Una panoramica sui modelli di Intelligenza artificiale generativa[*]

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

1. *Foundation Models*: *Large* e *Small Language Models*

L'IAG 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*), ossia 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[1], riducendo la necessità di sviluppare modelli a partire 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.

[*] 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

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

L'incremento della potenza computazionale e della dimensione del dataset di addestramento tende a migliorare progressivamente le capacità del modello, seguendo un andamento prevedibile[2]. 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[3]. 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, come specificato 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 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 raggiungere prestazioni migliori. Il numero di parametri è importante perché consente al modello di generalizzare meglio, ossia 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 che utilizzano gli LLM come "cervello" per

[2] G. Sastry et al. (2024), *Computing power and the governance of artificial intelligence*, https://arxiv.org/abs/2402.08797.
[3] J. Wei et al. (2022), *Emergent abilities of large language models*, https://arxiv.org/abs/2206.07682.

svolgere *tasks* multimodali[4]. 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 (Fig. 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 abbia subito un'accelerazione significativa a partire dal 2022.

Fig. 1: Timeline Multimodal Large Language Models[5].

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 infrastrutture avanzate che non sono sempre accessibili a tutte le organizzazioni o applicabili in ogni scenario. È in questo contesto che emergono gli 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-4 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[6], un numero estremamente superiore se confrontato con un SLM come DistilBERT, il cui numero di parametri si aggira attorno ai 66 milioni[7]. 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

[4] S. Yin et al. (2024), *A survey on multimodal large language models*, «National Science Review», 11(12), https://arxiv.org/abs/2206.07682.
[5] *Ibidem*.
[6] J.A. Baktash e M. Dawodi (2023), *GPT4: A Review on Advancements and Opportunities in Natural Language Processing*, https://arxiv.org/abs/2305.03195.
[7] V. Sanh, L. Debut, J. Chaumond e T. Wolf (2019), *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*, https://arxiv.org/abs/1910.01108.

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, la Tab. 1 presenta un confronto tra le principali caratteristiche dei modelli.

Tab. 1: Confronto tra LLM e SLM.

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

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 è essenziale per inquadrare il ruolo dei modelli generali e specializzati, in particolare perché molti di quelli specializzati (strumenti fondamentali in settori come sanità e finanza) tendono a essere *open source*.

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 [8], consentendo l'integrazione dei suoi modelli proprietari all'interno dei suoi servizi[9]. 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 prodotti o servizi, spesso venendo monetizzati tramite abbonamenti o modelli c.d. *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 (*Application Programming Interface*), ossia 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 *Cloud Service Providers* (*CSP*). 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 offrire una soluzione scalabile e flessibile che può essere integrata in applicazioni aziendali, piattaforme di *customer service* e molto altro.

[8] OpenAI (2019, July 22), *Microsoft invests in and partners with OpenAI to support us building beneficial AGI*, OpenAI, https://openai.com/index/microsoft-invests-in-and-partners-with-openai/.
[9] Nel mese di ottobre 2025, Microsoft ha acquisito circa il **27%** della società a scopo di lucro, denominata **OpenAI Group Pbc**, mentre una quota del 26% sarà detenuta dalla OpenAI Foundation, l'organizzazione no profit originaria.

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 business 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.

2.1. Sicurezza

Il controllo centralizzato degli sviluppatori nei modelli *closed source* permette di applicare delle restrizioni che riducono il rischio di manomissione dei dati e minimizzano la possibilità di utilizzo improprio del modello. Al contrario, nei modelli *open source*, una volta resi pubblici i pesi, gli sviluppatori perdono quasi completamente il controllo sull'uso che ne viene fatto a valle. Seppur quindi ci siano dei benefici nel permettere di visionare e modificare il modello, risulta difficile prevedere limitazioni efficaci per evitare che gli utenti ne facciano un uso improprio o malevolo. Inoltre, il rilascio dei pesi costituisce un'azione irreversibile in quanto, nonostante lo sviluppatore possa interrompere l'accesso al modello, non può né revocare le copie dei pesi ormai create, 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. Innovazione

Un'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'accesso più ampio e l'inferenza locale permette agli sviluppatori di addestrare i modelli su dati proprietari favorendo una personalizzazione più ampia, che non ha eguali nei modelli *closed source* a causa delle limitazioni imposte dallo sviluppatore, e permettono 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 accesso ai *feedback* degli utilizzatori del modello, che rappresentano una fonte di dati estremamente importante per il miglioramento continuo dello stesso.

