi come Kaspar, sviluppato dalla University of Hertfordshire, sono utilizzati per insegnare ai bambini autistici a interagire in modo appropriato. Questi avanzamenti consentono un approccio all'apprendimento più personalizzato e flessibile, impossibile da realizzare con i metodi tradizionali, e aprono nuove prospettive per l'istruzione a tutti i livelli.

3. *Finanziario*: il sondaggio di NVIDIA *State of AI in Financial Services* condotto nel febbraio 2024 ha rivelato che il 91% delle aziende di servizi finanziari sta valutando o ha già implementato l'IA per migliorare l'efficienza operativa. L'IA è particolarmente utile nel wealth management, in cui vengono utilizzati model-

li per decidere e suggerire cambiamenti di portafoglio basati su dati come età, propensione al rischio e reddito, rendendo la gestione patrimoniale più accessibile e personalizzata. Oltre a questo, l'IA viene ampiamente utilizzata per la rilevazione delle frodi. Infatti, modelli avanzati analizzano enormi quantità di dati per individuare schemi anomali e potenziali casi di frode più rapidamente e con maggiore precisione rispetto ai metodi tradizionali. Ad esempio, Feedzai utilizza modelli generativi per analizzare transazioni finanziarie in tempo reale, identificando schemi anomali e transazioni sospette attraverso il riconoscimento di pattern e la previsione comportamentale. Questo consente agli istituti finanziari di intervenire rapidamente per prevenire le frodi;

4. *Intrattenimento*: nel settore cinematografico, l'IA supporta attività precedentemente manuali, come la selezione di clip per i trailer, e la creazione di sceneggiature basate sull'analisi delle trame. Un esempio notevole è Fable Simulation, che ha sviluppato lo strumento di IA "Showrunner", in grado di generare interi episodi da semplici input testuali, utilizzato anche per un episodio della serie animata South Park. Nello stesso contesto, Amazon ha recentemente annunciato lo sviluppo di uno strumento di IAG per il doppiaggio di film e serie tv. Nel settore videoludico, invece, l'IA ha permesso di passare da personaggi con comportamenti predefiniti a quelli capaci di apprendere e sviluppare personalità proprie. Ad esempio, NVIDIA ACE è una piattaforma basata su tecnologie generative che permette di creare avatar digitali realistici e interattivi all'interno di ambienti videoludici. Utilizzando i LLM e la generazione vocale, questa tecnologia consente ai personaggi non giocanti (NPC) di dialogare in tempo reale con i giocatori, producendo conversazioni dinamiche e contestualizzate, invece delle interazioni predefinite tipiche dei videogiochi tradizionali;

5. *Scientifico*: anche diversi ambiti scientifici, come la fisica, la chimica, la matematica e la biologia, stanno subendo una profonda trasformazione grazie all'implementazione di modelli di intelligenza artificiale. AlphaFold, sviluppato da DeepMind, ha rivoluzionato la comprensione delle strutture proteiche, prevedendo la loro conformazione tridimensionale con un'accuratezza senza precedenti, un risultato fondamentale per la scoperta di nuovi farmaci e lo sviluppo di terapie innovative. Allo stesso modo, AlphaGeometry, un sistema di intelligenza artificiale progettato sempre da DeepMind, si distingue per la capacità di risolvere problemi complessi di geometria euclidea, raggiungendo un

livello paragonabile a quello dei medagliati d'oro delle Olimpiadi Internazionali di Matematica. Addestrato su un ampio set di dati sintetici, AlphaGeometry ha dimostrato prestazioni straordinarie, riuscendo a risolvere 25 problemi su 30 di geometria olimpica entro i limiti di tempo standard, avvicinandosi così alle capacità dei migliori studenti a livello globale. Nel campo dell'astrofisica, modelli come AstroBERT trovano applicazione nell'analisi di grandi cataloghi di galassie e nell'identificazione di esopianeti, accelerando processi che, in assenza di tali strumenti, richiederebbero anni di analisi manuale.

Per illustrare in modo sintetico l'ampiezza e la varietà di applicazioni dei modelli di ' nei diversi settori, la tabella 2.3 raccoglie alcuni esempi rappresentativi, oltre a quelli già evidenziati. La tabella fornisce una panoramica dei modelli più rilevanti, evidenziando il nome del modello, il suo sviluppatore, i parametri e il settore di utilizzo. Questa rappresentazione intende evidenziare come i modelli specializzati siano tendenzialmente più piccoli in termini di parametri rispetto ai modelli generali e che la loro natura sia prevalentemente open source.

Tabella 2.3: Alcuni modelli specializzati del mercato.

Modello	**Sviluppatore**	**Parametri**	**Settore**
FinBERT	Prosus AI	330 milioni	Finanziario
BloombergGPT	Bloomberg	50 miliardi	Finanziario
FinGPT	AI4Finance Foundation	3,67 milioni	Finanziario
ProteinBERT	Oxford Academic	170 milioni	Sanitario
BiomedGPT	Prosus AI	330 milioni	Sanitario
Amazon Comprehend Medical	Amazon Web Services	-	Sanitario
EduBERT	University of Edinburgh	-	Istruzione
PianoBART	Sun Yat-sen University	225 milioni	Intrattenimento
PhysBERT	Lawrence Berkeley National Laboratory, University of Naples Federico II	330 milioni	Fisica
Alpha Code	Google DeepMind	40 miliardi	Informatica

I modelli generali e specializzati stanno rivoluzionando il panorama tecnologico e industriale, con un impatto significativo sulla competizione tra settori esistenti e la possibile creazione di nuovi mercati in cui competere. Nel capitolo seguente, si analizzeranno le principali sfide legate all'implementazione e allo sviluppo di questi

modelli, evidenziando gli ostacoli tecnologici, economici e organizzativi che le aziende devono affrontare.

Capitolo 3

Gli economics dei modelli di IAG

3.1 La catena del valore

Lo sviluppo e la distribuzione di un FM si articolano lungo una catena del valore (figura 3.1) che coinvolge tre fasi principali: la costruzione dell'infrastruttura tecnologica, lo sviluppo dei modelli e la loro distribuzione:

1. *Infrastruttura*: la prima fase della catena del valore riguarda l'infrastruttura necessaria allo sviluppo e alla distribuzione dei modelli. Questa fase si basa su tre componenti essenziali: risorse computazionali, dati e competenze specializzate. L'addestramento dei FMs richiede l'utilizzo di chip acceleratori, di supercomputer o infrastrutture cloud, il cui costo rappresenta una voce di spesa significativa per le aziende del settore. I dati utilizzati nel processo di addestramento si suddividono in dati di pre-addestramento, che privilegiano la quantità e servono a sviluppare modelli generali, e dati di fine-tuning, che enfatizzano la qualità e vengono utilizzati per adattare i modelli a specifiche applicazioni. Infine, la disponibilità di risorse altamente qualificate nello sviluppo di modelli di IA, rappresenta un altro fattore critico, poiché la progettazione e ottimizzazione di un FM richiedono competenze avanzate lungo tutte le fasi della catena del valore;

2. *Sviluppo*: la seconda fase della catena del valore riguarda principalmente l'addestramento dei modelli. In questa fase, i dati e le risorse computazionali raccolte vengono impiegate per il pre-addestramento dei FMs. Una volta ad-

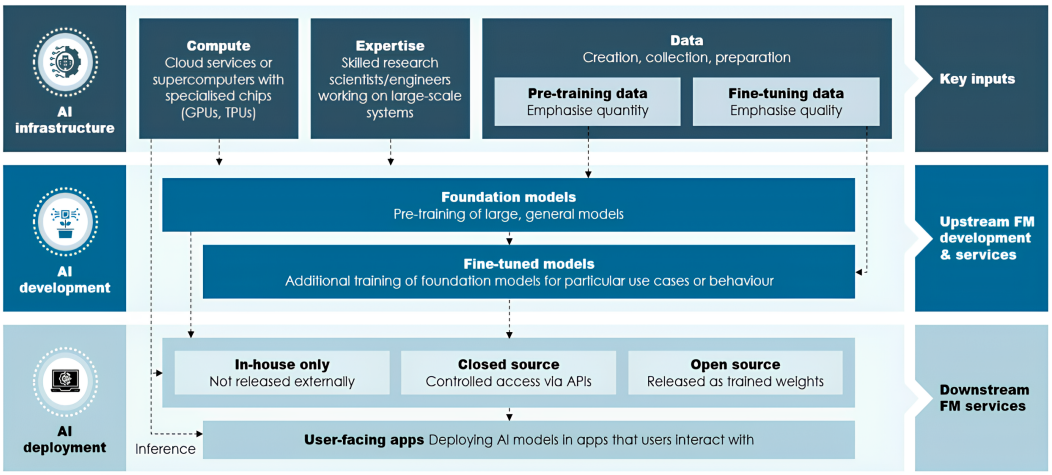

Figura 3.1: Catena del valore dell'IAG.[10]

destrati, questi modelli possono essere perfezionati attraverso le tecniche di fine-tuning, adattandoli a compiti specifici (*task-specific models*) o settori verticali (*domain-specific models*).

3. *Distribuzione*: l'ultima fase della catena del valore riguarda la distribuzione dei modelli e la loro modalità di utilizzo nel mercato. Una volta completato l'addestramento, i modelli possono essere mantenuti internamente (in-house only), senza essere resi disponibili all'esterno, oppure rilasciati in modalità closed source, con accesso controllato tramite API, che limitano l'uso del modello. In alternativa, alcuni modelli vengono distribuiti in modalità open source, consentendo a sviluppatori terzi di accedere ai pesi e di personalizzare il modello. Infine, i FMs possono essere integrati in applicazioni finali (user-facing apps), consentendo agli utenti di interagire direttamente con l'intelligenza artificiale attraverso prodotti e servizi basati sull'IAG.

Nelle sezioni successive, verranno analizzate nel dettaglio tutte le componenti della catena del valore dei FMs per evidenziarne i costi associati a ciascuna fase. Successivamente, si esamineranno le diverse modalità di distribuzione dei FMs quali SaaS e API. Infine, verranno presentati studi che valutano l'impatto delle scelte effettuate in fase di sviluppo sui livelli prestazionali dei modelli, al fine di comprendere come fattori quali la quantità e qualità dei dati, le risorse computazionali impiegate e la specializzazione del modello tramite fine-tuning influenzino le capacità dei fundation models in diverse applicazioni.

[10] Generative Artificial Intelligence: the Competitive Landscape. Copenhagen Economics (2024).

3.2 Costi per lo sviluppo di un Foundation Model

3.2.1 Dati per il pre-addestramento

Come anticipato, le prestazioni dei FMs segue un andamento prevedibile in funzione dell'aumento delle risorse computazionali e della quantità dei dati. Tuttavia, negli ultimi anni, lo sviluppo dell'intelligenza artificiale ha visto uno spostamento significativo dall'ottimizzazione delle architetture algoritmiche[11] al miglioramento della qualità e diversità dei dati, un approccio noto come *data-centric AI*. Tale cambiamento deriva dal fatto che l'incremento delle dimensioni e della complessità dei modelli ha cominciato a produrre rendimenti decrescenti, ovvero benefici marginali sempre più ridotti rispetto ai costi sostenuti. Di conseguenza, l'attenzione si è progressivamente spostata sul miglioramento della qualità dei dati utilizzati per l'addestramento, al fine di ottenere miglioramenti prestazionali più efficaci e sostenibili[12]. Questo cambiamento è ulteriormente motivato dalle limitazioni date dall'uso di dataset fissi, che non solo rischiano di amplificare bias e disuguaglianze, ma compromettono anche la capacità dei modelli di generalizzare efficacemente, esponendoli al rischio di overfitting[13] e riducendone le prestazioni su dati nuovi. In un approccio data-centric la raccolta dei dati assume quindi un ruolo centrale, soprattutto nel pre-addestramento del modello. In questa fase viene generalmente utilizzato un vasto dataset di dati grezzi non etichettati la cui tipologia varia a seconda del modello da sviluppare. Le diverse fonti utilizzate per la raccolta dei dataset incidono direttamente sulle caratteristiche che i dati devono possedere, quali:

1. *Dimensione e diversificazione*: più il dataset è vasto e diversificato più il modello riesce a generalizzare e coprire un numero più ampio di domini;

2. *Qualità*: i dati devono essere privi di errori grossolani, duplicati o contenuti dannosi, per evitare che il modello impari informazioni scorrette o distorte;

3. *Bilanciamento*: se i dati sono bilanciati si riduce il rischio di bias intrinseco, ovvero che nessun dominio o gruppo linguistico sia sovrarappresentato rispetto a un altro.

[11] Le architetture algoritmiche si riferiscono alla struttura dei modelli di intelligenza artificiale, come i transformer o altre configurazioni, che determinano il modo in cui i dati vengono elaborati per apprendere schemi e fare previsioni.

[12] Il salto qualitativo tra Chat GPT 3 e 4 è il risultato di una combinazione di interventi eseguiti sia sull'architettura del modello sia sulla qualità del set di dati di addestramento.

[13] Fenomeno in cui il modello dimostra prestazioni eccellenti sul dataset di pre-addestramento ma fallisce nel generalizzare su dati mai visti.

Risulta quindi necessario analizzare le metodologie di raccolta dei dataset in quanto esse, oltre a influenzare direttamente le prestazioni del modello, rappresentano una parte considerevole dei costi di sviluppo:

1. *Dataset pubblici*: offrono un accesso gratuito o a basso costo a grandi volumi di dati eterogenei per l'addestramento dei modelli. Ad esempio, l'Università di Harvard ha recentemente rilasciato un dataset contenente quasi un milione di libri di dominio pubblico, utilizzabile da chiunque per addestrare modelli linguistici di grandi dimensioni. Questi dataset sono spesso curati da istituzioni accademiche, organizzazioni no-profit o comunità open source. Tuttavia, la qualità dei dati può essere variabile, con possibili problemi di rumore[14], incompletezza e bias, che possono influire negativamente sulle prestazioni e sull'affidabilità del modello comportando, inoltre, investimenti aggiuntivi in termini di tempo e risorse per la pulizia e preparazione del dataset. Di seguito alcuni esempi di dataset pubblici più utilizzati per il pre-addestramento (tabella 3.1):

   Tabella 3.1: Esempi di dataset pubblici per l'addestramento dei FMs.

Dataset	**Descrizione**
Common Crawl	Dati testuali raccolti dal web
C4	Testo pulito derivato da Common Crawl
Books3	Dataset contenente 196.400 libri in formato testuale
MINT-1T	Dataset contenente 1.000 miliardi di token di testo, 3,4 miliardi di immagini e documenti HTML, PDF e ArXiv
LAION-5B	Un dataset contenente oltre cinque miliardi di coppie immagine-testo

2. *Dataset proprietari*: i dataset proprietari sono dati raccolti e curati da un'organizzazione che ne detiene il controllo esclusivo. Tale esclusività ne determina il valore superiore rispetto ai dati provenienti da fonti pubbliche, rendendoli più preziosi sul mercato. Il possesso di tali dati permette alle organizzazioni di sviluppare modelli più personalizzati e potenzialmente in grado di generare risultati più accurati. Tuttavia, la raccolta, l'archiviazione e la manutenzione di dataset proprietari richiedono ingenti investimenti in infrastrutture tecnologiche, strumenti di gestione dati e competenze specializzate, con costi operativi elevati dovuti anche alla necessità di aggiornamenti continui per mantenerli rilevanti e allineati ai cambiamenti del mercato;

[14]Errori o variazioni nei dati che non rappresentano informazioni rilevanti o utili per l'addestramento del modello.

3. *Dataset su licenza*: rappresentano un'alternativa per accedere a dati proprietari di terze parti di alta qualità, curati e settoriali, riducendo i costi legati alla raccolta e alla pre-elaborazione. Tuttavia, i costi delle licenze sono particolarmente elevati generando un impatto significativo sia sul mercato dei fornitori di dati sia su quello degli sviluppatori dei modelli. Infatti, tali licenze sono accessibili a grandi imprese dotate di risorse economiche rilevanti, e, parallelamente, solo i principali fornitori di dati riescono a stipulare accordi di questo tipo. Ad esempio, Open AI, solo nel 2024 ha stretto diversi accordi di licenza con importanti fornitori di dati, tra cui TIME, Reddit e il Financial Times (box 1). L'attuale quadro normativo sui dati solleva inoltre diverse criticità che potrebbero ostacolare l'innovazione tecnologica e l'accesso equo ai dataset proprietari, spingendo alcune organizzazioni a sviluppare nuove proposte per superare tali limitazioni[15];

Box 1: Licenze OpenAI nel 2024

Tra aprile e giugno 2024, OpenAI ha stipulato accordi di licenza con diverse società di media, tra cui TIME, The Atlantic, Vox Media, News Corps, Dotdash Meredith, Financial Times, Le Monde e Prisa Media. Questi accordi permettono a OpenAI di addestrare i suoi modelli utilizzando i contenuti degli articoli delle testate, in modo da fornire agli utenti informazioni autorevoli e aggiornate. Ad esempio, tramite l'accordo pluriennale con TIME, OpenAI ha accesso sia ai contenuti attuali che agli articoli degli ultimi 101 anni contenuti negli archivi della società. Le risposte generate a partire da questi documenti vengono evidenziate all'interno di ChatGPT con una citazione alla fonte originale del TIME. Dall'altra parte, il TIME ha ottenuto accesso alle tecnologie di OpenAI per migliorare il proprio giornalismo e sviluppare nuovi prodotti basati sull'IA. Quindi, se da un lato OpenAI aumenta l'accuratezza dei propri modelli tramite i dati forniti dalle testate giornalistiche, le società di media riescono a ottenere visibilità grazie all'indicizzazione delle risposte

[15]Nel giugno 2024 viene fondata la *Dataset Providers Alliance (DPA)* per affrontare le sfide legate all'uso e alla regolamentazione dei dati nell'ambito dell'IA. La DPA propone cinque modelli di licenza alternativi per favorire trasparenza e accesso equo ai dati: (i) licensing basato sull'uso, (ii) licensing basato sui risultati, (iii) modello di abbonamento, (iv) licensing ibrido e (v) licensing specifico per settore. Tali modelli mirano a ridurre i costi elevati delle licenze, che attualmente limitano l'accesso ai dataset alle sole grandi aziende, e a creare un mercato più competitivo e sostenibile per lo sviluppo dei modelli.

alla fonte.

Gli accordi di licenza stipulati da OpenAI rispondono principalmente all'esigenza di ottenere dati di alta qualità per migliorare le prestazioni dei modelli linguistici. Tuttavia, potrebbero anche essere interpretati come una risposta alla causa legale intentata nei primi mesi del 2024 dal New York Times e altre testate giornalistiche, che hanno accusato OpenAI di aver utilizzato contenuti protetti da copyright durante l'addestramento dei suoi modelli, ottenuti tramite tecniche di web scraping senza autorizzazione o compensazione economica. Alla fine del 2024, altre otto testate giornalistiche si sono aggiunte alla causa, rafforzando le pressioni sul tema del rispetto della proprietà intellettuale.

Nel frattempo, OpenAI ha cercato di espandere le proprie collaborazioni con altre realtà europee, provando ad avviare una partnership con il gruppo editoriale italiano GEDI, editore di testate come la Repubblica e La Stampa. Tuttavia, questa partnership ha incontrato ostacoli significativi a causa delle restrizioni imposte dal *Garante per la protezione dei dati personali (GPDP)* italiano. Il GPDP ha espresso preoccupazioni per la condivisione di archivi digitali contenenti dati personali e sensibili, che potrebbero violare le normative del *GDPR (Regolamento Generale sulla Protezione dei Dati)*. Di conseguenza, al momento la partnership è sospesa, riflettendo le sfide che gli sviluppatori di modelli devono affrontare per bilanciare innovazione tecnologica e rispetto delle regolamentazioni europee.

4. *Data Scraping*: rappresenta una delle principali metodologie di raccolta dati utilizzate per addestrare modelli di IAG. Questa tecnica consiste nell'estrazione automatizzata di informazioni da siti web e piattaforme online mediante bot, noti anche come *crawler* o *spider*, capaci di simulare la navigazione web umana. I dati raccolti, spesso pubblicamente disponibili, includono testi, immagini, e altre tipologie di contenuti, successivamente memorizzati e analizzati per sviluppare dataset utili all'addestramento. Ad esempio, il dataset Common Crawl precedentemente citato è stato completamente costruito tramite questo metodo. Uno dei principali vantaggi del data scraping è la possibilità di accedere a enormi volumi di dati a costi relativamente contenuti rispetto all'acquisto

di dataset curati tramite licenze. Inoltre, questa tecnica consente di attingere a fonti diversificate e aggiornate, garantendo maggiore varietà. Tuttavia, i dati raccolti attraverso lo scraping possono presentare problemi significativi in termini di liceità[16] sull'utilizzo dei dati raccolti e qualità, tra cui la presenza intrinseca di rumore e informazioni non strutturate o incomplete, che necessitano di ulteriori processi di pulizia e preparazione;

5. *Dataset sintetici*: i dataset sintetici rappresentano una risorsa innovativa nel campo dell'intelligenza artificiale, in quanto vengono generati tramite modelli di IA. La generazione di dataset sintetici può avvenire attraverso diverse modalità: (i) il dataset può essere completamente generato artificialmente; (ii) si può partire da un dataset reale e integrarlo o modificarlo con dati sintetici; (iii) si può adottare un approccio ibrido, che prevede la combinazione casuale di record provenienti da dataset reali e sintetici. Questa tecnologia è in costante crescita grazie alla sua capacità di simulare scenari rari o complessi che sarebbero difficili, se non impossibili, da catturare con dati reali. Inoltre, rappresentano una soluzione in grado di coniugare lo sviluppo dei modelli con la tutela della privacy[17]. Chiaramente, la possibilità di creare un modello che generi un dataset sintetico utile successivamente alla creazione di un atro modello di IA comporterebbe ingenti risorse sia in termini di infrastrutture tecnologiche che di risorse umane, il che risulterebbe una tecnica accessibile solo a pochissime aziende. Tuttavia, ci sono diverse soluzioni che offrono la possibilità di produrre questi dataset, come la famiglia di modelli open source Nemotron-4 340B di NVIDIA o il servizio cloud Amazon SageMaker, che forniscono una pipeline per generare e perfezionare dati sintetici abbassando notevolmente i costi di accesso a questi dataset.

