# Gli economics dei modelli di Intelligenza artificiale generativa*

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

## 1. La catena del valore

Lo sviluppo e la distribuzione di un FM si articolano lungo una catena del valore (Fig. 1) che coinvolge tre fasi principali, ovvero 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 addestrati, questi modelli possono essere perfezionati attraverso le tecniche di *fine-tuning*, adattandoli a compiti specifici (*task-specific models*) o a 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

* Si tratta del capitolo II del volume ASTRID in corso di pubblicazione, *Modelli di intelligenza artificiale generativa* a cura di R. Arcano, C. Cambini, P. Lupi, A, Manganelli, A, Perrucci, G, Trotta - Passigli Editore, 2026

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 evidenziare 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 *foundation models* in diverse applicazioni.

**Fig. 1:** Catena del valore dell'IAG[1].

## 2. Costi per lo sviluppo di un *Foundation Model*

### *2.1. Dati per il pre-addestramento*

Come anticipato, le prestazioni dei FMs seguono 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[2] 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, ossia 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[3]. 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,

[1] *Generative Artificial Intelligence: The Competitive Landscape* (2024), «Copenhagen Economics».

[2] 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.

[3] 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.

esponendoli al rischio di *overfitting*[4] 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, duplicazioni o contenuti dannosi, per evitare che il modello apprenda informazioni scorrette o distorte.
3. Bilanciamento: se i dati sono bilanciati, si riduce il rischio di *bias* intrinseco, cioè che nessun dominio o gruppo linguistico sia sovrarappresentato rispetto a un altro.

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[5], 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 (Tab. 1).

[4] Fenomeno in cui il modello dimostra prestazioni eccellenti sul dataset di pre-addestramento, ma fallisce nel generalizzare se applicato a dati mai prima utilizzati.
[5] Errori o variazioni nei dati che non rappresentano informazioni rilevanti o utili per l'addestramento del modello.

**Tab. 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 | Dataset contenente 1.000 miliardi di *token* di testo, 3,4 miliardi di |
| 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.
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[6].

**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[a], The Atlantic[b], Vox Media[c], News Corps[d], Dotdash Meredith[e], Financial Times[f], Le Monde e Prisa Media[g]. Questi accordi permettono a OpenAI di addestrare i 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 la propria azione giornalistica 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 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[h] 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[i], 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[j], editore di testate come la Repubblica e La Stampa. Tuttavia,

[6] 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.

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.

[a] OpenAI (2024, June 27), *Strategic content partnership with TIME*, OpenAI, https://openai.com/index/strategic-content-partnership-with-time/.

[b] OpenAI (2024, May 29), *A content and product partnership with The Atlantic*, OpenAI, https://openai.com/index/ enhancing-news-in-chatgpt-with-the-atlantic/.

[c] OpenAI (2024, May 29), *A content and product partnership with Vox Media*, OpenAI, https://openai.com/index/ a-content-and-product-partnership-with-vox-media/.

[d] OpenAI (2024, May 22), *NewsCorp and OpenAI sign landmark multiyear global partnership*, OpenAI, https://openai.com/index/ news-corp-and-openai-sign-landmark-multi-year-global-partnership/.

[e] *Dotdash Meredith Announces Strategic Partnership with OpenAI, Bringing Iconic Brands and Trusted Content to ChatGPT* (2024, May 7), Meredith Corporation MediaRoom, https://dotdashmeredith.mediaroom.com/2024-05-07-Dotdash-Meredith-Announces-Strategic-Partnership-with-OpenAI,-Bringing-Iconic-Brands-and-Trusted-Content-to-ChatGPT.

[f] OpenAI, *We're bringing the Financial Times' world-class journalism to ChatGPT*, OpenAI, https://openai.com/index/content-partnership-with-financial-times/.

[g] OpenAI (2024, March 8), *Global news partnerships: Le Monde and Prisa Media*, OpenAI, https://openai.com/index/global-news-partnerships-le-monde-and-prisa-media/.

[h] M.M. Grynbaum e R. Mac (2023, December 27), *New York Times sues OpenAI and Microsoft over use of copyrighted work*, «The New York Times», https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

[i] K. Robertson (2024, April 30), *Daily newspapers sue OpenAI and Microsoft over A.I*, «The New York Times», https://www.nytimes.com/2024/04/30/business/me-dia/newspapers-sued-microsoft-openai.html.