Quelli appena esposti rappresentano i principali metodi di accesso ai dataset per il pre-addestramento di un modello di IAG. Per offrire una visione d'insieme di questi, la tabella 3.2 compara i diversi dataset sotto tre dimensioni: costo, qualità e accessibilità.

[16] La raccolta di dati tramite scraping può violare normative come il GDPR (General Data Protection Regulation) in Europa e il CCPA (California Consumer Privacy Act) negli USA, comportando sanzioni e costi di compliance per le aziende.

[17] I dati sintetici vengono menzionati in diversi punti all'interno dell'AI Act, in particolare come strumento che permette l'addestramento etico dei modelli in quanto, non essendo legati a nessun individuo, non violano la tutela della privacy.

Tabella 3.2: Tabella comparativa dei diversi dataset.

Dataset	Costo	Qualità dei dati	Accessibilità
Pubblici	Basso	Variabile (spesso rumorosa)	Alta (spesso gratuiti)
Licensing	Alto	Alta (curati e specifici)	Limitata
Web scraping	Medio-basso	Variabile (rumore intrinseco)	Media (legato a risorse tecniche)
Sintetici	Medio-alto	Media (controllabile)	Media
Proprietari	Molto alto	Molto alta	Molto limitata

I costi legati ai dati non si limitano alla sola fase di raccolta: affinché possano essere utilizzati, infatti, i dati devono essere sottoposti a una pipeline di preparazione che li renda idonei all'addestramento dei modelli[18].

3.2.2 Hardware per lo sviluppo

Una volta raccolto e preparato il dataset, la seconda componente fondamentale è un'infrastruttura computazionale ad alte prestazioni, necessaria per gestire il numero considerevole di operazioni richieste per il pre-addestramento di un FM. Questa infrastruttura è suddivisa in diverse componenti, quali:

1. Chip acceleratori come GPU o TPU utili sia nella fase di preparazione dei dati che nell'addestramento del modello;
2. Infrastrutture per l'archiviazione di grandi quantità di dati;
3. Infrastruttura di rete per la connessione;
4. Altre componenti hardware come sistemi di raffreddamento e alimentazione.

La componente più rilevante è rappresentata dai chip acceleratori, in quanto costituiscono il cuore delle infrastrutture necessarie per l'addestramento di modelli avanzati. La loro importanza non è esclusivamente economica, vista l'elevata incidenza sui costi totali, ma anche strategica, dal momento che determinano in larga misura le prestazioni e la scalabilità dei modelli, influenzando così direttamente la competitività delle organizzazioni impegnate nel loro sviluppo.

[18] In appendice B alcuni esempi di processi a cui sono sottoposti i dati e il loro impatto sui costi di sviluppo.

I chip maggiormente utilizzati sono le GPU (Graphics Processing Units), acceleratori progettati inizialmente per il rendering grafico, come la modellazione 3D nel settore videoludico. A differenza delle CPU (Central Processing Units), ottimizzate per gestire un numero limitato di operazioni sequenziali, le GPU eccellono nell'elaborazione parallela grazie alla loro architettura composta da migliaia di core[19] meno potenti, ma altamente specializzati. Questa caratteristica rende le GPU ideali per i compiti di addestramento e inferenza nei modelli di IAG. L'addestramento di modelli di grandi dimensioni, come i LLM, richiede un'elevata quantità di queste unità computazionali, che possono variare da migliaia a decine di migliaia di GPU, comportando costi infrastrutturali elevati. Ad esempio, le GPU più utilizzate per lo sviluppo di modelli di IAG sono gli acceleratori di punta sviluppati da NVIDIA, le GPU A100 e H100, che hanno un prezzo a unità medio compreso tra $10.000 e $15.000 per le A100 e oltre i $30.000 per le H100.

Tabella 3.3: Numero di GPU impiegate per addestrare alcuni modelli di punta.[20]

Modello	**Sviluppatore**	**Modello GPU**	**Numero di GPU**
LLama 3.1	Meta AI	NVIDIA H100 SXM5 80GB	16.384
GPT 4	OpenAI	NVIDIA A100 SXM4 40 GB	25.000

Alla luce dell'elevato costo di adozione di infrastrutture hardware interne, è essenziale effettuare un confronto con la principale alternativa che le aziende hanno per accedere alle risorse computazionali, ovvero le piattaforme di cloud computing. L'infrastruttura on-premise richiede ingenti spese in conto capitale (CapEx), con investimenti iniziali rilevanti per l'acquisto di hardware, ai quali si aggiungono i costi ricorrenti di manutenzione e aggiornamento. Al contrario, il cloud computing adotta un modello di spesa operativa (OpEx), in cui le aziende pagano per i servizi utilizzati su base di abbonamento o secondo un modello "pay-as-you-go". Questo approccio elimina la necessità di grandi investimenti iniziali e consente una maggiore flessibilità, poiché le risorse computazionali possono essere scalate dinamicamente in base alle esigenze. Inoltre, per incentivare l'adozione di infrastrutture cloud, piattaforme come AWS Active e Google Cloud offrono programmi dedicati a startup, che includono crediti gratuiti per l'utilizzo degli acceleratori, contribuendo così a ridurre i costi iniziali.

[19] Unità di elaborazione fondamentali all'interno di una GPU, responsabili dell'esecuzione di calcoli e funzioni logiche.

[20] Epoch AI, Data on Notable AI Models'. Published online at epoch.ai. Retrieved from https://epoch.ai/data/notable-ai-models' [online resource].

Per quanto riguarda i costi indiretti, le infrastrutture on-premise richiedono una manutenzione regolare, che comprende riparazioni hardware, aggiornamenti software e monitoraggio costante dei sistemi. Questo implica la necessità di un team IT dedicato, con un conseguente aumento dei costi legati alla manodopera specializzata. Inoltre, i componenti hardware delle infrastrutture on-premise sono soggetti a deprezzamento nel tempo e richiedono aggiornamenti o sostituzioni periodiche, generando costi significativi nel lungo termine. Al contrario, i servizi cloud riducono drasticamente questi costi indiretti, poiché la manutenzione, il supporto e gli aggiornamenti sono gestiti direttamente dal fornitore del servizio.

Tuttavia, il solo confronto tra i costi diretti e indiretti non è sufficiente per determinare quale approccio sia più vantaggioso. Sebbene il cloud computing offra flessibilità, scalabilità ed elimini la necessità di investimenti iniziali, presenta alcune problematiche, tra cui la latenza della rete, che può rappresentare un limite per applicazioni che richiedono elaborazioni in tempo reale. Inoltre, il controllo dei dati passa al fornitore del servizio, con implicazioni significative in termini di sicurezza e privacy. Vi è, inoltre, il rischio di lock-in, poiché cambiare fornitore o migrare a un'infrastruttura on-premise può risultare costoso e complesso. Infine, nel lungo periodo, i costi operativi del cloud possono superare quelli di un'infrastruttura on-premise, soprattutto se l'addestramento di un modello si protrae per lunghi periodi. Dall'altro lato, l'approccio on-premise offre un controllo completo sui dati, un aspetto cruciale per settori con normative stringenti in materia di privacy. Inoltre, la latenza è generalmente inferiore poiché i dati non devono essere trasferiti a server remoti e la personalizzazione è maggiore, consentendo configurazioni specifiche per esigenze particolari.

Oltre a questi due approcci principali, esistono ulteriori possibilità per accedere alle infrastrutture necessarie allo sviluppo di modelli di intelligenza artificiale, come le collaborazioni con centri di ricerca: queste nascono con l'obiettivo di promuovere l'innovazione nel panorama dell'IA offrendo l'accesso a infrastrutture computazionali, tipicamente supercomputer, utili allo sviluppo dei modelli. Ad esempio, il progetto europeo *EuroHPC* ha tra gli obiettivi quello di costruire delle "fabbriche di intelligenza artificiale", ovvero strutture che intendono includere supercomputer, data center e servizi di supercalcolo orientati all'intelligenza artificiale. Queste strutture saranno aperte a utenti sia pubblici che privati, con condizioni di accesso dedicate a startup e piccole imprese.

3.2.3 Risorse umane

Lo sviluppo e la gestione dei modelli richiedono risorse umane altamente qualificate, che rappresentano una parte considerevole dei costi totali di sviluppo e comprendono i costi per il reclutamento, la formazione e la gestione del personale. Il panorama delle competenze coinvolte è dominato da figure chiave come data scientist, machine learning engineer o NLP engineer ciascuna delle quali contribuisce in modo essenziale alla progettazione, addestramento e gestione dei modelli su larga scala.

La rapida espansione del mercato dell'IAG ha determinato un aumento significativo della domanda di talenti specializzati. La figura 3.2 evidenzia questa tendenza, mostrando che le professioni che subiranno una crescita più rapida fanno riferimento a ruoli chiave nel campo dell'intelligenza artificiale. Tra i ruoli in maggiore espansione troviamo specialisti in Big Data, ingegneri FinTech e specialisti in IA e Machine Learning, figure chiave per lo sviluppo e l'ottimizzazione dei modelli di IAG.

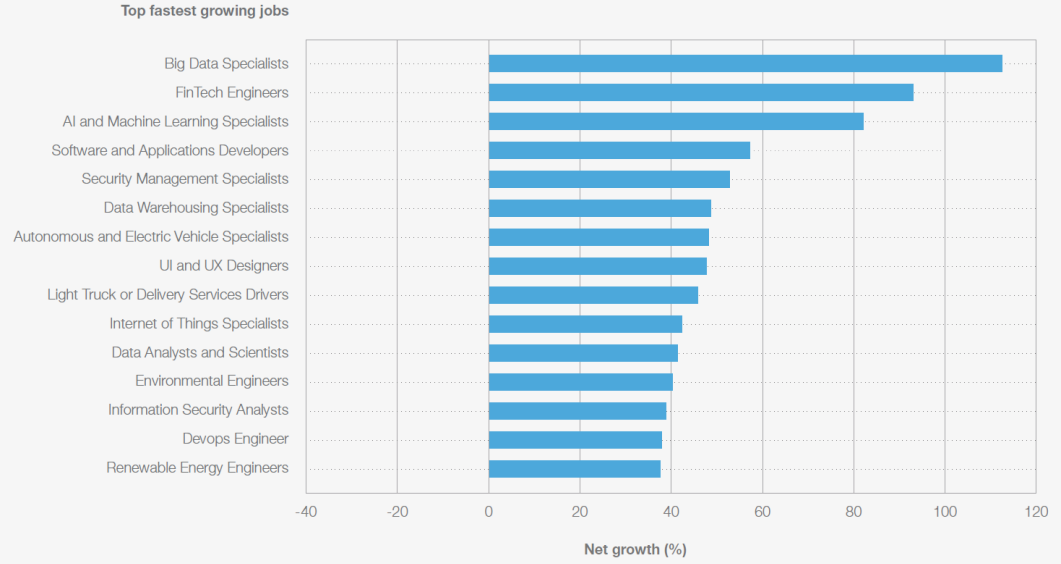

Figura 3.2: Lavori con la maggior crescita netta, 2025-2030.[21]

I salari per queste posizioni sono tra i più alti nel settore tecnologico: un report del 2024[22] analizzando i salari dei professionisti nel campo del data science dal 2020 al 2024, in particolare, ha evidenziato come vi sia stato un aumento costante negli anni, con un picco nel 2023, seguito da una leggera diminuzione nel 2024, con un salario medio annuo passato da $102.251 nel 2020 a $153.733 nel 2023 (nel 2024, $151.510). Ciò che contraddistingue queste professioni dalle altre è l'alto livello di esperienza

[21] Future of Jobs Report 2025, World Economic Forum.
[22] E. A. Bagyam, "Analysis of data science job salaries from 2020 to 2024: trends and influencing factors", Surya Publications, 2024.

e il titolo professionale: i ruoli senior ed executive percepiscono stipendi quasi doppi rispetto alle posizioni entry-level (mediamente $12.500 in più per gli executive), mentre figure altamente specializzate come AI Architect e AI Engineer sono tra le più remunerate. La dimensione aziendale influisce significativamente sui compensi, con le aziende di medie dimensioni che offrono salari più competitivi, seguite da quelle grandi. Inoltre, alcune stime per il 2025[23] indicano un ulteriore aumento, con retribuzioni previste tra $130.000 e $155.000, a conferma della crescente competizione tra le aziende per attrarre e trattenere professionisti altamente qualificati. Un ulteriore aspetto importante è la modalità di lavoro in questo settore, che incide sulle remunerazioni: le posizioni full remote tendono a essere pagare meno rispetto a quelle in sede (riduzione media di $3.700).

Accanto al vantaggio di specializzazioni sempre più ricercate e performanti per rispondere alle richieste dell'IA, vi è il rischio di un'eccessiva automazione che potrebbe portare a un turnover dannoso per i dipendenti. Per evitare ciò, le aziende investono sempre più in programmi di retention e welfare aziendale per ridurre il rischio di turnover e trattenere i talenti più qualificati[24]. In risposta, molte organizzazioni adottano strategie di formazione continua, piani di sviluppo personalizzati e maggiore flessibilità lavorativa.

3.2.4 Energia

L'addestramento dei FMs rappresenta uno dei processi più dispendiosi in termini energetici[25]. Per ridurre i consumi diverse aziende stanno adottando soluzioni come

[23]Refonte Learning. (2024, November 25). AI Salary Trends 2025: Unlocking High-Paying Careers in Artificial intelligence and Related Roles. https://www.linkedin.com/pulse/ai-salary-trends-2025-unlocking-high-paying-careers-artificial-scjoe/

[24]Secondo un recente report della Banca d'Italia, le professioni maggiormente a rischio di automazione si collocano prevalentemente nei due quintili più alti della distribuzione salariale, in particolare nel settore dei servizi. Questo fenomeno genera effetti ambigui sulla disuguaglianza economica: se da un lato la sostituzione di alcune mansioni potrebbe ridurre il divario salariale, dall'altro il passaggio verso occupazioni meno esposte all'IA si associa spesso a una perdita di reddito. In particolare, i lavoratori che lasciano posizioni altamente esposte e sostituibili per ruoli meno esposti subiscono una riduzione salariale, mentre coloro che riescono a spostarsi verso occupazioni complementari all'IA registrano aumenti retributivi più significativi. Inoltre, il livello di istruzione gioca un ruolo chiave nella capacità di adattamento al cambiamento tecnologico: i lavoratori più qualificati hanno maggiori opportunità di ricollocarsi in ruoli complementari all'IA, ma il loro vantaggio salariale si riduce qualora si spostino verso professioni meno esposte. Per ulteriori informazioni si veda: Banca d'Italia, "An assessment of occupational exposure to artificial intelligence in Italy", ottobre 2024.

[25]Secondo Lazzaro, D. et al. (2023) nello studio *Minimizing energy consumption of deep learning models by Energy-Aware training*, la quantità enorme di calcoli necessari per processare i dati all'interno delle reti neurali porta a un notevole consumo di energia, soprattutto quando si utilizzano potenti processori come GPU e TPU. Per ridurre l'impatto energetico introducono un metodo che permette di ridurre il numero di attivazioni neurali non necessarie durante l'addestramento. Questo metodo, definito *Energy-Aware Training (EAT)*, applicato a modelli conosciuti, potrebbe ridurre il consumo energetico fino al 27%, mantenendo un'accuratezza comparabile a quella delle

l'*ASIC (Application-Specific Integrated Circuits)*, una tecnologia che consente di ridurre i consumi energetici fino a dieci volte rispetto alle tradizionali GPU. Questi circuiti personalizzati sono progettati per evitare operazioni ridondanti e migliorare l'efficienza computazionale, ottimizzando il processo di addestramento.

Inoltre, le aziende che sviluppano modelli su larga scala tendono a localizzare i propri data center in paesi con una maggiore disponibilità di energia rinnovabile e tariffe elettriche più contenute, riducendo così l'impatto economico e ambientale del processo. Molti provider cloud stanno già investendo in soluzioni carbon neutral, con data center alimentati interamente da fonti rinnovabili, per diminuire la loro impronta ecologica

Parallelamente, la necessità di ottimizzare il consumo energetico ha portato allo sviluppo del concetto di Green AI, un movimento che promuove l'adozione di pratiche di sviluppo sostenibili. Tra le strategie più efficaci vi è la riduzione della complessità dei modelli e l'uso di tecniche come il pruning, che possono abbattere il consumo energetico fino al 50% senza compromessi significativi in termini di prestazioni[26].

Oltre alla fase di addestramento, il consumo energetico di un modello dipende anche dalla sua tipologia e dalle applicazioni in cui viene utilizzato. I modelli generali, progettati per eseguire un'ampia gamma di compiti, risultano significativamente più onerosi rispetto ai modelli specializzati, che vengono ottimizzati per applicazioni specifiche e, quindi, più efficienti dal punto di vista computazionale. Inoltre, il tipo di attività svolta influisce notevolmente sui costi energetici: la generazione di immagini, ad esempio, è molto più intensiva rispetto alla generazione di testo, richiedendo risorse computazionali superiori per la produzione di contenuti complessi.

Un altro aspetto critico è il consumo energetico della fase di inferenza, che in alcuni casi può eguagliare o addirittura superare quello dell'addestramento. Questo è particolarmente evidente nei modelli distribuiti su larga scala, con milioni di utenti che utilizzano il modello simultaneamente. Un caso emblematico è quello di ChatGPT, il cui utilizzo massivo richiede ingenti risorse energetiche per garantire risposte in tempo reale. Secondo Sam Altman, CEO di OpenAI, la fase di inferenza generera costi talmente elevati a tal punto che i prezzi della sottoscrizione al servizio OpenAI Pro non coprirebbero integralmente le spese operative, a causa dell'intenso utilizzo da parte degli utenti.

reti standard, con una perdita massima del 3% nelle prestazioni predittive.

[26]Yarally, T. et al. (2023) Uncovering Energy-Efficient Practices in Deep Learning Training: Preliminary Steps towards Green AI. https://arxiv.org/abs/2303.13972.

L'elevata domanda di FMs e la loro crescente applicazione in diversi settori rendono dunque essenziale un approccio più efficiente e sostenibile, sia in termini di hardware che di ottimizzazione degli algoritmi, al fine di contenere i costi e ridurre l'impatto ambientale dell'IAG.

3.2.5 Addestramento

L'addestramento di un FM rappresenta il cuore dello sviluppo di un modello di IAG, in quanto è la fase in cui il modello apprende come identificare pattern, relazioni e conoscenze dai dati grezzi. Come anticipato nel primo capitolo, questo processo si basa su tecniche di apprendimento non supervisionato che ottimizzano miliardi di parametri attraverso l'elaborazione di grandi dataset, rendendo il modello capace di risolvere problemi complessi su un vasto numero di domini. Durante la fase di addestramento vengono utilizzati ulteriori dataset oltre a quelli di pre-addestramento: (i) dataset di validazione[27], utili a regolare i parametri e identificare eventuali problemi di overfitting e (ii) dataset di test[28], impiegati alla fine del processo di addestramento per valutare le prestazioni del modello su dati o casi mai visti (edge-cases).

La combinazione delle risorse evidenziate nei precedenti paragrafi (dati, hardware, risorse umane ed energia) rappresento i costi legati alla fase di addestramento del modello. Queste componenti non solo determinano le prestazioni del modello, ma rappresentano anche la maggior parte dei costi associati all'intero processo di sviluppo. Studi recenti hanno evidenziato come questi costi stiano subendo una crescita esponenziale, rendendo sempre più complessa e onerosa la competizione nel settore dei FMs. Infatti, Cottier, B. et al. in *The rising costs of training frontier AI models, (2024)* hanno effettuato una stima quantitativa dei costi di addestramento per i modelli di frontiera, evidenziando come a partire dal 2016 questi siano aumentati di 2,4 volte all'anno (figura 3.3). Questa crescita, guidata dall'esigenza di infrastrutture computazionali avanzate, quantità immense di dati e risorse umane altamente qualificate, implica che solo le organizzazioni con notevoli risorse finanziarie possono permettersi di sviluppare modelli su larga scala.

[27] Generalmente, questi dataset sono un sottoinsieme di quelli di addestramento, in quanto servono per valutare le prestazioni del modello sui dati che rientrano nello stesso dominio.

[28] Dati completamente nuovi che non vengono generati a partire dal dataset di addestramento.

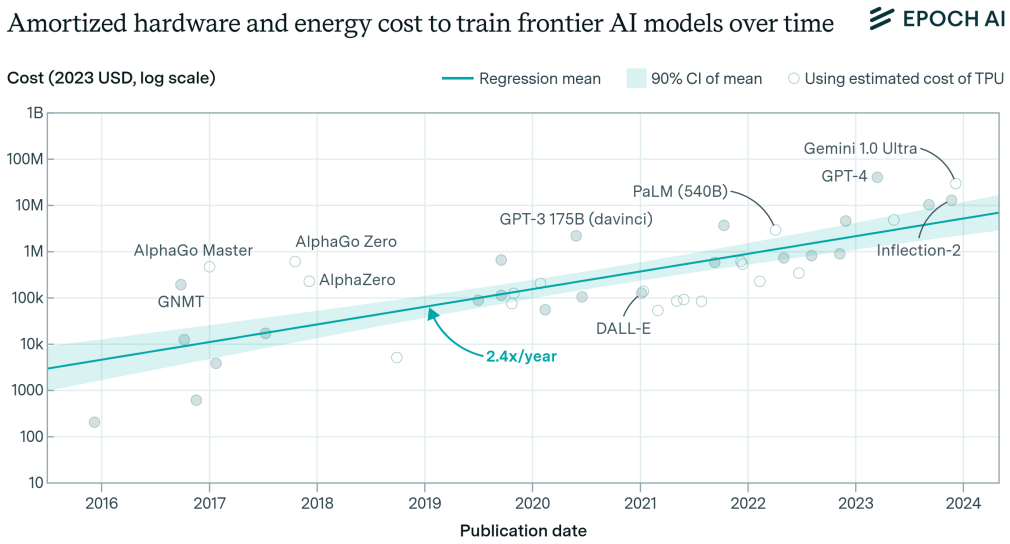

Figura 3.3: Andamento dei costi hardware ed energetici nel tempo per addestrare i modelli di frontiera.[29]

Lo studio evidenzia che i costi principali nell'addestramento sono rappresentati dalle risorse umane e dall'hardware, che risultano anche le componenti più variabili tra i diversi modelli (figura 3.4). Le risorse umane costituiscono una quota significativa, variando dal 29% in GPT-4 al 49% in Gemini 1.0 Ultra. L'hardware, comprendente chip acceleratori, componenti server e interconnessioni tra cluster, rappresenta oltre la metà dei costi totali in tutti i modelli, con un'incidenza che varia dal 53% in Gemini 1.0 Ultra al 61% in GPT-3 175B. Al contrario, i costi energetici sono non solo la componente meno variabile, ma anche quella con l'incidenza minore nella fase di addestramento.

[29]Cottier, B. et al. (2024) The rising costs of training frontier AI models. https://arxiv.org/abs/2405.21015.

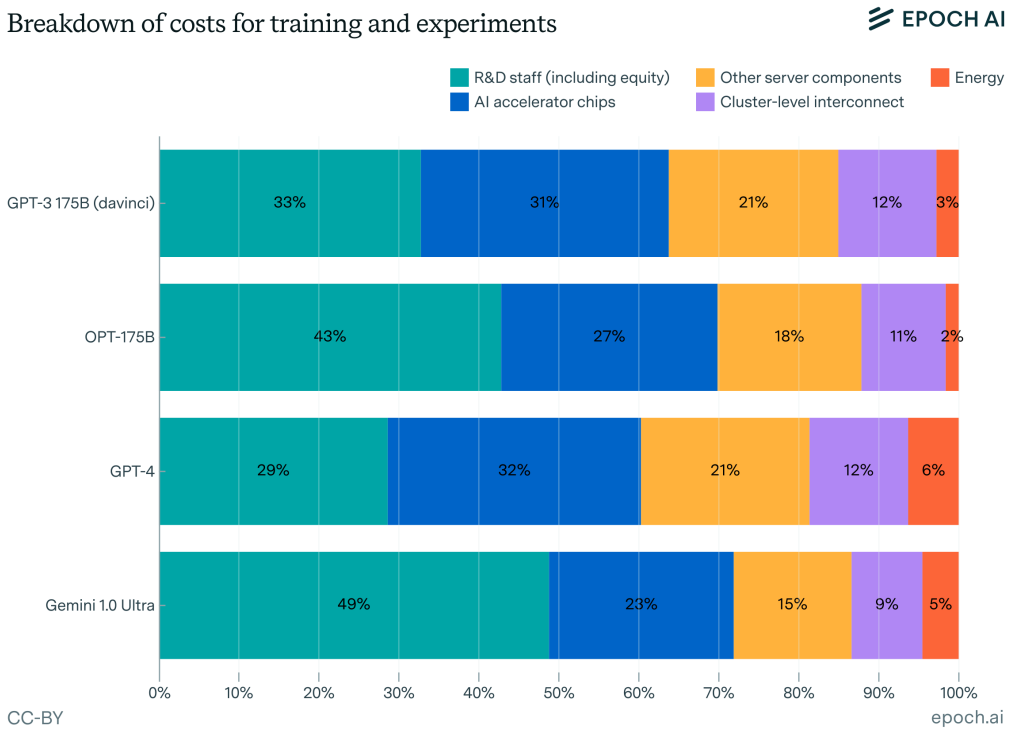

Figura 3.4: Struttura dei costi per alcuni dei modelli di frontiera.[30]

I costi, che oggi possono superare i 40 milioni di dollari per un singolo modello come GPT-4, potrebbero raggiungere il miliardo di dollari entro il 2027, ponendo una barriera significativa per i nuovi entranti e consolidando il dominio delle poche aziende con le risorse necessarie. Tuttavia, nel capitolo finale di questo elaborato verrà discusso e analizzato il caso della società cinese DeepSeek, che sembrerebbe aver messo in dubbio la necessità di sostenere costi così elevati per lo sviluppo di un modello competitivo.

3.2.6 Fine-tuning

Prima di essere distribuito, generalmente un FM viene sottoposto a un processo che prende il nome di fine-tuning. Tale attività consiste nell'adattamento del modello pre-addestrato a compiti o domini specifici, mediante ulteriori operazioni di apprendimento supervisionato o non supervisionato, che sfruttano dataset più circoscritti rispetto a quelli impiegati nella fase iniziale. Questi dataset hanno caratteristiche comuni con quelli di pre-addestramento: entrambi devono essere di alta qualità, di-

[30] Cottier, B. et al. (2024) The rising costs of training frontier AI models. https://arxiv.org/abs/2405.21015.

versificati e bilanciati. Tuttavia, hanno caratteristiche distintive che li differenziano dai precedenti, ovvero:

1. *Alta specificità*: i dati devono essere pertinenti al dominio o al compito specifico a cui il modello sarà applicato. Ad esempio, per un modello utilizzato in ambito medico il dataset di fine-tuning includerà termini tecnici, casi clinici e documenti scientifici del settore sanitario;

2. *Dimensionalità minore*: il dataset è generalmente di dimensioni ridotte, in quanto focalizzato, mirato e rappresentativo dello specifico dominio.

È fondamentale sottolineare che l'utilizzo di dataset troppo ristretti nella fase di pre-addestramento potrebbe compromettere la capacità del modello di generalizzare, aumentando il rischio di overfitting. Analogamente ai dati di pre-addestramento, i dataset per il fine-tuning possono essere ottenuti attraverso diverse modalità, tra cui l'utilizzo di dataset pubblici, il data licensing, lo scraping, i dataset sintetici o dataset proprietari. Tuttavia, vi sono differenze significative nel modo in cui queste modalità vengono impiegate per il fine-tuning, principalmente a causa delle caratteristiche e degli obiettivi distintivi di questa fase. Infatti, vengono principalmente utilizzati dataset proprietari e i dataset su licenza che garantiscono un'elevata qualità e specializzazione, caratteristiche difficilmente ottenibili tramite, ad esempio, dataset pubblici che offrono invece una elevata quantità di dati e diversità, caratteristiche più utili per il pre-addestramento del modello.

L'impiego di dataset circoscritti comporta una riduzione della complessità del modello in termini di numero di parametri. Inoltre, poiché il fine-tuning consiste nell'adattare un modello precedentemente addestrato (operazione che implica che la maggior parte del lavoro computazionale sia già stata svolta durante il pre-addestramento) esso richiede esclusivamente l'ottimizzazione del modello per un dominio o un task specifico. Di conseguenza, le risorse computazionali necessarie in questa fase risultano significativamente inferiori[31]. Tuttavia, nel caso in cui non si disponga di un FM proprietario, ma si voglia sviluppare un modello specializzato, esistono principalmente due modalità per personalizzare o adattare un modello pre-addestrato esistente:

1. *Personalizzare un modello per-addestrato tramite API*: è un servizio offerto dalle grandi aziende che possiedono modelli pre-addestrati, come Open AI o

[31]Talvolta, il processo di fine-tuning può richiedere una singola GPU.

Google, che permettono di adattare i loro modelli senza la necessità di investire in infrastrutture hardware. Questo avviene fornendo direttamente il dataset tramite le interfacce API. Dal punto di vista dei costi, questi sono distribuiti su base operativa, anche in questo caso secondo un modello "pay-as-you-go";

2. *Addestramento di modelli open-source*: rappresenta un approccio on-premise con la differenza che avviene su un modello pre-addestrato disponibile gratuitamente o a costi ridotti disponibile su piattaforme come Huggin Face. Questo approccio permette un elevato livello di personalizzazione ma richiede la costruzione di un'infrastruttura hardware, oltre che alle competenze tecniche per gestire tutto il processo di fine-tuning. È anche possibile adottare un approccio ibrido utilizzando risorse computazioni in cloud tramite piattaforme come AWS EC2.

Seppur il processo di fine-tuning richieda meno risorse computazionali, introduce costi significativi legati alla raccolta, ad esempio tramite licenza, di dataset di alta qualità. Inoltre, potrebbe sorgere la necessità di etichettare il dataset per assicurare una maggiore accuratezza del modello al dominio specifico. Questa attività risulta particolarmente onerosa in quanto in alcuni casi è necessario coinvolgere annotatori esperti o implementare strumenti avanzati per automatizzare il processo[32]. Infine, in aggiunta figurano i costi per l'impiego di tecniche avanzate come il *Reinforcement Learnig From Human Feedback (RLHF)*[33] che, sebbene permetta di migliorare le performance del modello basandosi sul feedback umano, richiede il coinvolgimento di annotatori qualificati e l'addestramento di modelli di ricompensa aggiungendo costi operativi al processo di sviluppo.

3.3 Modalità di distribuzione di un Foundation Model

Una volta completato l'addestramento, i FMs proprietari possono essere distribuiti secondo diverse strategie. In questa sezione vengono approfonditi i concetti precedentemente esposti nel primo capitolo, quale la distribuzione del modello come servizio principale o l'integrazione in altri servizi per sviluppare applicazioni specializzate.

[32] In appendice C una panoramica delle diverse tecniche di etichettatura.

[33] È una tecnica di affinamento dei modelli di intelligenza artificiale che prevede l'utilizzo di feedback umani per ottimizzare il comportamento del modello. In particolare, il modello viene addestrato a massimizzare un "premio" basato sulle preferenze espresse da valutatori umani, consentendo così di generare risposte più coerenti e allineate ai valori e alle aspettative degli utenti rispetto ai modelli addestrati esclusivamente su dati non supervisionati o parzialmente supervisionati.

3.3.1 Foundation Models come Software-as-a-Service (SaaS)

L'adozione del modello *Software-as-a-Service (SaaS)* nella distribuzione dei FMs rappresenta una delle strategie più diffuse e consolidate nel settore. Questo approccio consente alle aziende sviluppatrici di offrire l'accesso ai loro modelli attraverso piattaforme cloud, senza che gli utenti debbano installare o gestire direttamente l'infrastruttura sottostante. La distribuzione del modello avviene prevalentemente tramite piattaforme proprietarie che forniscono un'interfaccia utente per l'accesso alle capacità del modello. In questa configurazione, gli utenti possono interagire direttamente con il modello attraverso applicazioni web o software dedicati, senza richiedere risorse computazionali locali.

La monetizzazione dei FMs distribuiti via SaaS può avvenire attraverso diverse modalità:

1. *Modello freemium*: viene offerto l'accesso a una versione gratuita del modello con funzionalità limitate, incentivando gli utenti a sottoscrivere piani a pagamento per ottenere prestazioni superiori o l'accesso a modelli più avanzati;

2. *Partnership*: alcune aziende stringono accordi con provider cloud o piattaforme di terze parti per integrare i loro modelli in ecosistemi più ampi, basandosi spesso su un modello di revenue-sharing. Un caso emblematico è la collaborazione tra OpenAI e Microsoft, in cui GPT-4 è reso disponibile tramite Azure, con una ripartizione dei ricavi generati dall'utilizzo del modello[34];

3. *Licenze esclusive*: gli sviluppatori forniscono servizi di personalizzazione, modificando i modelli per adattarli alle specifiche esigenze delle aziende che richiedono soluzioni su misura. Questo approccio è particolarmente diffuso nei settori regolamentati dove l'integrazione di dati proprietari risulta essenziale per ottimizzare le prestazioni al contesto specifico.

Box 2: OpenAI e Morgan Stanley

Nel 2023, Morgan Stanley annuncia un accordo con OpenAI rendendo la banca il primo istituto finanziario al mondo a ottenere un accesso esclusivo a una versione personalizzata di GPT-4, distinta dall'offerta standard disponibile pubblicamente. L'obiettivo dell'accordo era sviluppare Morgan Stanley Assistant, un chatbot addestrato sui dati proprietari della

[34]Questo accordo verrà approfondito nel box 6 del capitolo 4.

banca in grando di elaborare rapidamente le informazioni per i consulenti finanziari. Nonostante OpenAI offra la possibilità di utilizzare e personalizzare i suoi modelli tramite fine-tuning e API, la scelta strategica della partnership risiede soprattutto sul controllo dei dati. Infatti, la partnership ha consentito a Morgan Stanley di integrare il modello direttamente nei sistemi aziendali evitando l'invio dei dati sensibili finanziari alle API di OpenAI, ottenendo la potenza del modello GPT-4 senza rinunciare al controllo totale dei dati e dell'infrastruttura.

L'implementazione dei modelli di IAG all'interno dei servizi aziendali ha creato un vero e proprio mercato di intelligenza artificiale SaaS, il quale era pari a 71,54 miliardi di dollari nel 2024 e si stima che nel 2031 possa arrivare a 775,44 miliardi, con un tasso di crescita annuo composto (CAGR) del 38,28%[35].

3.3.2 Foudation Models come Application Programming Interfaces (API)

Le *API* rappresentano strumenti e protocolli che forniscono interfacce standardizzate, consentendo a un'applicazione di accedere ai dati o alle funzionalità di un'altra applicazione, piattaforma o servizio, senza la necessità di interagire direttamente con l'intero sistema sottostante. Dal punto di vista tecnico, un'API può essere considerata un ponte che collega applicazioni e servizi, fornendo un linguaggio comune per lo scambio di informazioni in modo efficiente e scalabile.

Nel contesto dell'IAG, le API costituiscono un canale per accedere alle funzionalità avanzate dei modelli, eliminando la necessità di addestrare o gestire direttamente l'infrastruttura necessaria. Attraverso l'utilizzo delle API, sviluppatori e aziende possono integrare le capacità generative dei modelli, quali la creazione di testi, immagini o codice, all'interno delle proprie applicazioni o piattaforme (un esempio è l'integrazione dei modelli di OpenAI da parte della società Notion discussa nel box 3).

Box 3: Notion AI

Notion AI rappresenta un esempio di come le API di modelli pre-addestrati possono essere integrati all'interno di servizi per offrire valore aggiuntivo ai

[35] AI created SaaS market size and share analysis - growth trends and forecast (2024-2031), Coherent Market Insights, 2024.

clienti, ottenere vantaggio competitivo e aumentare la monetizzazione del servizio. Nel 2023 Notion, una piattaforma SaaS per la gestione di progetti, note e documentazione, ha introdotto Notion AI, una funzionalità avanzata sviluppata utilizzando le API di OpenAI. Questo ha permesso alla piattaforma di includere strumenti di IAG per supportare gli utenti in diverse attività. Attraverso Notion AI, gli utenti possono generare contenuti testuali, riassumere documenti complessi e migliorare la qualità dei testi direttamente all'interno della piattaforma. Queste funzionalità si basano sull'elaborazione delle richieste inviate alle API di GPT, che restituiscono risposte in tempo reale. L'elaborazione delle richieste avviene tramite infrastrutture cloud, mentre l'interfaccia utente è stata personalizzata per adattarsi alle esigenze della piattaforma. Questo approccio ha consentito a Notion di utilizzare un modello avanzato senza sviluppare una propria tecnologia proprietaria, riducendo i costi di implementazione. Le funzionalità di IA sono disponibili come servizio premium per gli utenti con abbonamento a pagamento, rappresentando una strategia per aumentare il valore percepito del prodotto e differenziarlo nel mercato.

Questo approccio rende le tecnologie di IAG accessibili anche a quei soggetti che non dispongono delle risorse economiche e tecniche per sviluppare modelli proprietari, favorendo così una democratizzazione dell'accesso a queste tecnologie avanzate.

Il modello di business più comune per le API è il già citato pay-as-you-go. Si basa su un sistema di tariffazione che addebita agli utenti i costi in base all'effettivo utilizzo delle risorse o delle funzionalità offerte garantendo flessibilità e scalabilità.

Le API rappresentano uno strumento strategico non soltanto per la distribuzione dei modelli, ma anche per l'implementazione di strategie di differenziazione di prodotto basate sull'offerta di più modelli con capacità generative differenti. Ad esempio, la piattaforma Azure OpenAI, permette di implementare le API dei modelli sviluppati da OpenAI, come GPT-4o per la generazione del testo in cui la tariffazione è calcolata su token in input e output generato o DALL-E per la generazione delle immagini in cui la tariffazione è calcolata sulla risoluzione e quantità di immagini generate. Allo stesso modo, le organizzazione possono offrire diverse versioni di un modello che variano per prestazioni, velocità, accuratezza e capacità di elaborazione, consentendo così di soddisfare esigenze differenti di clienti e segmenti di mercato. Infatti, generalmente vengono offerti sia modelli ad alte prestazioni a prezzi più ele-

vati rivolti a grandi aziende con esigenze complesse, sia modelli più semplici e veloci a prezzo inferiore, rivolgendosi così a piccole e medie imprese o startup. In figura 3.5 vengono confrontati alcuni modelli dei principali sviluppatori sul mercato per evidenziare le dinamiche di differenziazione del prodotto.

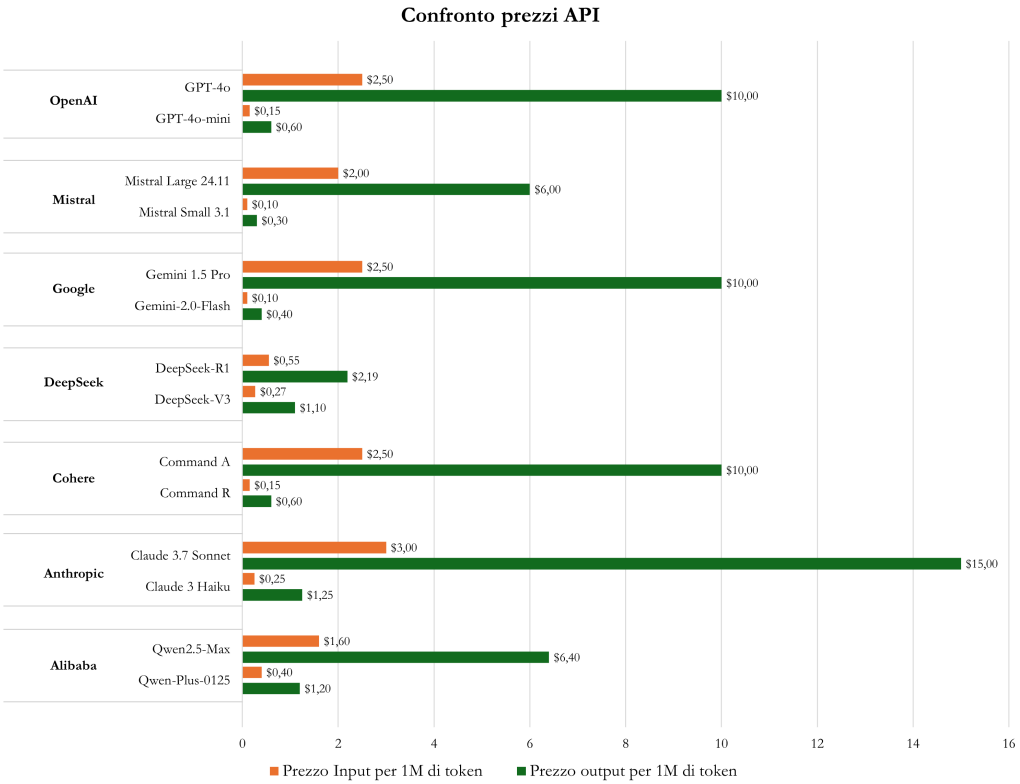

Figura 3.5: Confronto prezzi API di alcuni sviluppatori sul mercato.[36]

L'analisi del confronto tra i prezzi delle API dei principali sviluppatori evidenzia chiaramente strategie di segmentazione del mercato attraverso la differenziazione di prodotto. Alcuni modelli, quali GPT-4o (OpenAI), Gemini 1.5 Pro (Google), Claude 3.7 Sonnet (Anthropic) e Command A (Cohere), presentano tariffe significativamente più elevate che variano da 2$ a 3$ per milione di token in input e da 10$ a 15$ per milione di token in output, posizionandosi come i modelli più costosi per quanto riguarda i modelli per la generazione del testo. Questi rappresentano i modelli più avanzati degli sviluppatori risultando chiaramente indirizzati ad aziende e organizzazioni con esigenze sofisticate in termini di accuratezza e capacità di elaborazione linguistica, giustificando così un prezzo superiore. Parallelamente, vengono offerte versioni più economiche degli stessi modelli o modelli alternativi ottimizzati per usi

[36]Tuttavia, esistono anche casi limite caratterizzati da prezzi estremamente elevati, come nel caso dei modelli di OpenAI GPT-4.5 e o1-pro che raggiungono tariffe rispettivamente di 75$ e 150$ per milione di token in input e fino a 150$ e 600$ per milione di token di output, nonché il principale competitor per questa fascia di modelli Claude 3 Opus di Anthropic con tariffe di 15$ e 75$.

meno complessi. Ad esempio, GPT-4o-mini (OpenAI), Gemini-2.0-Flash (Google), Claude 3 Haiku (Anthropic) e Command R (Cohere), presentano costi drasticamente inferiori rendendoli accessibili a startup, PMI o sviluppatori indipendenti. Inoltre, le differenze dei prezzi tra gli sviluppatori già affermati e quelli più piccoli ed emergenti, evidenzia come questi ultimi (Mistral, DeepSeek e Alibaba) adottino strategie basate su un rapporto qualità-prezzo particolarmente competitivo, probabilmente con l'obiettivo di attrarre i consumatori e acquisire quota di mercato.