[j] OpenAI (2024, May 22), *OpenAI e GEDI annunciano una partnership strategica per rendere accessibili contenuti news in lingua italiana all'interno di ChatGPT*, OpenAI, https://openai.com/index/gedi/.

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[7] 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

[7] 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.

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[8]. Chiaramente, la possibilità di costruire un modello che generi un dataset sintetico utile successivamente alla creazione di un altro 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, riducendo 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 Tab. 2 compara i diversi dataset sotto tre dimensioni: costo, qualità e accessibilità.

**Tab. 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[9].

[8] 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 ad alcun individuo, non violano la tutela della privacy.
[9] In appendice B alcuni esempi di processi a cui sono sottoposti i dati e il loro impatto sui costi di sviluppo.

### *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.
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*[10] 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 dei prezzi dell'ordine di migliaia di dollari (Tab. 3).

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

**Tab. 3:** Numero di GPU impiegate per addestrare alcuni modelli di punta[11].

| 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, ossia il ricorso alle 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.

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 di IT (*Information Technologies*) 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*[11] ha tra gli obiettivi quello di costruire "fabbriche di intelligenza artificiale" (*AI factories*), ossia strutture che intendono includere supercomputer, *data center* e servizi di supercalcolo orientati all'intelligenza artificiale[12]. Queste strutture saranno aperte a utenti sia pubblici che privati, con condizioni di accesso dedicate a startup e piccole imprese.

### *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 *natural language processing 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 Fig. 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

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

[12] La Commissione europea ha promosso di recente la costruzione di *gigafactories* di IA, definite come «large-scale facilities dedicated to the development and training of next-generation AI models containing trillions of parameters».

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.

**Fig. 2:** Lavori con la maggior crescita netta, 2025-2030[13].

I salari per queste posizioni sono tra i più alti nel settore tecnologico: un report del 2024[14] 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 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 [15] 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 [16]. In risposta, molte

[13] *Future of Jobs Report 2025*, World Economic Forum.

[14] E.A. Bagyam (2024), *Analysis of data science job salaries from 2020 to 2024: trends and influencing factors*, Surya Publications.

[15] 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/.

[16] 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

organizzazioni adottano strategie di formazione continua, piani di sviluppo personalizzati e maggiore flessibilità lavorativa.

### *2.4. Energia*

L'addestramento dei FMs rappresenta uno dei processi più dispendiosi in termini energetici[17]. Per ridurre i consumi diverse aziende stanno adottando soluzioni come 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[18].

---

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 (ottobre 2024), *An assessment of occupational exposure to artificial intelligence in Italy*.

[17] Secondo D. Lazzaro 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 reti standard, con una perdita massima del 3% nelle prestazioni predittive.

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

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[19], 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 genera 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.
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.

### *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[20], utili a regolare i parametri e identificare eventuali problemi di *overfitting*

[19] La fase di inferenza è quel processo in cui un modello di *machine learning* addestrato usa la conoscenza acquisita per fare previsioni, prendere decisioni o generare risposte su dati nuovi e sconosciuti.
[20] 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.

e (ii) dataset di test[21], 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) rappresenta 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, B. Cottier et al. (2024)[22] 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 (Fig. 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.

**Fig. 3:** Andamento dei costi hardware ed energetici nel tempo per addestrare i modelli di frontiera[23].

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 (Fig. 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.

**Fig. 4:** Struttura dei costi per alcuni dei modelli di frontiera[24].

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

[21] Dati completamente nuovi che non vengono generati a partire dal dataset di addestramento.
[22] B. Cottier et al. (2024), *The rising costs of training frontier AI models*, https://arxiv.org/ abs/2405.21015.
[23] B. Cottier et al. (2024), *The rising costs of training frontier AI models*, https://arxiv.org/ abs/2405.21015.
[24] *Ibidem*.

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.

### *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, diversificati e bilanciati. Tuttavia, hanno caratteristiche distintive che li differenziano dai precedenti, cioè:

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[25]. 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 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 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[26]. Infine, in aggiunta figurano i costi per l'impiego di tecniche avanzate come il *Reinforcement Learnig From Human Feedback* (*RLHF*)[27] 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.

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

[26] In Appendice C una panoramica delle diverse tecniche di etichettatura.

[27] 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.