Le API, inoltre, giocano un ruolo fondamentale anche nella creazione di modelli specializzati, consentendo alle aziende di personalizzare i modelli per specifici casi d'uso. Un esempio rilevante di API per la creazione di modelli specializzati è rappresentato da Azure AI Foundry, una piattaforma che consente alle aziende di personalizzare modelli tramite un processo guidato e integrato. Il flusso di lavoro include la selezione di un modello di base, l'inserimento di dati di addestramento e, opzionalmente, di validazione, la configurazione dei parametri di addestramento e l'avvio del fine-tuning. Una volta completato il processo, il modello ottimizzato può essere distribuito immediatamente tramite API, consentendo un'integrazione rapida e scalabile in applicazioni aziendali.

3.4 I livelli prestazionali

Le differenze prestazionali tra le diverse tipologie di modelli costituiscono un elemento cruciale per comprendere le dinamiche competitive all'interno di questo mercato. In particolare, i modelli closed source proprietari di punta hanno dimostrato, nella maggior parte dei casi, di offrire prestazioni superiori rispetto a modelli open source o a soluzioni di dimensioni più ridotte.

Nelle sezioni precedenti sono stati analizzati i principali fattori che influenzano le prestazioni, come la qualità dei dati di addestramento e la complessità architetturale. In questa sezione, invece, l'attenzione sarà rivolta a studi quantitativi sulle differenze prestazionali, con un primo confronto tra i modelli generali closed source e open source e, successivamente, con un secondo confronto sulle capacità dei modelli specializzati rispetto a quelli generali in termini di prestazioni quando applicati a contesti specifici o domini verticali.

3.4.1 Modelli generali closed source vs open source

Come anticipato nel primo capitolo, i modelli open source hanno dimostrato prestazioni inferiori rispetto ai modelli closed source. Per analizzare più approfonditamente queste differenze, Huang et al.[37], (2024) hanno sviluppato un benchmark innovativo, l'*OlympicArena*, progettato per valutare sia Large Language Models che Multimodal Large Language Models. Il benchmark include un totale di 11.163 problemi, suddivisi in sette categorie disciplinari (matematica, fisica, chimica, biologia, geografia, astronomia e informatica) e comprende sia contenuti testuali sia multimediali, offrendo così una valutazione completa delle prestazioni dei modelli. L'OlympicArena adotta un approccio di valutazione basato su due metodologie complementari: la valutazione a livello di *risposta* e la valutazione a livello di *processo*.

La valutazione a livello di risposta si focalizza sull'accuratezza delle soluzioni fornite dai modelli. Per i problemi con risposte fisse (ad esempio, numeriche), vengono utilizzati metodi basati su regole che verificano la corrispondenza con la risposta corretta. Per problemi più complessi, invece, le risposte vengono valutate dal modello GPT-4V e sottoposte successivamente a revisione di esperti per garantire l'affidabilità. Questo metodo ha evidenziato che un campione di 100 risposte presenta circa l'80% di accordo tra la valutazione umana e quella del modello. Dato che circa solo il 5% dei problemi compresi nel benchmark richiedono questo tipo di valutazione, il tasso di errore può essere considerato circa dell'1%.

La valutazione a livello di processo, invece, è progettata per analizzare la correttezza dei passaggi di ragionamento e valutare in modo approfondito le capacità cognitive dei modelli. Per tale scopo, tramite GPT-4 vengono convertiti sia le soluzioni fornite che quelle date dal modello in un formato strutturato "passo dopo passo". Successivamente, questi passaggi vengono sottoposti alla valutazione di GPT-4V che valuta la correttezza di ogni passaggio assegnando 0 se il passaggio è errato e 1 se corretto. Infine, il punteggio per ogni problema viene ottenuto tramite la media dei punteggi di ogni passaggio. Per convalidare le valutazioni, viene utilizzato il giudizio umano su alcuni campioni, il cui risultato ha dimostrato un accordo modello-annotatore pari all'83%.

Questo approccio di valutazione, che combina analisi a livello di risposta e di processo, consente di ottenere una misurazione affidabile e approfondita delle ca-

[37]Huang, Z. et al. (2024) OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI. https://arxiv.org/abs/2406.12753.

pacità cognitive e delle prestazioni dei modelli. L'OlympicArena emerge così come uno dei benchmark più completi e rigorosi attualmente disponibili per confrontare le prestazioni di modelli open source e closed source, fornendo un quadro dettagliato delle loro capacità.

L'analisi delle prestazioni dei modelli condotta attraverso questo benchmark ha evidenziato significative differenze tra le due categorie. Lo studio iniziale ha mostrato come i modelli closed source abbiano raggiunto prestazioni nettamente superiori rispetto ai modelli open source, con un'accuratezza complessiva vicina al 40%, contro valori mediamente inferiori al 20% registrati dai modelli open source. Per fornire una visione più granulare e intuitiva delle differenze prestazionali, in uno studio successivo[38] e aggiornato con ulteriori modelli di punta, è stato introdotto un sistema di classificazione dei modelli simile a quello delle Olimpiadi. Vale a dire, le medaglie d'oro, d'argento e di bronzo vengono assegnate ai modelli che ottengono i primi tre punteggi in una determinata disciplina.

I risultati ottenuti dal sistema di ranking (tabella 3.4) mostrano chiaramente la predominanza dei modelli closed source nelle prime posizioni.

[38]Huang, Z., Wang, Z., Xia, S. and Liu, P. (2024) OlympicArena Medal ranks: Who is the most intelligent AI so far? https://arxiv.org/abs/2406.16772.

Tabella 3.4: Tabella delle medaglie dei principali modelli open e closed source.[39]

Pos	Modello	Accesso	Oro	Argento	Bronzo	Totale	Punteggio
1	GPT-4o	Closed	4	3	0	7	40,47
2	Claude-3.5-Sonnet	Closed	3	3	0	6	39,24
3	GPT-4V	Closed	0	1	1	2	33,17
4	Gemini-1.5-Pro	Closed	0	0	6	6	35,09
5	Claude-3-Sonnet	Closed	0	0	0	0	25,53
6	Qwen1.5-32B-Chat	Open	0	0	0	0	24,36
7	Qwen-VL-Max	Open	0	0	0	0	21,24
8	Gemini-Pro-Vision	Closed	0	0	0	0	21,02
9	LLaVA-NeXT-34B	Open	0	0	0	0	18,16
10	Yi-34B-Chat	Open	0	0	0	0	18,01
11	InternVL-Chat-V1.5	Open	0	0	0	0	17,39
12	InternLM2-Chat-20B	Open	0	0	0	0	17,33
13	Yi-VL-34B	Open	0	0	0	0	15,07
14	Qwen-VL-Chat	Open	0	0	0	0	7,34
15	Qwen-7B-Chat	Open	0	0	0	0	4,34

GPT-4o si distingue come il modello con le migliori prestazioni complessive, accumulando il maggior numero di medaglie d'oro e ottenendo un punteggio totale di 40,47, seguito da Claude-3.5-Sonnet e GPT-4V, anch'essi modelli closed source. Nelle prime cinque posizioni si trovano esclusivamente modelli proprietari, sottolineando il divario prestazionale tra le due categorie. In netto contrasto, i modelli open source, come Qwen1.5-32B-Chat e Qwen-VL-Max, iniziano ad apparire solo dalla sesta posizione in poi e non riescono a ottenere medaglie in alcuna disciplina. Questo risultato evidenzia un significativo divario in termini di capacità tra i modelli open source e quelli closed source, rafforzando l'impressione che i modelli proprietari mantengano un vantaggio competitivo significativo nel panorama attuale.

Analizzando un'applicazione più pratica, Irugalbandara et al. (2024)[40] utilizzano *myca.ai* come caso applicativo per valutare le differenze di performance tra l'utilizzo dell'API di un LLM proprietario e l'utilizzo di un SLM open-source self-hosted.

[39]Huang, Z., Wang, Z., Xia, S. and Liu, P. (2024) OlympicArena Medal ranks: Who is the most intelligent AI so far? https://arxiv.org/abs/2406.16772.

[40]Irugalbandara, C. et al. (2023) Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs in Production. https://arxiv.org/abs/2312.14972.

Myca.ai è un'applicazione di gestione personale in cui una delle funzionalità centrali è il *"Daily Pep Talk"*, un messaggio motivazionale personalizzato generato ogni mattina sulla base delle attività completate il giorno precedente, del piano per la giornata successiva e dei progressi verso gli obiettivi. Per valutare le performance dei modelli, vengono implementare tre tipologie di valutazione:

1. valutazione umana in modalità *blind test*, in cui ai valutatori vengono forniti il problema, il prompt originale e le risposte generate dai modelli, senza rivelare quale modello ha prodotto ciascuna risposta. Le risposte sono presentate in ordine casuale per ridurre i bias, e vengono valutate per qualità e pertinenza rispetto al problema;

2. per integrare e ridurre il carico della valutazione umana, vengono utilizzate tecniche di valutazione automatizzata basate su GPT-4. Questi metodi includono: (i) GPT-Scorer in cui GPT-4 valuta le risposte generate dai modelli SLM, basandosi su qualità e rilevanza, seguendo criteri simili a quelli impiegati nella valutazione umana; (ii) GPT-Comparer in cui GPT-4 confronta le risposte degli SLM con una risposta di riferimento generata tramite l'API di OpenAI, fornendo punteggi e spiegazioni sul ragionamento alla base delle sue decisioni; (iii) GPT-4 Multi-Choice Selector, in cui GPT-4 seleziona la migliore risposta tra quelle generate da più SLM per uno stesso input, valutando qualità, accuratezza e rilevanza;

3. Infine, viene implementata una valutazione basata sulla similarità semantica per quantificare quanto le risposte generate dagli SLM siano vicine a una risposta di riferimento.

Confrontando i risultati ottenuti dall'API di Open AI e di 29 SLM open source di Huggin Face, si ottiene che: utilizzando la valutazione umana per valutare la qualità e la rilevanza delle risposte, GPT-4 ottiene il punteggio medio più alto (5,9) ma diversi SLM hanno raggiunto punteggi comparabili, dimostrando che possono generare risposte di qualità competitiva. Tra questi, Vicuna:7b-q3 ha ottenuto un punteggio medio di 5,6, Starling-LM:7b-q4 di 5,1 e Neural-chat:7b-q2 di 5,3. Un'osservazione chiave dello studio riguarda i modelli quantizzati, che, pur essendo più leggeri in termini di dimensione, hanno mostrato prestazioni comparabili o in alcuni casi superiori rispetto ai modelli base. Questo risultato evidenzia il potenziale degli SLM

quantizzati per applicazioni pratiche, dove dimensioni ridotte e maggiore efficienza computazionale sono cruciali.

Dal punto di vista della latenza, i risultati mostrano che, mentre GPT-4 registra la latenza media per richiesta più bassa, molti SLM, come Mistral:7b-instruct, Orca-mini:3b e StableLM-zephyr:3b, si avvicinano a meno di un secondo di differenza rispetto alla latenza di GPT-4. Nel confronto sulla latenza per token, diversi SLM, tra cui StableLM-zephyr:3b, Mistral-7b e StarlingLM, si sono dimostrati più rapidi di GPT-4 nella generazione di singoli token, evidenziando un potenziale vantaggio in contesti applicativi che richiedono tempi di elaborazione ridotti.

Questi risultati suggeriscono che gli SLM self-hosted possono offrire prestazioni competitive o addirittura migliori rispetto all'API di GPT-4 per quanto riguarda la latenza, sebbene con una variabilità maggiore nella lunghezza delle risposte e nei tempi complessivi. Inoltre, la latenza delle richieste dell'API di OpenAI varia notevolmente nel corso delle 24 ore, oscillando tra 3,4 secondi e 8,6 secondi per richiesta. Al contrario, i modelli self-hosted, come StarlingLM e Orca-mini-3b, hanno dimostrato una latenza molto più costante durante l'arco della giornata, con una distribuzione più stabile delle prestazioni sia tra le richieste all'interno della stessa ora sia nell'intero periodo di 24 ore.

Questi risultati evidenziano che, sebbene l'API di OpenAI offra prestazioni di alta qualità e consistenza nelle risposte, i modelli SLM open-source rappresentano un'alternativa valida e competitiva. In particolare, i modelli quantizzati si dimostrano adatti per scenari applicativi dove le risorse computazionali sono limitate. Tuttavia, la scelta tra API e SLM dipende da diversi fattori, tra cui la necessità di prevedibilità delle prestazioni, i requisiti di latenza e i costi operativi. In contesti dove la stabilità e l'efficienza a lungo termine sono cruciali, gli SLM self-hosted possono rappresentare un'opzione più affidabile e scalabile.

3.4.2 Modelli generali vs modelli specializzati

Le differenze prestazionali tra modelli linguistici generali e specializzati rappresentano un elemento fondamentale per valutare l'efficacia dell'intelligenza artificiale in domini complessi. Nel campo medico, ad esempio, Singhal et al. (2023)[41] hanno sviluppato Med-PaLM 2, un modello specializzato basato sul FM PaLM 2. Questo modello è stato testato su nove dataset contenenti domande a risposta multipla,

[41] Singhal, K. et al. (2023) Towards Expert-Level Medical Question Answering with Large Language Models. https://arxiv.org/abs/2305.09617.

come MMLU-Anatomy, focalizzato su tematiche di anatomia, e MMLU-Medical Genetics, dedicato a domande sulla genetica medica. Inoltre, è stato valutato su quattro dataset con domande a risposta estesa, tra cui MultiMedQA, che include quesiti complessi tipici del settore sanitario.

Per la valutazione delle domande a risposta multipla, sono state utilizzate tecniche avanzate di prompting per migliorare il ragionamento e l'accuratezza delle risposte. Tra queste, il (i) *few-shot prompting*, che consiste nel fornire esempi di domande e risposte prima della domanda target, il (ii) *Chain-of-Thought (CoT)*, che aggiunge spiegazioni passo-passo agli esempi forniti, consentendo al modello di affrontare problemi complessi condizionandosi sui propri passaggi intermedi, la (iii) *self-consistency*, che prevede la generazione di risposte multiple per una stessa domanda, selezionando quella più frequente o coerente tramite un voto di maggioranza e, infine, (iv) l'*ensemble refinement (ER)*, una strategia che combina e raffina risposte precedentemente generate per produrre una versione finale più accurata e coerente, integrando i vantaggi del CoT e della self-consistency.

Per la valutazione delle risposte in forma estesa, Med-PaLM 2 è stato sottoposto a metodi di valutazione umana. Le risposte sono state generate utilizzando specifici prompt progettati per garantire coerenza e sono state successivamente sottoposte a diversi medici chiamati ad analizzare ogni risposta in modo indipendente, senza sapere il modello di appartenenza delle risposte. Le valutazioni si basano su diversi criteri, ad esempio l'accuratezza della risposta, l'omissione di contenuti importanti o il rischio di bias.

Dai risultati emerge che Med-PaLM 2 ha superato GPT-4 in diversi dataset, dimostrando i vantaggi di un modello specializzato rispetto a un modello generalista nel contesto dell'applicazione in uno specifico dominio. In particolare, su MedQA, un benchmark che valuta la conoscenza medica generale in domande nello stile dell'esame di licenza medica statunitense, Med-PaLM 2 ha raggiunto un'accuratezza dell'86,5%, rispetto all'86,1% di GPT-4. Un altro risultato significativo si osserva su PubMedQA, che richiede di rispondere a domande basate su abstract scientifici: Med-PaLM 2 ha ottenuto un'accuratezza dell'81,8%, superando GPT-4 (80,4%). Anche sul dataset MMLU Professional medicine ha ottenuto un punteggio superiore, pari a 95,2% contro il 93,8% di GPT-4. Tuttavia, in determinati domini il modello ha ottenuto risultati inferiori ma comparabili, ad esempio in MedMCQA, ha ottenuto il 72,3% di accuratezza, un risultato pressoché identico al 72,4% di GPT-4. Oppure,

nel dominio della biologia (MMLU College Biology), GPT-4 ha leggermente superato MedPaLM-2 con un' accuratezza del 97,2% rispetto al 95,8%.

Risultati simili sono stati ottenuti anche da Chen et al. (2024)[42], i quali hanno sviluppato PharmaGPT, un modello specializzato nel dominio biofarmaceutico e chimico. Questo modello è stato testato sul NAPLEX (North American Pharmacist Licensure Examination), un esame standardizzato progettato per valutare le competenze fondamentali richieste per la pratica farmacologica. Il test comprende domande a scelta multipla che coprono tre aree principali: la gestione della terapia farmacologica (ad esempio, identificare trattamenti ottimali per condizioni mediche specifiche), la preparazione e distribuzione dei farmaci, e la gestione della pratica farmacologica. Su domande tratte da database ufficiali di simulazione del NAPLEX, che includono sia domande cliniche complesse, sia scenari che richiedono ragionamento critico, il modello ha ottenuto una precisione compresa tra il 70% e l'80% in tutte le sezioni dell'esame, superando sia GPT-3.5 Turbo che GPT-4 in diversi domini. Questa valutazione è stata condotta valutando l'accuratezza e l'allineamento delle risposte del modello con quelle corrette fornite dagli esperti del settore.

Tuttavia, è importante sottolineare che non sempre i modelli specializzati hanno performance migliori rispetto ai generali. Uno studio che lo evidenzia è quello di Li et al. (2023)[43], che ha come obiettivo quello di valutare le prestazioni tra modelli generali come ChatGPT-4 nel risolvere compiti nel dominio finanziario rispetto a modelli specializzati come BloombergGPT. Lo studio si propone di valutare i modelli su cinque categorie di compiti di difficoltà e conoscenza del dominio finanziario crescente: (i) Sentiment Analysis, (ii) Classification, (iii) Named Entity Recognition (NER), (iv) Relation Extraction (RE) e (v) Question Answering (QA).

1. I risultati mostrano che nei task più semplici come la Sentiment Analysis e la Classification, GPT-4 ha dimostrato prestazioni superiori rispetto ai modelli specializzati. Ad esempio, nel dataset FiQA, GPT-4 ha ottenuto un punteggio sull'indice di valutazione F1 ponderato pari a 88,11% rispetto al 75,07% di BloombergGPT. Nei task di classificazione testuale finanziaria, GPT-4 ha ottenuto un punteggio dell'86%, superando BloombergGPT (82,2%). Tuttavia, il modello specializzato di BERT ha raggiunto la performance migliore con un punteggio del 95,36%, superiore di 9 punti percentuali rispetto a GPT-4;

[42]Chen, L. et al. (2024) PharmaGPT: Domain-Specific Large Language Models for Bio-Pharmaceutical and Chemistry. https://arxiv.org/abs/2406.18045.

[43]Li, X. et al. (2023) Are ChatGPT and GPT-4 General-Purpose solvers for financial text analytics? A study on several typical tasks. https://arxiv.org/abs/2305.05862.

2. Nei task strutturati come il NER e RE, i modelli specializzati hanno avuto prestazioni migliori rispetto ai generali. Nei task NER, GPT-4 ha mostrato risultati limitati (56,71% in modalità few-shot) mentre BloombergGPT ha raggiunto un punteggio F1 di 60,82%. La performance migliore è stata ottenuta dal modello CRF, ottimizzato su dati specifici (FIN5), con un punteggio di 82,7%. È importante sottolineare che CRF ha mostrato una forte sensibilità al dominio dei dati: quando addestrato su un dataset fuori dominio (CoNLL), il suo punteggio è sceso drasticamente a 17,2%, evidenziando l'importanza della pertinenza dei dati di addestramento. Allo stesso modo, nei task RE, il modello specializzato Luke-base ha ottenuto un punteggio Macro F1 pari al 56,3%, maggiore di circa 10 punti percentuali rispetto a GPT-4 (46,87%);

3. Nel task di question answering (QA) finanziario, che rappresenta una delle applicazioni più complesse per i modelli linguistici, GPT-4 ha dimostrato prestazioni significativamente superiori rispetto ad altri modelli generali e specializzati. L'attività di QA finanziario richiede ai modelli non solo di comprendere la conoscenza del dominio, ma anche di effettuare operazioni numeriche e ragionamenti logici complessi, come il calcolo dei tassi di crescita del profitto da tabelle finanziarie. Lo studio ha utilizzato due dataset specifici: FinQA, focalizzato su singole domande e risposte, e ConvFinQA, che aggiunge un elemento di conversazione e ragionamento multi-turno. Nei test su FinQA, GPT-4 ha ottenuto un'accuratezza del 68,79% in modalità zero-shot e ha raggiungo un valore massimo di 78,03% utilizzando la strategia Chain-of-Thought (CoT). Su ConvFinQA, GPT-4 ha confermato la sua superiorità con un'accuratezza del 76,48%, dimostrando notevoli capacità di ragionamento numerico, significativamente superiori a quelle di BloombergGPT (43,41%). Un risultato particolarmente sorprendente è che GPT-4, con la strategia CoT, ha superato FinQANet, un modello specializzato per il task, che ha raggiunto un'accuratezza del 68,9%. Ciò evidenzia come la scala dei parametri e le tecniche avanzate di pre-addestramento di GPT-4 possano colmare il divario con i modelli specializzati anche in compiti complessi.

Le differenze prestazionali tra modelli generali e specializzati non sono scontate e dipendono in modo significativo dal tipo di compito e dal dominio applicativo. Sebbene modelli specializzati come Med-PaLM 2 o PharmaGPT tendano a superare quelli generali in contesti altamente specifici, i risultati dello studio nel dominio finanziario

hanno dimostrato che modelli avanzati come GPT-4 si distinguono nei task meno strutturati, che richiedono capacità di ragionamento trasversale. Al contrario, nei compiti più strutturati, i modelli specializzati hanno mantenuto un chiaro vantaggio, grazie alla loro ottimizzazione per esigenze specifiche. Queste differenze evidenziano che l'efficacia di un modello non dipende esclusivamente dalla sua specializzazione, ma anche dalla natura e complessità del task, nonché dalle competenze richieste.

Capitolo 4

Gli assetti di mercato

Nel capitolo precedente sono stati analizzati i fattori necessari allo sviluppo di un modello di IAG, evidenziando i fattori di costo riguardanti i diversi input, le modalità di distribuzione e le differenze tra i livelli prestazionali. Nel corso di questo capitolo verranno analizzate le dinamiche competitive all'interno di ogni segmento della catena del valore per comprendere eventuali dinamiche di concentrazione e quali siano le strategie che portano un'azienda ad assumere una posizione di vantaggio nel mercato.

4.1 Dinamiche competitive lungo la catena del valore

4.1.1 Il mercato dei chip e del cloud

Come anticipato, il principale hardware necessario per lo sviluppo di modelli di IAG è rappresentato dai chip accelleratori. Attualmente, il leader mondiale nello sviluppo di GPU per applicazioni di IA è NVIDIA, la cui posizione dominante è sostenuta dalla sua architettura chiusa (box 4), dall'efficienza dei suoi acceleratori e dalla diffusione dell'architettura CUDA. Infatti, la figura 4.1 mostra che ad ottobre 2024, NVIDIA ha registrato un fatturato record di 35,08 miliardi di dollari, con una crescita significativa rispetto ai trimestri precedenti. Il segmento data center, che include le tecnologie per l'intelligenza artificiale e il calcolo accelerato, ha generato 30,77 miliardi di dollari, più del doppio rispetto allo stesso trimestre dell'anno precedente. Questa crescita esponenziale è attribuita principalmente alla crescita dell'IA e alla domanda di chip specializzati.

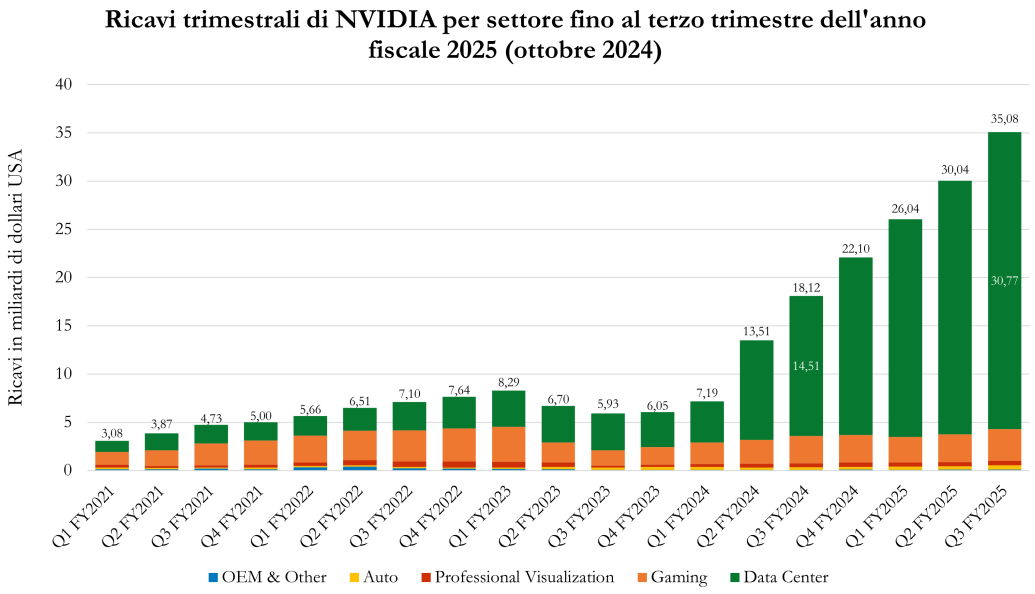

Figura 4.1: Ricavi trimestrali di NVIDIA.[44]

Box 4: NVLink e UALink

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

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

Altri produttori, come AMD e Intel, stanno cercando di acquisire quote di mercato offrendo chip alternativi; tuttavia, le loro soluzioni attuali non riescono a competere con le prestazioni dei chip acceleratori prodotti da NVIDIA. Per ridurre le dipendenze dall'azienda statunitense, i principali fornitori di infrastrutture hardware cloud per lo sviluppo di modelli, ovvero Microsoft, Google e Amazon, stanno sviluppando internamente chip proprietari, trasformandosi così da clienti principali[45] a potenziali concorrenti. Un esempio sono gli acceleratori TPU di Google che rappresentano una soluzione hardware progettata specificamente per ottimizzare le operazioni di addestramento e inferenza nei modelli di intelligenza artificiale. Queste possono essere sfruttate tramite la piattaforma Cloud TPU di Google che permette di utilizzare i suoi acceleratori in base ad una tariffa a consumo misurata in ora chip. Anche Amazon Web Services ha rilasciato tramite la sua piattaforma cloud la possibilità di utilizzare i chip proprietari Trainium che permettono di ridurre i costi di addestramento dei modelli ottenendo allo stesso tempo elevate prestazioni. Tuttavia, questa trasformazione, sebbene in corso, non sembra rappresentare una minaccia immediata per NVIDIA. Infatti, oltre ad essere la principale fornitrice di GPU per l'addestramento dei modelli, è anche la principale cliente della *Taiwan Semiconductor Manufacturing Company (TSMC)*, la più grande produttrice di semiconduttori al mondo. Il ruolo di TSMC è cruciale perché fornisce a NVIDIA le capacità produttive più avanzate, tra cui l'uso della tecnologia *Chip-on-Wafer-on-Substrate (CoWoS)*, un sistema di packaging che migliora l'integrazione tra i componenti dei chip, aumentando l'efficienza e la velocità di elaborazione. La TSMC prevede di raddoppiare la sua attuale capacità produttiva a seguito della crescente domanda da parte di NVIDIA, Microsoft, Amazon e Alphabet, di cui più del 50%[46] verrebbe occupata esclusivamente da NVIDIA. È quindi probabile che questo controllo sulla catena di fornitura permetterà a NVIDIA di rimanere nella sua posizione dominante nel lungo periodo. A suggerirlo è anche l'adozione della nuova *GPU Blackwell* da parte di Microsoft, Google e Amazon, che hanno annunciato l'integrazione di questi chip nelle loro infrastrutture cloud per migliorare le capacità computazionali dedicate all'intelligenza artificiale. Questo ulteriore investimento da parte dei principali hyperscaler dimostra come, nonostante i tentativi di sviluppare soluzioni proprietarie, la dipen-

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

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

denza da NVIDIA rimanga ancora un elemento centrale nell'evoluzione del mercato dei chip per l'IA.

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

4.1.2 Il mercato dei dati

Le stime riportate da recenti studi evidenziano un problema strutturale che potrebbe limitare il futuro sviluppo dei modelli di IAG. Secondo queste analisi, la disponibilità di dati generati dall'uomo è destinata ad esaurirsi entro pochi anni. Tuttavia, il momento esatto in cui questi dati verranno utilizzati appieno dipende in larga misura dalla politica di scalabilità adottata durante l'addestramento dei modelli. Se i modelli venissero addestrati in modo ottimale dal punto di vista computazionale, ci sarebbero dati sufficienti per addestrare modelli fino al 2028. Tuttavia, le pratiche di addestramento recenti, come nel caso di Llama 3-70B, mostrano una tendenza al "sovra addestramento", in cui i modelli utilizzano meno parametri e più dati per ottimizzare l'efficienza computazionale durante l'inferenza. In scenari di sovra addestramento più modesti, ad esempio di 5 volte, si stima che lo stock di dati generati dall'uomo sarà completamente utilizzato entro il 2027. Invece, se si adotta una politica più aggressiva, come il sovra addestramento di 100 volte, lo stock di dati potrebbe esaurirsi già entro il 2025[47].

[47]Villalobos, P. (2024) 'Will we run out of data? Limits of LLM scaling based on Human-Generated Data,' Epoch AI, 6 Giugno. https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

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

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

4.1.3 Il mercato dei Foundation Models: numerosità dei modelli e concentrazione

Per analizzare il mercato dei FMs e offrire una panoramica completa, in questa sezione vengono presentati diversi grafici costruiti a partire dal database *Ecosystem Graphs del CRFM (Center for Research on Foundation Models)* di Stanford che, introdotto nel 2023, traccia l'ecosistema dei FMs, includendo modelli, applicazioni e dataset. La figura 4.2 mostra il numero di modelli rilasciati globalmente dal 2019 al 2024 suddivisi in base alle modalità di accesso (closed, limited e open) anticipate nel secondo capitolo di questo elaborato. L'analisi dell'evoluzione del numero di modelli evidenzia in primo luogo una crescita esponenziale nella loro diffusione. In particolare, il numero totale è aumentato da pochi esemplari negli anni 2019 e 2020 fino a un massimo di 184 modelli totali nel 2023. La riduzione a 133 modelli nel 2024 potrebbe suggerire che, dopo una rapida espansione iniziale, il mercato stia entrando in una fase di consolidamento. Dalla suddivisione tra le modalità di accesso, emerge invece che, nonostante i modelli open source offrano generalmente prestazioni inferiori rispetto ai modelli closed[48], lo sviluppo di questi ha subito una crescita

[48]Vedere sezione 2.3.1 del capitolo precedente

significativamente maggiore, raggiungendo nel 2023 un numero quasi quattro volte superiore a quello dell'anno precedente. Anche nel 2024 il numero dei modelli open source rappresenta la percentuale maggiore, pari a circa il 72% del totale. Questo mette in luce la tendenza nel mercato all'innovazione e alla ricerca, in quanto anche se i modelli open non offrano un elevato controllo e protezione, consentono ad un numero ampio di attori di testare, personalizzare e innovare, contribuendo alla crescita complessiva del settore.

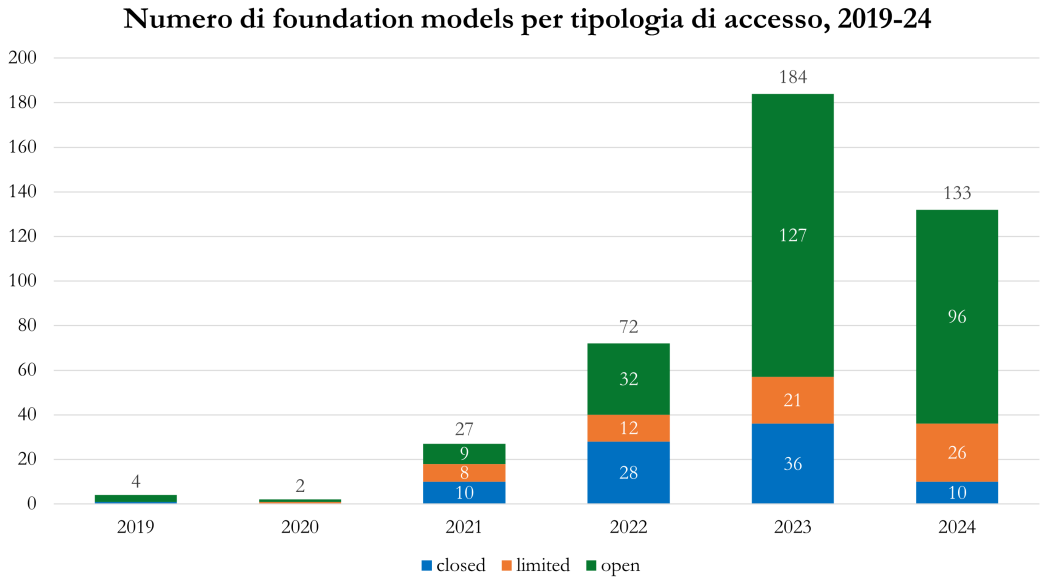

Figura 4.2: Numero di FMs per tipologia di accesso, 2019-24.[49]

Oltre alla suddivisione per tipologia di accesso, un ulteriore aspetto da analizzare è il numero di modelli rilasciati dai diversi players nel mercato. I successivi due grafici (figura 4.3 e 4.4) mostrano, rispettivamente, il numero di modelli rilasciati dai primi 15 player più attivi nel periodo 2019-2024 e il numero totale rilasciato, sempre dai 15 player più attivi, solo nel 2024. Quest'analisi permette di evidenziare, oltre che i principali attori, anche eventuali tendenze di concentrazione del mercato e l'evoluzione dei diversi sviluppatori nel tempo.

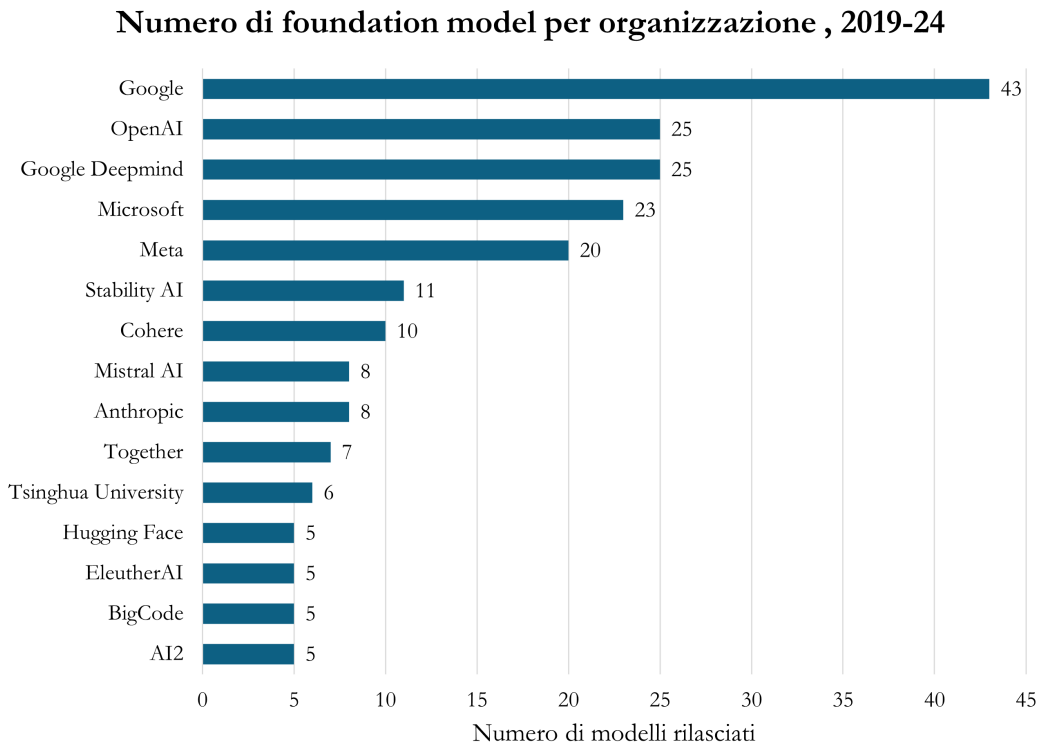

Figura 4.3: Numero di FMs dei primi 15 player più attivi rilasciati nel periodo 2019-2024.[50]

L'analisi della distribuzione dei modelli tra il 2019 e il 2024 evidenzia che la maggior partecipazione all'interno del mercato è data da 5 principali società: Google (43), OpenAI (25), Google DeepMind (25), Microsoft (23) e Meta (20). Complessivamente, questi attori hanno rilasciato 136[51] modelli nell'arco di questi 6 anni. Questo dato assume particolare importanza se confrontato con il numero dei modelli totali a livello globale e il numero di sviluppatori a questi associati. Infatti, il numero totale di attori che hanno sviluppato almeno un FM nei 6 anni passati ammonta a circa 610 e comprende aziende private, istituzioni accademiche, organizzazioni no-profit, enti governativi e collaborazioni scientifiche internazionali. I cinque precedenti attori rappresentano quindi circa lo 0,08% del panorama mondiale e solamente i loro modelli costituisco quasi un terzo (32%) di tutti quelli sviluppati globalmente.

Questi dati suggeriscono che storicamente il settore è stato fortemente dominato da un ristretto gruppo di aziende tecnologiche, le quali, probabilmente grazie al loro accesso privilegiato a risorse computazionali, dati e competenze altamente specializzate, riescono a mantenere una posizione predominante nello sviluppo dell'IAG.

[50] Ecosystem graphs for foundation models, CRM Stanford.
[51] Da questo dato sono esclusi i modelli rilasciati in collaborazione con altre società o con istituzioni accademiche.

Questa elevata concentrazione del mercato solleva interrogativi sulla competitività del settore, sulle barriere all'ingresso per nuovi attori e sul possibile impatto delle politiche regolatorie nel riequilibrare la distribuzione del potere tra i vari sviluppatori di modelli. Tuttavia, l'analisi della distribuzione dei FMs sviluppati nel 2024 rispetto all'intero periodo 2019-2024 suggerisce una diversificazione tra gli attori coinvolti. Se nel lungo periodo il mercato è stato dominato dalle big tech, con Google, OpenAI, Google DeepMind, Microsoft e Meta responsabili di oltre un terzo dei modelli sviluppati globalmente, il 2024 mostra un panorama più distribuito e competitivo.

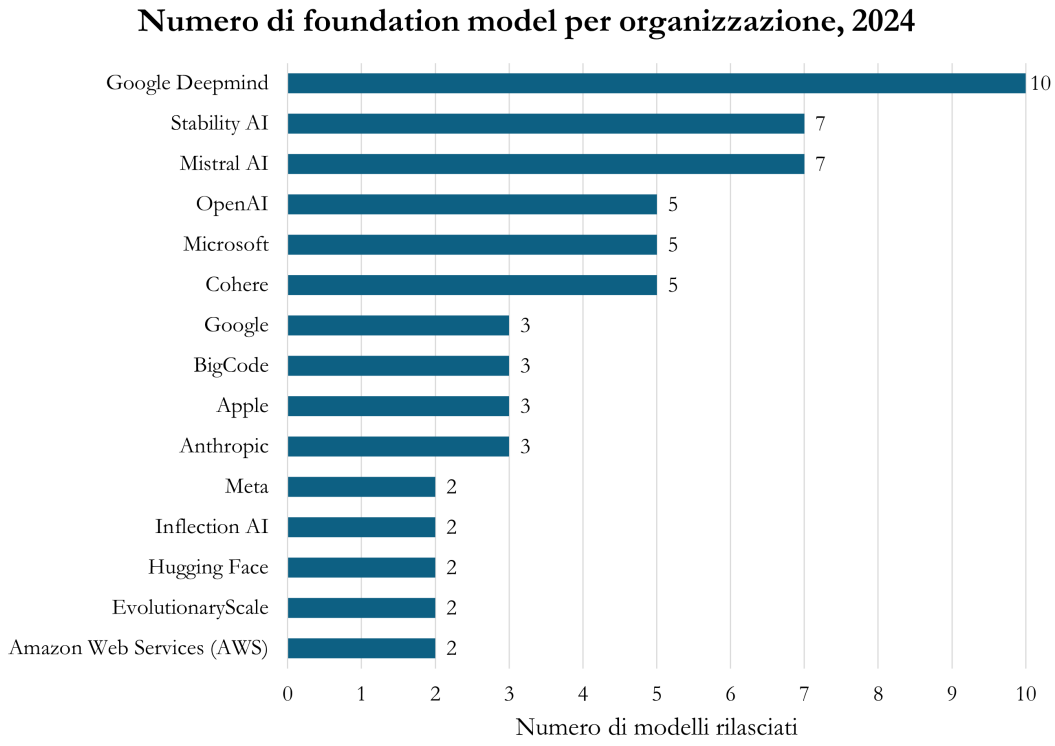

Figura 4.4: Numero di FMs dei primi 15 player più attivi rilasciati nel 2024.[52]

Google DeepMind emerge come il principale sviluppatore, con 10 modelli rilasciati, superando la casa madre Google, che nel lungo periodo aveva accumulato 43 modelli, ma che nell'ultimo anno ha sviluppato solo 3 modelli. Parallelamente, Stability AI e Mistral AI si posizionano tra i leader del 2024 con 7 modelli ciascuna, evidenziando l'ingresso di nuovi attori che potenzialmente potrebbero competere con le grandi aziende tecnologiche. Inoltre, è importante evidenziare come questi due attori, rispettivamente con sede nel Regno Unito e in Francia, abbiano sviluppato un totale di 8 e 11 modelli nell'arco dei 6 anni di cui la maggior parte solamente del 2024. Questo potrebbe suggerire una partecipazione più attiva da parte di nazioni

[52]Ecosystem graphs for foundation models, CRM Stanford.

differenti in un mercato fino a questo momento dominato esclusivamente da aziende statunitensi e cinesi.

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

Per la seguente analisi vengono confrontate le piattaforme di 8 chatbot di IAG: ChatGPT (OpenAI), Le Chat (MistralAI), Gemini (Google), Meta AI (Meta), Claude (Anthropic), Copilot (Microsoft), Qwen2.5-Max (Alibaba) e DeepSeek (DeepSeek AI). La figura 4.5 mostra la percentuale di traffico a livello globale delle piattaforme[53] sul totale del traffico generato da queste ultime nei mesi di dicembre 2024, gennaio 2025 e la prima settimana di febbraio 2025.

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

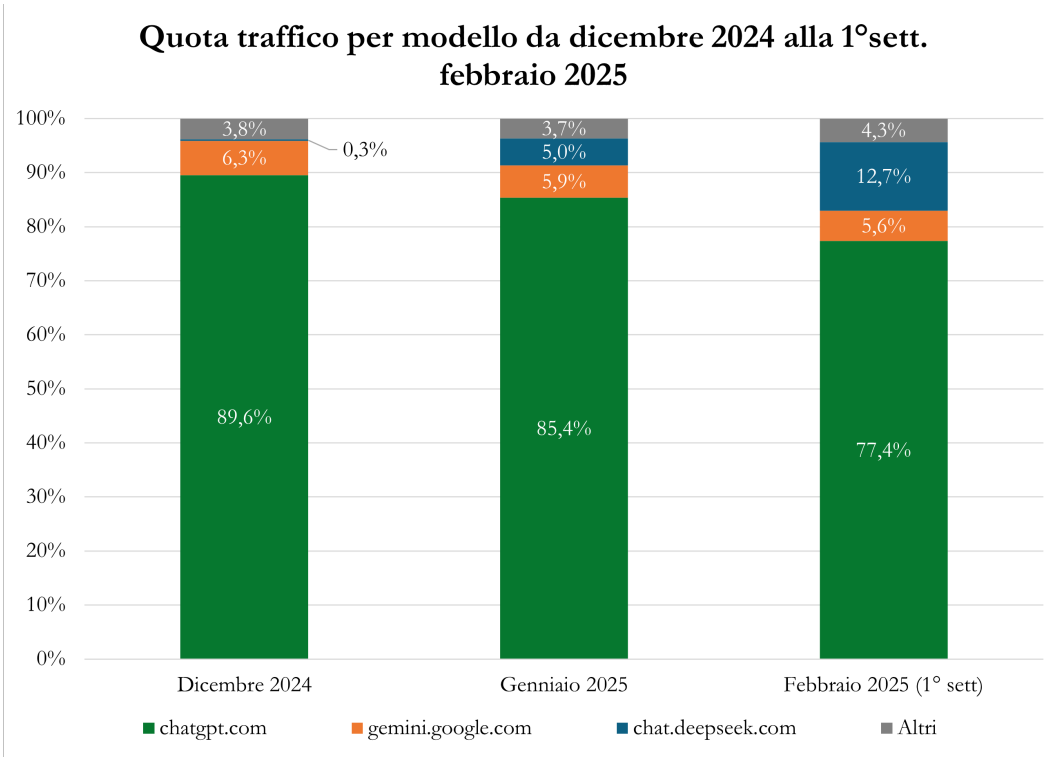

Figura 4.5: Quota traffico generato dalle piattaforme di alcuni FM.[54]

L'analisi combinata dei dati sul traffico web delle principali piattaforme che permettono lo sfruttamento dei FMs e del numero di modelli sviluppati dalle aziende tra il 2019 e il 2024 evidenzia dinamiche di mercato particolarmente rilevanti. In primo luogo, ChatGPT di OpenAI continua a dominare il settore, con una quota di traffico che, nonostante un lieve calo dal 89,56% di dicembre 2024 al 77,36% nella prima settimana di febbraio 2025, rimane nettamente superiore a quella di qualsiasi altro competitor. Questo dato è particolarmente significativo se confrontato con il numero di modelli sviluppati dalle diverse aziende: Google, leader per numero di FMs rilasciati negli ultimi sei anni (43), non riesce a tradurre questa superiorità in un dominio effettivo sul mercato, con il suo modello di punta Gemini che raccoglie appena il 5,62% del traffico nella prima settimana di febbraio 2025, confermando una posizione marginale rispetto a ChatGPT.

Un altro elemento di particolare rilievo è la rapida ascesa di DeepSeek AI, la cui piattaforma chat.deepseek.com, che permette l'utilizzo dei suoi modelli open source, è passata dallo 0,26% di dicembre 2024 al 12,7% nella prima settimana di febbraio 2025, registrando una crescita esponenziale in pochissimo tempo. Questo incremento rappresenta un'anomalia nel mercato, suggerendo che DeepSeek abbia trovato

[54] Similarweb

una strategia altamente efficace di penetrazione. Nel box 5 viene approfondita la strategia di DeepSeek e le motivazioni per il quale il rilascio dei suoi modelli abbia drasticamente cambiato le dinamiche di mercato.

Box 5: DeepSeek AI

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

Tuttavia, il costo così ridotto per lo sviluppo dei suoi modelli ha sollevato alcuni dubbi sulla veridicità delle dichiarazioni della startup: ad esempio il CEO di Scale AI, Alexandr Wang, durante un'intervista ha sostenuto che in realtà DeepSeek possieda 50.000 chip GPU H100 che non possono essere dichiarate proprio a causa delle restrizione sulle esportazioni. Non solo, la stessa OpenAI ha accusato la startup cinese su una possibile violazione della proprietà intellettuale, in quanto secondo una speculazione DeepSeek avrebbe usato i modelli di OpenAI per addestrare i suoi modelli tramite tecniche chiamate di

distillazione, il che violerebbe i termini d'uso di OpenAI.
Nonostante questo, DeepSeek ha completamente sconvolto i mercati, causando la più grande perdita azionaria giornaliera nella storia: NVIDIA a seguito del rilascio dei modelli di DeepSeek, ha perso 590 miliardi di dollari di capitalizzazione (calo del 17%). Questo calo potrebbe rispecchiare la sfiducia degli investitori in un mercato che fino a quel momento si pensava potesse essere dominato esclusivamente da chi possedeva un numero elevato di modelli di GPU NVIDIA di punta.
L'ingresso di DeepSeek ha quindi segnato il panorama globale, in quanto la strategia open source basata sull'efficienza, in netto contrasto con quella adottata dalle aziende statunitensi, ha sollevato sia interrogativi sulle modalità di sviluppo sia sulle attuali regolamentazioni, dando il via a una corsa allo sviluppo a livello globale che verrà analizzata nel seguito di questo elaborato.

L'analisi congiunta di questi dati suggerisce dunque che il numero di modelli sviluppati non rappresenta, di per sé, un indicatore sufficiente per determinare il successo di mercato o la concentrazione dello stesso. Infatti, tale numero può essere fuorviante anche per ulteriori ragioni: spesso, le organizzazioni sviluppano un singolo modello base accompagnato da numerosi altri modelli derivati, progettati per essere più economici o più veloci rispetto al modello principale[55]. Questi modelli rispondono prevalentemente a logiche di differenziazione del prodotto piuttosto che riflettere una reale frammentazione competitiva del mercato. Tuttavia, se da un lato OpenAI, con un numero inferiore di modelli rispetto a Google, ha consolidato la propria leadership attraverso un singolo prodotto dominante, dall'altro la rapida crescita di DeepSeek dimostra che nuovi entranti possono scalare rapidamente se dotati di un prodotto competitivo e di una strategia efficace di distribuzione. Questo scenario potrebbe prefigurare una maggiore frammentazione del mercato nel prossimo futuro, con il potenziale emergere di nuovi concorrenti in grado di sfidare il predominio consolidato di OpenAI, derivato in parte dalle sue strategie di integrazione verticale.

4.2 Partnerships e dinamiche di integrazione verticale

L'integrazione verticale rappresenta una delle strategie più rilevanti nel mercato dell'IAG, permettendo alle aziende di acquisire un maggiore controllo sulla catena del

[55] Come anticipato nella sezione 3.3.2.

valore e di ottenere vantaggi competitivi significativi. In questo contesto, l'integrazione può avvenire sia a monte che a valle. L'integrazione a monte fa riferimento al controllo delle infrastrutture computazionali nonché sui dati necessari all'addestramento del modello. In quella a valle, invece, l'integrazione avviene con le piattaforme di distribuzione dei servizi basati sui FMs. Queste tipologie di integrazioni, che possono avvenire singolarmente o in modo congiunto, forniscono vantaggi competitivi significativi per le aziende integrate ma creano potenziali barriere di ingresso per i nuovi partecipanti. Partendo dalle integrazioni a monte, di seguito vengono esaminate due casistiche:

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

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

Box 6: Accordo tra OpenAI e Microsoft

La collaborazione tra Microsoft e OpenAI ha avuto inizio nel 2019, quando Microsoft ha effettuato un investimento iniziale di un miliardo di dollari in OpenAI, con l'obiettivo di accelerare le innovazioni nel campo dell'intelligenza artificiale e di democratizzare l'accesso a queste tecnologie avanzate. Questa partnership è stata ulteriormente rafforzata nel gennaio 2023, con un investimento pluriennale e multimiliardario da par-

te di Microsoft, volto a estendere la collaborazione nella ricerca sull'IA e nello sviluppo di supercomputer specializzati.
I termini dell'accordo prevedevano che Microsoft diventasse il fornitore cloud esclusivo per OpenAI, con Azure che alimentava tutti i carichi di lavoro. In cambio, Microsoft ha ottenuto diritti sulla proprietà intellettuale sui modelli di OpenAI, consentendone l'integrazione nei propri prodotti, come Copilot, e l'offerta di servizi basati su questi modelli attraverso l'Azure OpenAI Service. L'accordo inoltre prevedeva una condivisione dei profitti bilaterale in modo che entrambe le aziende potessero trarre vantaggio dall'utilizzo più diffuso dei modelli di OpenAI. Questa collaborazione ha portato vantaggi significativi a entrambe le aziende: per OpenAI, l'accesso alle vaste risorse computazionali di Microsoft ha facilitato l'addestramento di modelli sempre più avanzati, per Microsoft, invece, l'integrazione delle tecnologie di OpenAI ha arricchito la propria offerta di prodotti e servizi con funzionalità di IA all'avanguardia, rafforzando la sua posizione nel mercato del cloud computing e dell'intelligenza artificiale.
Tuttavia, a gennaio 2025 i due colossi hanno annunciato una revisione dell'accordo che ne lascia invariata la struttura portante ma che offre ad OpenAI più flessibilità. In particolare, a seguito del progetto Stargate, OpenAI potrà accedere a risorse di calcolo anche da fornitori di cloud diversi da Microsoft, tra questi figura Oracle, una delle società coinvolte nel progetto Stargate.

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

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

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

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

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

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

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

tali accordi favoriscono lo sviluppo tecnologico e l'efficienza operativa, dall'altro potrebbero ridurre la diversità dell'offerta e rallentare l'emergere di modelli innovativi indipendenti, con implicazioni rilevanti per la competizione e la regolamentazione del settore.

Per quanto riguarda l'integrazione a valle, come evidenziato sia dalla britannica *CMA (Competition & Market Authority)* nel report del 2023 sui FMs (ulteriormente approfondito nell'update paper del 2024) che dall*'OECD (Organisation for Economic Co-operation and Development)* nel paper *Artificial intelligence, data and competition* del 2024, questa non si limita esclusivamente alla distribuzione dei modelli attraverso interfacce proprietarie, ma si estende alla capacità delle grandi piattaforme di consolidare il proprio vantaggio competitivo sfruttando ecosistemi digitali già esistenti. Le aziende con una base utenti consolidata in mercati digitali adiacenti, come sistemi operativi, suite di produttività o motori di ricerca, possono integrare le proprie soluzioni di intelligenza artificiale direttamente nei loro servizi, rendendo più complessa la competizione per gli sviluppatori indipendenti. Esempi di questo fenomeno fanno riferimento alle aziende GAMMA (Google, Amazon, Microsoft, Meta e Apple), che, come evidenziato dalla figura 4.6 agiscono lungo tutta la catena del valore della GenAI.

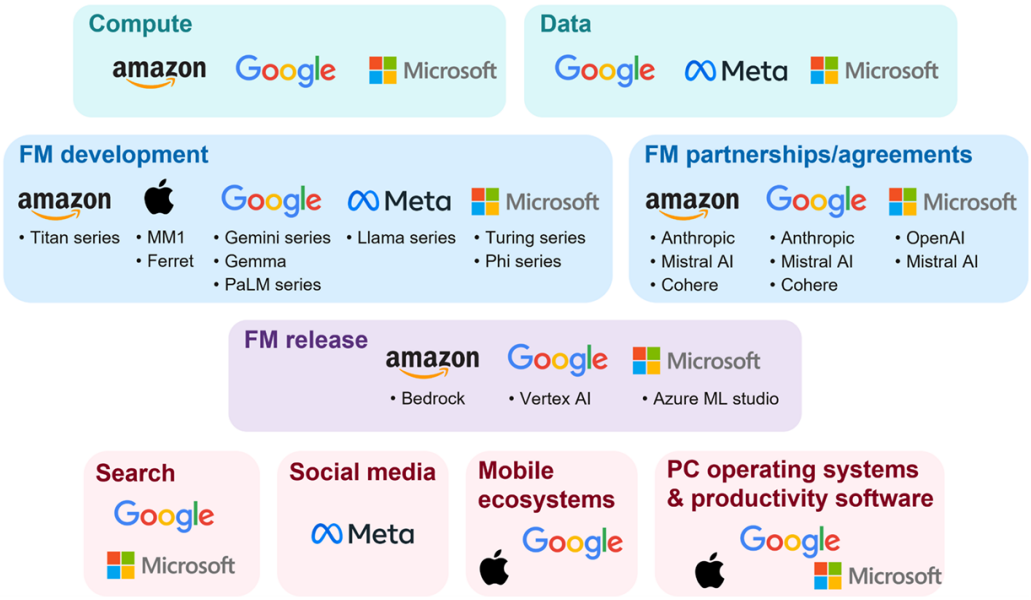

Figura 4.6: Presenza delle aziende GAMMA lungo la catena del valore della GenAI.[56]

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

Tutte queste società hanno integrato i loro modelli all'interno dei propri servizi, ad esempio Microsoft, attraverso la partnership con OpenAI, ha incorporato modelli di intelligenza artificiale nel proprio software di produttività (Copilot in Office), nel sistema operativo Windows e nel motore di ricerca Bing. Google segue una strategia simile, implementando i propri FM nella Search Generative Experience. Queste aziende, oltre ad avere il vantaggio di poter utilizzare per l'addestramento i dati generati dai propri servizi, come Meta con i dati di Facebook e Instagram o Google con le trascrizioni dei video di YouTube, posseggono l'ulteriore vantaggio legato alla possibilità di utilizzare i dati generati dagli utenti durante l'utilizzo del modello per migliorare il modello stesso, innescando dei data feedback loops[57].

Box 7: Data feedback loops

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

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

[57]Hagiu and Wright, (2025), Artificial intelligence and competition policy, International Journal of Industrial Organization.

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

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

1. *Bundling*: si verifica quando un'azienda vende due o più prodotti insieme come un pacchetto, spesso a un prezzo inferiore rispetto all'acquisto separato. Se il bundling avvantaggia in modo sleale un prodotto rispetto a quello di un concorrente (es. Microsoft che include gratuitamente Copilot in Office, rendendo più difficile per altri sviluppatori competere), può limitare la concorrenza;

2. *Tying*: si ha quando un'azienda impone ai clienti l'acquisto di un prodotto per poter accedere a un altro. Ad esempio, se un'azienda richiedesse l'uso del proprio FM per accedere a determinati servizi cloud, potrebbe impedire ai concorrenti di entrare nel mercato;

3. *Self-preferencing*: accade quando una piattaforma digitale o un'azienda favorisce i propri prodotti rispetto a quelli dei concorrenti. Un esempio potrebbe

essere Google che mostra il proprio FM prima di quelli di altri sviluppatori nei risultati di ricerca, limitando la visibilità di alternative competitive.

Queste possibili pratiche hanno attivato le autorità regolatorie mondiali, le quali hanno avviato indagini sulle grandi aziende integrate verticalmente, al fine di evitare una distorsione del mercato da parte delle imprese consolidate per permettere agli utenti di accedere liberamente ad una diversità di modelli senza il verificarsi dell'effetto di lock-in. In figura 4.7, secondo una recente analisi dell'Antitrust inglese, viene rappresentato il possibile feedback loop che le grandi aziende potrebbero innescare col fine di aumentare la loro posizione di mercato a discapito dei concorrenti:

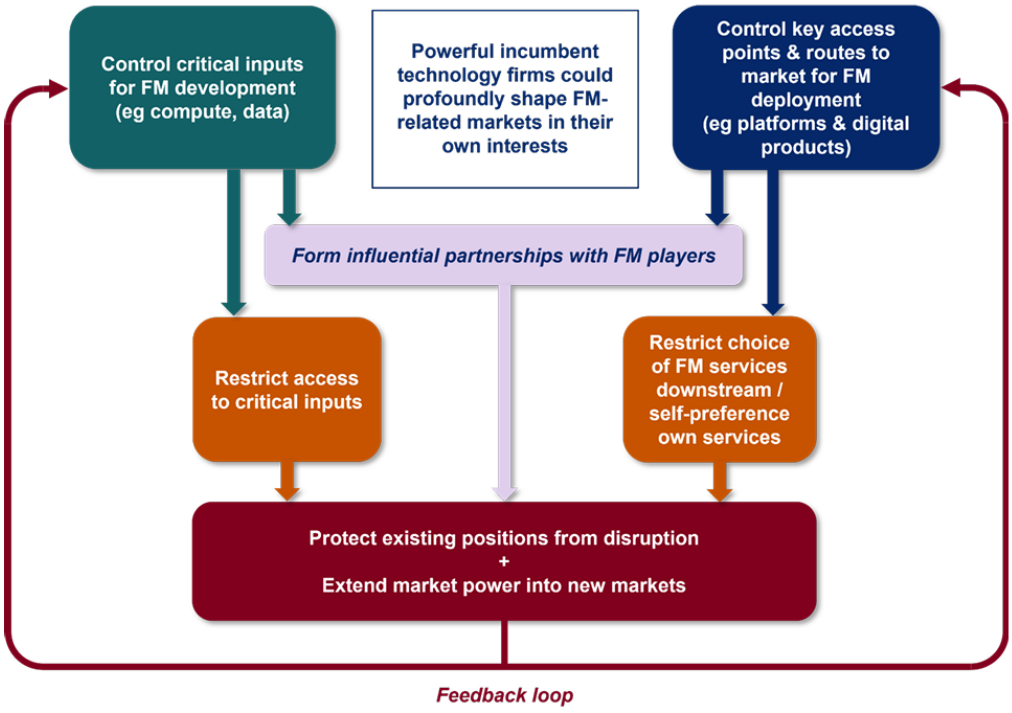

Figura 4.7: Possibile feedback loop che le aziende leader potrebbero innescare per favorire i loro interessi.[58]

4.3 Investimenti e sviluppo dei modelli in USA, Europa e Cina

Come è emerso dai dati presentati, i principali players all'interno di questo mercato provengono principalmente dagli Stati Uniti e dalla Cina, con concorrenti europei che faticano a penetrare nel mercato. Questa tendenza è visibile anche a livello

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

del numero di modelli rilasciati dalle diverse nazioni[59]; infatti, la figura 4.8 mostra chiaramente una dominanza da parte degli USA sul numero di modelli rilasciati (109 nel 2023 e 65 nel 2024), rispecchiando il suo dominio a livello globale. La tendenza interessante è quella relativa al 2024, con una diminuzione del numero di modelli rilasciati dagli Stati Uniti e dalla Cina, il che suggerisce, come precedentemente accennato, a un consolidamento dei modelli già rilasciati negli anni passati. L'Europa e il Regno Unito hanno invece una tendenza opposta che vede il numero dei modelli rilasciati aumentare da 16 nel 2023 a 28 nel 2024, da cui si intuisce una partecipazione più attiva da parte di queste ultime.

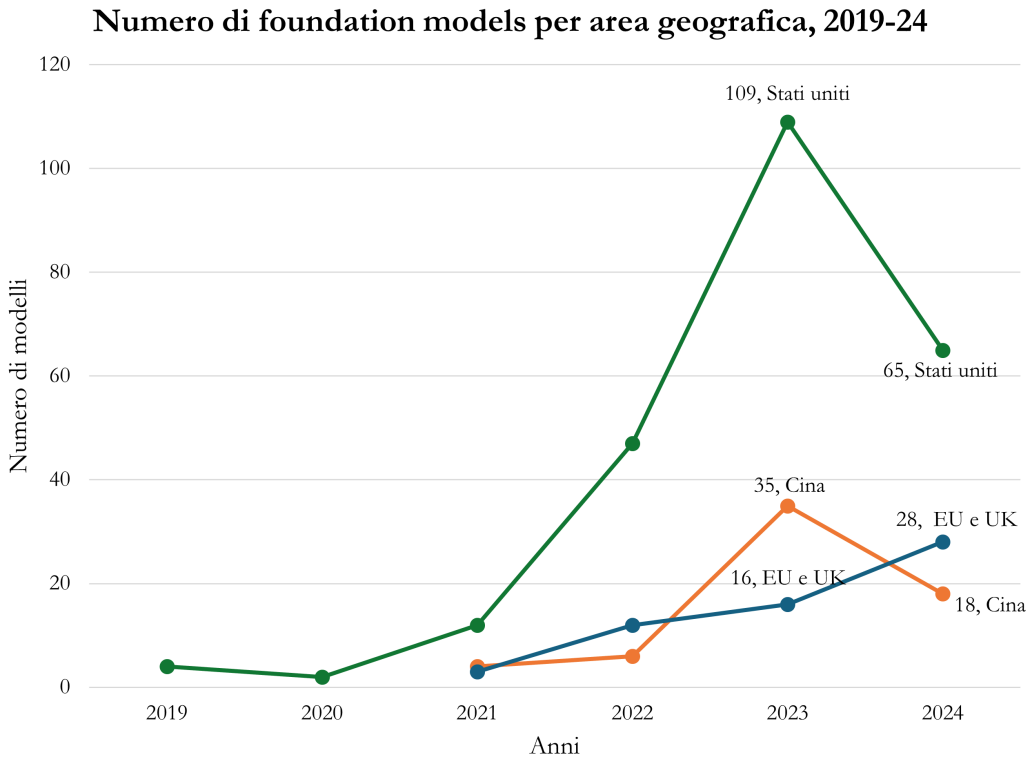

Figura 4.8: FMs riliasciati da US, UE, UK e Cina dal 2019 al 2024.[60]

Un elemento chiave di questa leadership è rappresentato dal volume degli investimenti privati (figura 4.9). Nel 2023, gli Stati Uniti hanno investito 22,46 miliardi di dollari nel settore della GenAI, una cifra significativamente superiore rispetto a quella di Europa e Regno Unito (0,74 miliardi di dollari) e Cina (0,65 miliardi di dollari). Questa disparità suggerisce che la capacità statunitense di attrarre capitali privati sia un fattore determinante per il rapido sviluppo del settore. Gli elevati

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

[60] Ecosystem graphs for foundation models, CRM Stanford.

investimenti consentono infatti di finanziare infrastrutture computazionali avanzate, l'acquisizione di talenti altamente specializzati e progetti di ricerca su larga scala, creando un vantaggio competitivo significativo.

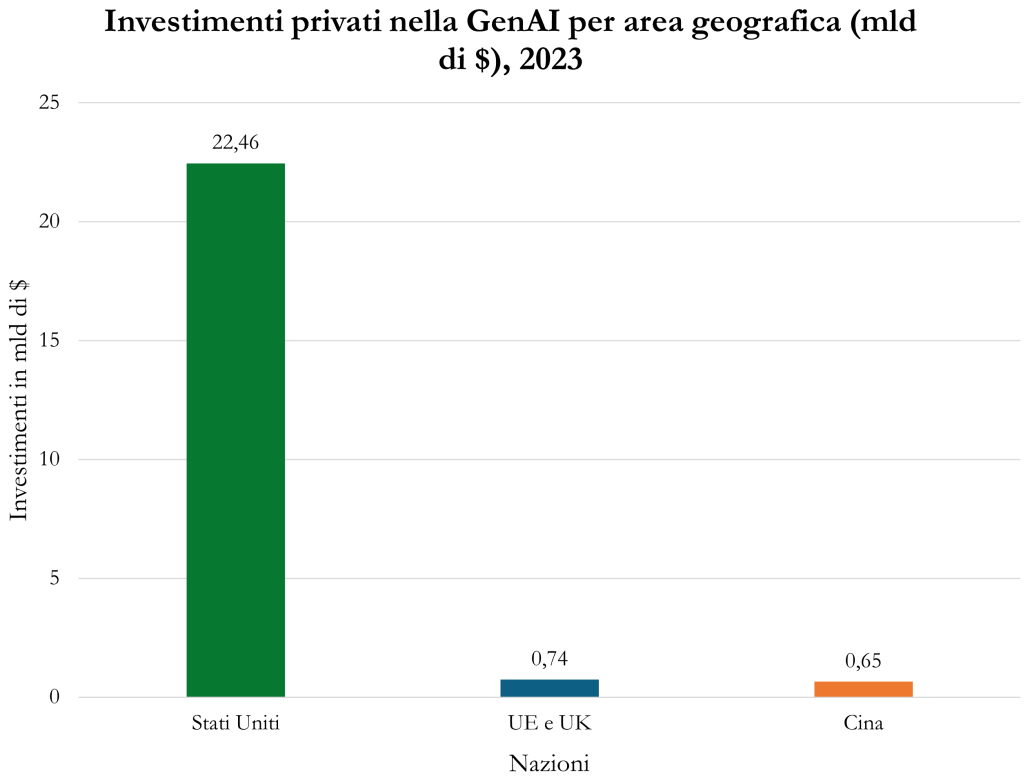

Figura 4.9: Investimenti privati in IAG negli USA, UE, UK e Cina nel 2023.[61]

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

L'effetto complessivo di questa regolamentazione meno flessibile è il rischio di una minore attrattività del mercato europeo per gli investitori privati, che probabilmente

[61] Artificial Intelligence Index Report, Stanford University (2024).

preferiscono destinare i propri capitali a contesti più favorevoli alla crescita dell'IA. Se da un lato l'UE punta a costruire un ecosistema di intelligenza artificiale etico e sicuro, dall'altro il rischio è che questa rigidità normativa possa limitare l'innovazione e la competitività delle aziende europee nel panorama globale dell'IAG.

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

1. *European AI Champion*: 150 miliardi di euro da fondi privati;
2. *InvestAI*: 50 miliardi di euro da fondi pubblici.

Inoltre, dei 200 miliardi di euro stanziati, 20 saranno destinati alla costruzione di 4 Gigafactories, ovvero dei centri di calcolo di grandi dimensioni per fornire la tecnologia necessaria all'addestramento dei modelli anche alle aziende più piccole. La Francia, che si è dimostrata essere il principale player europeo nel mercato dell'IAG globale grazie alla startup Mistral, conferma la sua posizione annunciando un piano nazionale di 109 miliardi di euro finanziato sia da investitori francesi che esteri con il principale obiettivo di costruire data center finalizzati allo sviluppo di modelli. Nonostante il coinvolgimento limitato dell'Italia in queste iniziative, sono numerosi i progetti realizzati legati all'IAG (tabella 4.1), tra i più rilevanti vi sono:

Sviluppatore	**Modello**	**Descrizione**
iGenius	Colosseum 355B	Un LLM progettato specificamente per operare in settori altamente regolamentati e nella pubblica amministrazione. Rispetta le rigorose normative sulla protezione dei dati, offrendo alle aziende anche la possibilità di ospitare il modello all'interno delle proprie infrastrutture, garantendo il pieno controllo sui dati sensibili
Almawave	Velvet 14B e Velvet 2B	Modelli addestrati tramite il supercomputer Leonardo e rilasciati in modalità open source che rispettano tutto il quadro regolatorio europeo garantendo efficienza e bassi consumi energetici
ASC27	Vitruvian-1	Un LLM addestrato con 8 GPU H100 di NVIDIA con un costo di addestramento dell'ordine delle decina di migliaia di euro. Progettato prevalentemente per compiti logici, ha ottenuto un punteggio di 94 sul benchmark MATH-500, dimostrandosi competitivo con i modelli di punta dell'aziende statunitensi e cinesi
Engineering	EngGPT	Un LLM proprietario sviluppato per le aziende e la pubblica amministrazione tramite tecniche che garantiscono la trasparenza e la sicurezza dei dati in conformità alle linee guida dell'AI Act

Tabella 4.1: Modelli di IAG italiani.

L'emergere di DeepSeek ha quindi influenzato i mercati globali, spingendo sia Europa che Stati Uniti ad intensificare gli investimenti nel settore. Tuttavia, se da un lato l'Europa ha stanziato 200 miliardi di euro tramite i piani EuropeanAI e InvestAI, negli Stati Uniti, il 21 gennaio 2025, è stato annunciato il progetto *Stargate* che mira a consolidare la leadership americana nel settore: una joint venture tra OpenAI, SoftBank, Oracle e MGX che investiranno 500 miliardi di dollari entro il 2029 per la costruzione di infrastrutture di calcolo per l'intelligenza artificiale. Inoltre, il progetto vede la partecipazione anche di NVIDIA, Microsoft e Arm come partner tecnici. Questa disparità di investimenti, con un volume di investimento americano superiore di circa 300 miliardi di dollari rispetto a quello europeo, potrebbe riflettere strategie diametralmente opposte. Mentre l'Europa si focalizza sulla regolamentazione e sulla creazione di un ecosistema di IA etico e conforme al quadro normativo, gli Stati

uniti puntano sulla creazione di infrastrutture computazionali massicce. Un esempio concreto è rappresentato da Grok 3, la nuova famiglia di modelli sviluppata dalla stratup xAI fondata da Elon Musk. Grok 3 è stato addestrato utilizzando una delle più grandi infrastrutture di calcolo mai realizzate sfruttando oltre 200.000 GPU.

Se da un lato l'approccio europeo mira a garantire trasparenza e conformità normativa, dall'altro rischia di rallentare la competitività del continente in un settore dominato da economie che investono in modo più aggressivo. La questione chiave per il futuro sarà comprendere se, attraverso questi piani di investimento e una regolamentazione armonizzata, l'Europa riuscirà a ridurre il gap tecnologico con gli Stati Uniti o se, al contrario, il divario continuerà ad ampliarsi.

Capitolo 5

Conclusioni

In conclusione, l'analisi degli assetti di mercato legati ai modelli di intelligenza artificiale generativa evidenzia tre aspetti principali. In primo luogo, nonostante la presenza di numerosi attori e l'apparente apertura alla concorrenza, emergono chiari fenomeni di concentrazione in diversi segmenti chiave della catena del valore. Nel mercato dei chip e delle infrastrutture cloud, sebbene vi siano diverse aziende che cercano di inserirsi, NVIDIA e i grandi Cloud Service Provider mantengono una posizione di assoluto rilievo, beneficiando di economie di scala e forti barriere tecnologiche che limitano di fatto la concorrenza.

Analogamente, nel segmento dei FMs, nonostante un crescente numero di player operanti sia con modelli proprietari sia open source, OpenAI emerge nettamente come leader di mercato, mentre altre realtà faticano ad acquisire significative quote di mercato e visibilità, dimostrando una dinamica competitiva fortemente sbilanciata. Queste dinamiche sono ulteriormente accentuate dalla diffusa pratica di integrazione verticale, mediante la quale i grandi attori consolidano la propria posizione dominante, integrando le diverse fasi della catena del valore e generando ulteriori barriere competitive nei confronti di competitor di dimensioni inferiori.

Inoltre, l'analisi evidenzia una sostanziale disparità nella capacità competitiva tra diverse aree geografiche. Gli Stati Uniti continuano a guidare il settore grazie ad investimenti privati consistenti, un ecosistema innovativo dinamico e infrastrutture avanzate, mentre la Cina rappresenta un concorrente strategico rilevante. Al contrario, l'Europa si presenta in ritardo rispetto ai principali competitor internazionali, con un mercato frammentato e una minore propensione agli investimenti privati, fattori che limitano significativamente il suo ruolo nello sviluppo e nell'adozione dei FMs. Alla luce di quanto esposto, le scelte strategiche e regolatorie rappresenteran-

no un fattore determinante per equilibrare un mercato attualmente caratterizzato da diversi squilibri competitivi.

Appendice A

Architetture dei modelli di intelligenza artificiale

Generative Adversarial Networks (GANs): i GANs sono modelli che vengono addestrati tramite un processo contradditorio sfruttando due modelli: (i) un generatore, che cerca di creare un campione di dati simili a quello fornito in ingresso e, (ii) un discriminatore, che cerca di distinguere i dati creati da quelli originali. Tramite questo processo iterativo, i due modelli migliorano le proprie capacità. Le GAN si sono dimostrate pratiche per attività complesse come l'elaborazione delle immagini e l'assistenza sanitaria, ma possono sorgere limitazioni come la mancata corrispondenza dei dati. I modelli DCGAN (Deep Convolutional Generative Adversarial Networks) e WGAN (Wasserstein Generative Adversarial Networks) superano queste limitazioni. I DCGAN introducono degli strati convoluzionali sia nel modello generatore che in quello discriminatore, questo permette di catturare delle dipendenze spaziali nei dati il che può migliorare la qualità delle immagini generate. I modelli WGAN, invece, introducono la distanza di Wasserstein utilizzata come funzione di perdita. La distanza di Wasserstein è una misura della differenza tra due distribuzioni di probabilità. In un WGAN, il modello generatore viene addestrato per produrre campioni da una distribuzione sintetica che siano il più vicino possibile ai campioni reali dei dati di training. Minimizzando questa funzione di perdita, il WGAN impara a generare campioni sintetici di alta qualità che siano il più simili possibile ai campioni reali dei dati di addestramento.

Variational Autoencoders (VAEs): i VAE combinano elementi di autoencoder[62] con modelli a variabili latenti[63] per la generazione di dati. Imparano una rappresentazione compressa dei dati in uno spazio latente, che consente loro di generare nuovi punti dati. Durante l'addestramento, i VAE mirano a ricostruire i dati di input garantendo al contempo che la distribuzione dello spazio latente assomigli a una distribuzione standard, tipicamente una gaussiana. Questo duplice obiettivo consente ai VAE di acquisire le caratteristiche essenziali dei dati, consentendo loro di generare campioni diversificati. I VAE hanno trovato un uso pratico nella compressione dei dati, nell'apprendimento delle rappresentazioni e nella generazione di immagini, rendendoli uno strumento versatile in vari settori.

Diffusion models: i modelli di diffusione generano i dati simulando un processo che trasforma gradualmente una distribuzione semplice, come quella gaussiana in una più complessa, simile alla distribuzione dei dati target. Questo processo iterativo comporta l'aggiunta di rumore ai dati e la sua progressiva rimozione. Questi modelli eccellono nel catturare pattern complessi nei dati e sono preziosi per la modellazione generativa che richiede campioni ad alta dimensionalità. Per migliorare l'output di questi modelli sono stati introdotti i DDPM, un'estensione dei diffusion models, che ottimizzano il processo di rimozione del rumore (denoising) utilizzando un modello probabilistico. In particolare, i DDPM applicano il denoising in modo iterativo con l'obiettivo di raffinare i campioni in modo incrementale, migliorando la fedeltà del campione generato. Ogni passaggio nel processo è modellato tramite una distribuzione gaussiana, e l'algoritmo cerca di apprendere i parametri ottimali per ridurre il rumore aggiunto in ciascun passaggio, portando a una generazione di immagini di alta qualità.

Trasformers models: i Trasformers sono emersi come una rivoluzione nel deep learning, in particolare nel Natural Language Processing (NLP). I Trasformers gestiscono enormi set di dati, eseguono calcoli paralleli in modo efficiente e acquisiscono relazioni a lungo raggio all'interno di dati sequenziali. Le loro componenti principali

[62] Gli autoencoder sono modelli composti da due parti, l'encoder e il decoder, che imparano a comprimere i dati di input in rappresentazioni più compatte (encoding) e poi a ricostruire l'output originale dall'encoding, minimizzando la perdita di informazioni durante questo processo.

[63] Le variabili latenti sono variabili nascoste che descrivono caratteristiche fondamentali dei dati ma che non sono osservabili direttamente. Nei VAE, queste variabili vengono modellate per apprendere e rappresentare tratti strutturali sottostanti dei dati, come la forma e la composizione visiva nelle immagini.

includono meccanismi di self-attention[64], un'architettura di decodifica e codifica posizionale. Il meccanismo di self-attention consente al modello di valutare l'importanza relativa delle diverse parti della sequenza di input quando si effettuano previsioni. Questa capacità e le competenze di parallelizzazione hanno fatto progredire significativamente i compiti di NLP. Le architetture dei Trasformers costituiscono la base per i Large Language Models (LLM). Questi modelli sfruttano le loro immense dimensioni e complessità per identificare pattern e sfumature contestuali estremamente complessi. Gli LLM sono in genere pre-addestrati su enormi set di dati per poi venire ottimizzati per compiti specifici tramite le tecniche di fine-tuning, rendendoli adattabili a varie applicazioni come la traduzione linguistica o la scrittura creativa. Gli LLM hanno migliorato in modo significativo le capacità dei chatbot e degli assistenti virtuali, consentendo loro di comprendere e generare risposte di qualità pari a quella umana.

Tabella A.1: Tabella riassuntiva dei modelli.

Modello	**Caratteristiche**	**Applicazione**
GAN	Due reti neurali che competono: generatore e discriminatore	Generazione immagini, Aumento dei dati[65]
DCGAN	Include una rete neurale convoluzionale sia nel generatore sia nel discriminatore	Generazione di immagini, aumento dei dati
WGAN	Include la distanza di Wasserstein	Campioni di alta qualità
VAE	Architettura encoder-decoder con spazio latente	Compressione dei dati, Representation Learning[66], Generazione di immagini
Modelli di Diffusione	Processo di diffusione che cambia la distribuzione di base nella distribuzione target	Campioni di alta qualità, Generazione di immagini
DDPM	Riduzione del rumore nel processo di diffusione	Campioni di alta qualità, Generazione di immagini
VDM	Include il principio variazionale nel modello	Campioni di alta qualità, Generazione di immagini
Transformer	Utilizza il meccanismo di self-attention	Natural Language Processing

[64]Il meccanismo di self-attention dei Transformer consente di catturare dipendenze tra parole che sono distanti tra loro nella sequenza di input, permettendo quindi al modello di comprendere meglio le relazioni a lungo raggio.

[66]Per aumento dei dati si fa riferimento ad una serie tecniche utilizzate per aumentare la quantità e la varietà dei dati di addestramento senza dover raccogliere nuovi dati reali.

[66]Un modello di intelligenza artificiale impara a estrarre automaticamente rappresentazioni significative e compatte dai dati grezzi, senza dover ricevere esplicite istruzioni su cosa cercare.

Appendice B

Costi legati alle tecniche di preparazione dei dataset

Oltre ai costi legati alla raccolta, figurano anche quelli legati alla preparazione di questi per il pre-addestramento. La preparazione è un processo articolato in diverse fasi che variano anche in base alla natura del dataset, di seguito vengono esposte le principali tecniche utilizzate nella pipeline di preparazione:

1. *Data preparation*: la preparazione dei dati rappresenta un passaggio fondamentale nello sviluppo di modelli di intelligenza artificiale, poiché i dati grezzi spesso non sono immediatamente pronti per essere utilizzati e richiedono interventi specifici per eliminare problematiche quali valori mancanti, duplicati e incongruenze;

2. *Data reduction*: consente di ridurre la complessità di un dataset mantenendo le informazioni essenziali. La riduzione può essere implementata attraverso la riduzione del numero di campioni o di caratteristiche, apportando benefici in termini di efficienza computazionale e interpretabilità del modello, dove per interpretabilità si intente la capacità di un essere umano di intelligenza media di comprendere il funzionamento del modello. Dal punto di vista dei campioni, la riduzione della dimensione del dataset semplifica l'elaborazione e consente di alleviare problemi legati alla memoria e al calcolo, oltre a contribuire al bilanciamento dei dati mediante il sotto campionamento delle classi maggioritarie. D'altro canto, la riduzione del numero di caratteristiche elimina variabili irrilevanti o ridondanti, riducendo il rischio di overfitting e migliorando la velocità sia nell'addestramento che nell'inferenza;

3. *Data augmentation*: consente di ampliare e migliorare la qualità dei dataset incompleti o insufficienti per l'addestramento di modelli complessi. Questo obiettivo viene raggiunto attraverso la generazione artificiale di nuovi dati a partire da quelli esistenti, incrementando non solo la dimensione, ma anche la diversità del dataset. Tale arricchimento contribuisce significativamente a migliorare l'ottimizzazione, la capacità di generalizzazione e la robustezza dei modelli, riducendo al contempo il rischio di overfitting.

Le tecniche appena descritte rappresentano le metodologie più comuni impiegate nella pipeline di preparazione dei dataset. Sebbene l'applicazione di tali processi consenta di ottimizzare i tempi di addestramento del modello, con conseguente riduzione dei costi complessivi, la loro implementazione richiede investimenti significativi sia in infrastrutture hardware avanzate sia in risorse umane altamente specializzate (è anche possibile affidarsi a servizi cloud, come Amazon SageMaker, che mettono a disposizioni strumenti per la preparazione dei dati). Tali costi possono variare considerevolmente in relazione alla tipologia di dataset utilizzato. Ad esempio, un dataset ottenuto tramite web scraping richiede una pipeline di preparazione più complessa, comportando costi superiori rispetto a un dataset acquisito tramite licenza. Quest'ultimo, pur presentando un costo iniziale elevato, garantisce standard qualitativi più elevati, riducendo così la necessità di interventi di pulizia e preparazione.

Appendice C

Costi legati alle tecniche di etichettatura dei dataset

Le strategie di etichettatura sono molteplici e combinano sia l'intervento umano che avanzate tecnologie automatizzate, adattandosi alle specifiche esigenze del modello e del dominio di applicazione. Le principali tecniche di etichettatura sono:

1. *Etichettatura crowdourcing*: è uno degli approcci più economici e rapidi che consiste nel suddividere il compito di etichettatura in task più piccole e distribuirli ad annotatori non esperti (crowd workers). Solitamente si utilizzano piattaforme online dedicate come Amazon Mechanical Turk che consente a più persone di contribuire parallelamente all'etichettatura dei dati. Un altro esempio rilevante di crowd sourcing è lo strumento reCAPTCHA di Google, originariamente ideato per prevenire frodi sul web, viene anche utilizzato per assegnare alle immagini delle etichette. Ovvero, quando un utente è chiamato a risolvere un reCAPTCHA identificando, ad esempio, le auto in una serie di immagini queste vengono etichettate e utilizzate per creare un database di etichette per le immagini. Queste tipologie di approcci permettono quindi una elevata scalabilità permettendo di annotare una grande quantità di dati ad un costo relativamente ridotto. Tuttavia, gli annotatori potrebbero generare etichette di bassa qualità, data la soggettività nel processo, e compromettere l'efficienza del modello. Ciononostante, esistono diverse tecniche per mitigare questi rischi, ad esempio richiedere a più crowd worker di annotare lo stesso campione per poi inferire su di questo ed estrarre un'etichetta comune;

2. *Etichettatura semi-supervisionata*: è un approccio ibrido che combina un pic-

colo insieme di dati etichettati manualmente con un ampio insieme di dati non etichettati. Questo metodo utilizza i dati etichettati come base per inferire o predire le etichette mancanti sui dati non etichettati attraverso algoritmi di apprendimento automatico. Tra le tecniche più diffuse rientrano il self-training, in cui un modello addestrato su dati etichettati genera pseudo-etichette per i dati non etichettati, e la label propagation, che diffonde etichette utilizzando rappresentazioni grafiche per identificare somiglianze tra campioni. L'etichettatura semi-supervisionata è vantaggiosa sia in termini di costi che di scalabilità, in quanto richiede meno annotazioni manuali rispetto ai metodi completamente supervisionati, risultando particolarmente adatta per dataset di grandi dimensioni. Tuttavia, la qualità delle pseudo-etichette dipende strettamente dall'accuratezza del modello e dalla rappresentatività del set iniziale di dati etichettati, il che può introdurre rumore o bias nel dataset;

3. *Apprendimento attivo*: è un procedimento iterativo che coinvolge sia gli esseri umani che un modello di intelligenza artificiale. In questo metodo, il modello identifica autonomamente i campioni di dati più informativi o incerti, che hanno il maggior potenziale di migliorare le sue prestazioni, e li sottopone ad annotatori umani per l'etichettatura. Questo approccio si rivela particolarmente utile quando si lavora con dataset di grandi dimensioni, poiché consente di ridurre significativamente il numero di campioni che devono essere etichettati manualmente, limitando i costi e i tempi di annotazione. Tuttavia, richiede un'infrastruttura computazionale avanzata per gestire l'iterazione continua tra il modello e il processo di selezione dei dati, comportando costi non trascurabili. In secondo luogo, implica la presenza di annotatori qualificati, in grado di gestire campioni particolarmente complessi o ambigui, il che può incrementare i costi del capitale umano;

4. *Supervisione a distanza*: è una tecnica di etichettatura automatica che assegna etichette per grandi quantità di dati sfruttando come riferimento fonti esterne in base a corrispondenze predefinite. Questo approccio è particolarmente utile quando l'etichettatura manuale è troppo costosa o. Permette di creare dataset di grandi dimensioni rapidamente e a costi significativamente inferiori rispetto a metodi completamente manuali in quanto non richiede l'intervento del capitale umano ma richiede un'infrastruttura tecnologica avanzata per integrare i dati da fonti esterne e implementare algoritmi di associazione. Tuttavia,

le etichette generate automaticamente possono essere rumorose o imprecise, soprattutto in caso di disallineamenti tra il dataset e la fonte di riferimento. Questo rumore può compromettere la qualità del dataset, richiedendo ulteriori interventi e di conseguenza costi per pulire i dati o correggere errori;

5. *Data programming*: è una tecnica di etichettatura automatizzata che si basa sull'utilizzo di regole euristiche, chiamate labeling functions, per generare etichette su grandi dataset. Queste regole sono progettate da esperti del dominio e vengono applicate ai dati grezzi per assegnare etichette in modo semi-automatizzato. L'obiettivo principale della data programming è ridurre la necessità di etichettatura manuale, sfruttando al contempo conoscenze esperte per migliorare l'accuratezza delle etichette generate. Come la supervisione a distanza, anche questa tecnica viene utilizzata in contesti in cui 'etichettatura manuale risulterebbe proibitiva in termini di tempo e costi a causa della dimensione del dataset. Ciò nonostante, le regole di etichettatura possono essere soggette a errori e generare rumore nei dati, soprattutto quando le labeling functions non sono sufficientemente robuste o dettagliate. Questo rumore può compromettere la qualità complessiva del dataset, richiedendo ulteriori interventi di pulizia e validazione. Inoltre, la progettazione delle regole richiede una conoscenza approfondita del dominio e può richiedere un investimento iniziale significativo in termini di competenze specialistiche.

Bibliografia

[1] Yue Hern Tan, Hui Na Chua, Yeh-Ching Low e Muhammed Basheer Jasser. «Current Landscape of Generative AI: Models, Applications, Regulations and Challenges». In: *2024 IEEE 14th International Conference on Control System, Computing and Engineering (ICCSCE)*. 2024, pp. 168–173. DOI: 10.1109/ICCSCE61582.2024.10696569.

[2] Kuldeep Singh Kaswan, Jagjit Singh Dhatterwal, Kiran Malik e Anupam Baliyan. «Generative AI: A Review on Models and Applications». In: *2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI)*. 2023, pp. 699–704. DOI: 10.1109/ICCSAI59793.2023.10421601.

[3] Roshan Mohammad. «The rise of task-tailored generative models: Redefining specialization in Artificial Intelligence». In: *International Journal of Engineering and Technical Research (IJETR)* Volume 9 (ago. 2024), pp. 19–28. DOI: 10.5281/zenodo.13604001.

[4] Or Sharir, Barak Peleg e Yoav Shoham. *The Cost of Training NLP Models: A Concise Overview*. Apr. 2020. URL: https://arxiv.org/abs/2004.08900.

[5] Lukas Ryll e Sebastian Seidens. *Evaluating the Performance of Machine learning Algorithms in Financial Market Forecasting: A Comprehensive survey*. Giu. 2019. URL: https://arxiv.org/abs/1906.07786.

[6] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes e Ajmal Mian. *A Comprehensive Overview of Large Language Models*. 2024. arXiv: 2307.06435 [cs.CL]. URL: https://arxiv.org/abs/2307.06435.

[7] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu e Enhong Chen. «A survey on multimodal large language models». In: *National Science Review* 11.12 (nov. 2024). ISSN: 2053-714X. DOI: 10.1093/nsr/nwae403. URL: http://dx.doi.org/10.1093/nsr/nwae403.

[8] Diederik P Kingma e Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML]. URL: https://arxiv.org/abs/1312.6114.

[9] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg e Gideon Mann. *BloombergGPT: A Large Language Model for Finance*. 2023. arXiv: 2303.17564 [cs.LG]. URL: https://arxiv.org/abs/2303.17564.

[10] Raghavendra Chalapathy e Sanjay Chawla. *Deep Learning for Anomaly Detection: A Survey*. 2019. arXiv: 1901.03407 [cs.LG]. URL: https://arxiv.org/abs/1901.03407.

[11] Jonathan Ho, Ajay Jain e Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: https://arxiv.org/abs/2006.11239.

[12] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao e Hoifung Poon. «Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing». In: *ACM Transactions on Computing for Healthcare* 3.1 (ott. 2021), pp. 1–23. ISSN: 2637-8051. DOI: 10.1145/3458754. URL: http://dx.doi.org/10.1145/3458754.

[13] Jason Wei et al. *Emergent Abilities of Large Language Models*. 2022. arXiv: 2206.07682 [cs.CL]. URL: https://arxiv.org/abs/2206.07682.

[14] Milad Moradi, Ke Yan, David Colwell, Matthias Samwald e Rhona Asgari. *Exploring the landscape of large language models: Foundations, techniques, and challenges*. 2024. arXiv: 2404.11973 [cs.AI]. URL: https://arxiv.org/abs/2404.11973.

[15] Johannes Schneider, Christian Meske e Pauline Kuss. «Foundation models». In: *Business Information Systems Engineering* 66.2 (gen. 2024), pp. 221–231. DOI: 10.1007/s12599-024-00851-0. URL: https://doi.org/10.1007/s12599-024-00851-0.

[16] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL]. URL: https://arxiv.org/abs/2302.13971.

[17] Nestor Maslej et al. *Artificial Intelligence Index Report 2024*. 2024. arXiv: 2405.19522 [cs.AI]. URL: https://arxiv.org/abs/2405.19522.

[18] *Artificial Intelligence: in-depth market analysis | Statista*. URL: https://www.statista.com/study/50485/in-depth-report-artificial-intelligence/.

[19] Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej e Percy Liang. *The 2024 Foundation Model Transparency Index*. 2025. arXiv: 2407.12929 [cs.LG]. URL: https://arxiv.org/abs/2407.12929.

[20] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

[21] Carl Doersch. *Tutorial on Variational Autoencoders*. 2021. arXiv: 1606.05908 [stat.ML]. URL: https://arxiv.org/abs/1606.05908.

[22] Sayash Kapoor et al. *On the Societal Impact of Open Foundation Models*. 2024. arXiv: 2403.07918 [cs.CY]. URL: https://arxiv.org/abs/2403.07918.

[23] Girish Sastry et al. *Computing Power and the Governance of Artificial Intelligence*. 2024. arXiv: 2402.08797 [cs.CY]. URL: https://arxiv.org/abs/2402.08797.

[24] Irene Solaiman. *The Gradient of Generative AI Release: Methods and Considerations*. 2023. arXiv: 2302.04844 [cs.CY]. URL: https://arxiv.org/abs/2302.04844.

[25] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang e Percy Liang. *The Foundation Model Transparency Index*. 2023. arXiv: 2310.12941 [cs.LG]. URL: https://arxiv.org/abs/2310.12941.

[26] Alex Singla, Alexander Sukharevsky, Lareina Yee e Michael Chui. *The state of AI in early 2024: Gen AI adoption spikes and starts to generate value*. Mag. 2024. URL: https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2024/the-state-of-ai-in-early-2024-final.pdf?shouldIndex=false.

[27] Michael Chui, Eric Hazan, Roger Roberts, Alex Singla, Kate Smaje, Alex Sukharevsky, Lareina Yee e Rodney Zemmel. *The economic potential of generative AI: The next productivity frontier*. Giu. 2023. URL: https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20economic%20potential%20of%20generative%20ai%

20the%20next%20productivity%20frontier/the-economic-potential-of-generative-ai-the-next-productivity-frontier.pdf?shouldIndex=false.

[28] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim e Marius Hobbhahn. *Will we run out of data? Limits of LLM scaling based on human-generated data*. 2024. arXiv: 2211.04325 [cs.LG]. URL: https://arxiv.org/abs/2211.04325.

[29] Chandra Irugalbandara, Ashish Mahendra, Roland Daynauth, Tharuka Kasthuri Arachchige, Jayanaka Dantanarayana, Krisztian Flautner, Lingjia Tang, Yiping Kang e Jason Mars. *Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs in Production*. 2024. arXiv: 2312.14972 [cs.SE]. URL: https://arxiv.org/abs/2312.14972.

[30] Cheonsu Jeong. «Domain-specialized LLM: Financial fine-tuning and utilization method using Mistral 7B». In: *Journal of Intelligence and Information Systems* 30.1 (mar. 2024), pp. 93–120. ISSN: 2288-4882. DOI: 10.13088/jiis.2024.30.1.093. URL: http://dx.doi.org/10.13088/jiis.2024.30.1.093.

[31] Anas Awadalla et al. *MINT-1T: Scaling Open-Source Multimodal Data by 10x: A Multimodal Dataset with One Trillion Tokens*. 2024. arXiv: 2406.11271 [cs.CV]. URL: https://arxiv.org/abs/2406.11271.

[32] Wayne Xin Zhao et al. *A Survey of Large Language Models*. 2024. arXiv: 2303.18223 [cs.CL]. URL: https://arxiv.org/abs/2303.18223.

[33] Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu e Sameena Shah. *Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks*. 2023. arXiv: 2305.05862 [cs.CL]. URL: https://arxiv.org/abs/2305.05862.

[34] Long Ouyang et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: 2203.02155 [cs.CL]. URL: https://arxiv.org/abs/2203.02155.

[35] Yung-Chieh Chan, George Pu, Apaar Shanker, Parth Suresh, Penn Jenks, John Heyer e Sam Denton. *Balancing Cost and Effectiveness of Synthetic Data Generation Strategies for LLMs*. 2024. arXiv: 2409.19759 [cs.CL]. URL: https://arxiv.org/abs/2409.19759.

[36] Tim Yarally, Luís Cruz, Daniel Feitosa, June Sallou e Arie van Deursen. *Uncovering Energy-Efficient Practices in Deep Learning Training: Preliminary Steps Towards Green AI*. 2023. arXiv: 2303.13972 [cs.LG]. URL: https://arxiv.org/abs/2303.13972.

[37] xiangyang Ju, Yunsong Wang, Daniel Murnane, Nicholas Choma, Steven Farrell e Paolo Calafiura. *Benchmarking GPU and TPU Performance with Graph Neural Networks*. 2022. arXiv: 2210.12247 [cs.LG]. URL: https://arxiv.org/abs/2210.12247.

[38] Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng e Jimmy Huang. *A Comprehensive Evaluation of Large Language Models on Benchmark Biomedical Text Processing Tasks*. 2024. arXiv: 2310.04270 [cs.CL]. URL: https://arxiv.org/abs/2310.04270.

[39] Chi Wang, Susan Xueqing Liu e Ahmed H. Awadallah. *Cost-Effective Hyperparameter Optimization for Large Language Model Generation Inference*. 2023. arXiv: 2303.04673 [cs.CL]. URL: https://arxiv.org/abs/2303.04673.

[40] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong e Xia Hu. *Data-centric Artificial Intelligence: A Survey*. 2023. arXiv: 2303.10158 [cs.LG]. URL: https://arxiv.org/abs/2303.10158.

[41] Lianmin Zheng et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023. arXiv: 2306.05685 [cs.CL]. URL: https://arxiv.org/abs/2306.05685.

[42] Zhen Huang et al. *OlympicArena: Benchmarking Multi-discipline Cognitive Reasoning for Superintelligent AI*. 2025. arXiv: 2406.12753 [cs.CL]. URL: https://arxiv.org/abs/2406.12753.

[43] Karan Singhal et al. *Towards Expert-Level Medical Question Answering with Large Language Models*. 2023. arXiv: 2305.09617 [cs.CL]. URL: https://arxiv.org/abs/2305.09617.

[44] Zhen Huang, Zengzhi Wang, Shijie Xia e Pengfei Liu. *OlympicArena Medal Ranks: Who Is the Most Intelligent AI So Far?* 2024. arXiv: 2406.16772 [cs.CL]. URL: https://arxiv.org/abs/2406.16772.

[45] Linqing Chen et al. *PharmaGPT: Domain-Specific Large Language Models for Bio-Pharmaceutical and Chemistry*. 2024. arXiv: 2406.18045 [cs.CL]. URL: https://arxiv.org/abs/2406.18045.

[46] Sasha Luccioni, Yacine Jernite e Emma Strubell. «Power Hungry Processing: Watts Driving the Cost of AI Deployment?» In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT 24. ACM, giu. 2024, pp. 85–99. DOI: 10.1145/3630106.3658542. URL: http://dx.doi.org/10.1145/3630106.3658542.

[47] Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu e David Owen. *The rising costs of training frontier AI models*. 2025. arXiv: 2405.21015 [cs.CY]. URL: https://arxiv.org/abs/2405.21015.

[48] Abi Aryan, Aakash Kumar Nain, Andrew McMahon, Lucas Augusto Meyer e Harpreet Singh Sahota. *The Costly Dilemma: Generalization, Evaluation and Cost-Optimal Deployment of Large Language Models*. 2023. arXiv: 2308.08061 [cs.CL]. URL: https://arxiv.org/abs/2308.08061.

[49] Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou e He Tang. *Synthetic Data in AI: Challenges, Applications, and Ethical Implications*. 2024. arXiv: 2401.01629 [cs.LG]. URL: https://arxiv.org/abs/2401.01629.

[50] DeepSeek-AI et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: 2501.12948 [cs.CL]. URL: https://arxiv.org/abs/2501.12948.

[51] Andrei Hagiu e Julian Wright. «Artificial intelligence and competition policy». In: *International Journal of Industrial Organization* (2025), p. 103134. ISSN: 0167-7187. DOI: https://doi.org/10.1016/j.ijindorg.2025.103134. URL: https://www.sciencedirect.com/science/article/pii/S0167718725000013.

[52] Antonio Dalla Zuanna, Davide Dottori, Elena Gentili e Salvatore Lattanzio. *An assessment of occupational exposure to artificial intelligence in Italy*. Rapp. tecn. 878. Banca d'Italia, ott. 2024. URL: https://www.bancaditalia.it/pubblicazioni/qef/2024-0878/QEF_878_24.pdf?language_id=1.

[53] Ebenesar Anna Bagyam. «ANALYSIS OF DATA SCIENCE JOB SALARIES FROM 2020 TO 2024: TRENDS AND INFLUENCING FACTORS». In: ott. 2024, pp. 89–97. ISBN: 978-81-971251-4-0. DOI: 10.5281/zenodo.13883851.

[54] Jai Vipra e Anton Korinek. *Market Concentration Implications of Foundation Models*. 2023. arXiv: 2311.01550 [cs.AI]. URL: https://arxiv.org/abs/2311.01550.

[55] «GENERATIVE ARTIFICIAL INTELLIGENCE: THE COMPETITIVE LANDSCAPE». In: URL: https://api.semanticscholar.org/CorpusID:273663280.

[56] Tobias Härlin, Gardar Björnsson Rova, Alex Singla, Oleg Sokolov e Alex Sukharevsky. *Exploring opportunities in the generative AI value chain*. en. Apr. 2023. URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/exploring-opportunities-in-the-generative-ai-value-chain.

[57] World Economic Forum. *Future of Jobs Report*. en. World Economic Forum, gen. 2025. URL: https://www.weforum.org/publications/the-future-of-jobs-report-2025/in-full/?_gl=1*m72py*_up*MQ..*_gs*MQ..&gclid=Cj0KCQiAz6q-BhCfARIsAOezPxncvuyMNX77kLI7zMzEJrYv_rwkWdEi6mmtJwz-II2cDG5MYt2SzvYaAn98EALw_wcB.

[58] Office of Technology Staff Federal Trade Commission. *Partnerships Between Cloud Service Providers and AI Developers: FTC Staff Report on AI Partnerships & Investments 6(b) Study*. Rapp. tecn. Federal Trade Commission, gen. 2025. URL: https://www.ftc.gov/system/files/ftc_gov/pdf/p246201_aipartnerships6breport_redacted_0.pdf.

[59] Nicolò Monti e Nicola Grandis. *Vitruvian-1 Technical Report: Data-Centric Chain-of-Thought Reasoning in Multilingual Language Models*. Rapp. tecn. Ver. Revision 1. ASC27, feb. 2025. URL: https://storage.googleapis.com/vitruvian-ui-assets/vitruvian-1-rev1.pdf.

[60] Richard May. *Artificial Intelligence, Data and Competition*. Rapp. tecn. No. 18. OECD, mag. 2024. URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/05/artificial-intelligence-data-and-competition_9d0ac766/e7e88884-en.pdf.

[61] Competition e Markets Authority (CMA). *AI Foundation Models: Initial Report*. Rapp. tecn. UK Competition e Markets Authority, set. 2023. URL: https://assets.publishing.service.gov.uk/media/65081d3aa41cc300145612c0/Full_report_.pdf.

[62] Competition e Markets Authority (CMA). *AI Foundation Models: Technical Update Report*. Rapp. tecn. UK Competition e Markets Authority, apr. 2024. URL: https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf.

[63] Competition e Markets Authority (CMA). *AI Foundation Models: Update Paper*. Rapp. tecn. UK Competition e Markets Authority, apr. 2024. URL: https://assets.publishing.service.gov.uk/media/661941a6c1d297c6ad1dfeed/Update_Paper__1_.pdf.

[64] Dataset Providers Alliance (DPA). *AI Data Licensing Position Paper*. Rapp. tecn. Dataset Providers Alliance, 2024. URL: https://5a5ee099-3141-4217-af47-c61b445c2269.filesusr.com/ugd/6112c3_4c700dd417044c4aa268a4a4a9080c88.pdf.

[65] Garante per la Protezione dei Dati Personali (GPDP). *Web scraping ed intelligenza artificiale generativa: nota informativa e possibili azioni di contrasto*. Rapp. tecn. Garante per la Protezione dei Dati Personali (GPDP), mag. 2024. URL: https://www.garanteprivacy.it/documents/10160/0/Web+scraping+ed+intelligenza+artificiale+generativa+-+nota+informativa+e+possibili+azioni+di+contrasto.pdf/40b4600c-80f5-a3ce-a74e-604a8e88e601?version=4.0.

[66] Maicon Roberto Martins. «From On-Premise to Cloud: Evolving IT Infrastructure for the AI Age». In: *World Journal of Advanced Research and Reviews* 20.03 (dic. 2023), pp. 1898–1934. DOI: 10.30574/wjarr.2023.20.3.1590. URL: https://wjarr.com/sites/default/files/WJARR-2023-1590.pdf.

[67] Nivedhaa N. «A Comprehensive Review of AI's Dependence on Data». In: *International Journal of Artificial Intelligence and Data Science (IJADS)* 1.1 (mar. 2024). This paper explores AI's reliance on data, analyzing the impact of data quality, quantity, and biases on AI model performance., pp. 1–11. DOI: 10.13140/RG.2.2.27033.63840. URL: https://iaeme.com/MasterAdmin/Journal_uploads/IJADS/VOLUME_1_ISSUE_1/IJADS_01_01_001.pdf.

[68] Dario Lazzaro, Antonio Emanuele Cinà, Maura Pintor, Ambra Demontis, Battista Biggio, Fabio Roli e Marcello Pelillo. *Minimizing Energy Consumption of Deep Learning Models by Energy-Aware Training*. 2023. arXiv: 2307.00368 [cs.LG]. URL: https://arxiv.org/abs/2307.00368.

[69] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So e Jaewoo Kang. «BioBERT: a pre-trained biomedical language representation model for biomedical text mining». In: *Bioinformatics* 36.4 (set. 2019). A cura di Jonathan Wren, pp. 1234–1240. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btz682. URL: http://dx.doi.org/10.1093/bioinformatics/btz682.

[70] Kai Zhang et al. «A generalist visionlanguage foundation model for diverse biomedical tasks». In: *Nature Medicine* 30.11 (ago. 2024), pp. 3129–3141. ISSN: 1546-170X. DOI: 10.1038/s41591-024-03185-2. URL: http://dx.doi.org/10.1038/s41591-024-03185-2.