2.3. Trasparenza

Per analizzare le differenze di trasparenza tra le due categorie di modelli è possibile far affidamento al *Foundation Model Transparency Index*, un *framework* elaborato dalle università di Stanford University, Harvard e Princeton University [vedi R. Bommasani 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 (Fig. 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).

Fig. 2: Foundation Model Transparency Index Scores, Maggio 2024[10].

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 source* ottengono un punteggio pari o superiore in 18 su 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 ottengono 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, 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.

[10] R. Bommasani et al. (2024), *The 2024 Foundation Model Transparency Index*, https://arxiv.org/abs/2407.12929.

3. Modelli generali e specializzati

La presente sezione ha lo scopo di approfondire le principali caratteristiche e utilizzi dei modelli generali e specializzati. I modelli generali, precedentemente definiti come *Foundation Models* (FMs), si contraddistinguono per la loro versatilità e capacità di adattamento a una vasta gamma di compiti. Già descritti nei capitoli precedenti, questi modelli possono essere impiegati in molteplici contesti applicativi, spesso anche al di là delle finalità originarie per cui sono stati progettati. Grazie al loro pre-addestramento su larga scala, vengono utilizzati come base per la costruzione di sistemi più specializzati. Attualmente, i modelli generali comprendono principalmente LLM e MLLM, e trovano applicazione in settori come medicina, finanza, chimica, istruzione, programmazione e molti altri. In Tab. 2 alcuni dei principali modelli generali nel mercato:

Tab. 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 di minori dimensioni in modo da poterli adattare a specifici contesti o settori. Ad esempio, un dataset contenente documenti legali potrebbe essere utilizzato per migliorare la capacità di un modello di fornire consulenze o generare documenti di carattere giuridico. 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 contenuti 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 implementati 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[11].

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* personalizzati. 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, CoGen[12] è un'architettura che utilizza modelli generativi per convertire video di programmazione in sessioni di tutoraggio interattivo e adattivo, guidando lo studente nell'acquisizione graduale delle competenze attraverso strategie pedagogiche strutturate. Tali approcci aprono prospettive concrete per l'implementazione di

[11] L'Osservatorio Astrid sulle dinamiche dell'intelligenza artificiale ha avviato specifici gruppi di ricerca sugli impatti dell'IA in particolari "mercati", quali la sanità, l'*education*, l'agro-alimentare, la pubblica amministrazione, il diritto d'autore.
[12] W. Li, R. Pea, N. Haber e H. Subramonyam (2025), *CogGen: A learner-centered generative AI architecture for intelligent tutoring with programming videos*, https://arxiv.org/abs/2506. 20600.

ambienti formativi scalabili e responsivi, capaci di superare i limiti dei modelli educativi tradizionali, soprattutto in termini di accessibilità, flessibilità e personalizzazione dell'insegnamento.

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 modelli 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. Ad esempio, Netflix ha recentemente annunciato di aver utilizzato l'IAG per la produzione di una scena della serie TV El Eternauta [13], affermando che la sequenza di effetti visivi è stata completata dieci volte più velocemente rispetto all'utilizzo dei tradizionali strumenti di VFX, ottenendo al contempo dei risultati sorprendenti. Nello stesso contesto, Amazon ha recentemente annunciato lo sviluppo di uno strumento di IAG per il doppiaggio di film e serie tv[14]. 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 [15] 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 (PNG) di dialogare in tempo reale con i giocatori, producendo conversazioni dinamiche e contestualizzate, invece delle interazioni predefinite tipiche dei videogiochi tradizionali.

[13] M. Sweney (2025, July 19), *Netflix uses generative AI in one of its shows for first time*, «The Guardian», https://www.theguardian.com/media/2025/jul/18/ netflix-uses-generative-ai-in-show-for-first-time-el-eternauta.

[14] A. Staff (2025, March 5), *Prime Video tests AI-powered dubbing in English, Spanish*, https: //www.aboutamazon.com/news/entertainment/prime-video-ai-dubbing-english-spanish.

[15] NVIDIA ACE for Games, *NVIDIA Developer*, https://developer.nvidia.com/ ace-for-games.

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 IAG. Ad esempio, SyntheMol[16], sviluppato congiuntamente da Stanford Medicine e dalla McMaster University, è un sistema di intelligenza artificiale generativa finalizzato a supportare il processo di scoperta e progettazione di nuovi composti chimici, in particolare a scopo farmaceutico. Il modello è in grado non solo di generare strutture molecolari potenzialmente efficaci, ma anche di fornire indicazioni plausibili per la loro sintesi sperimentale. In tal modo, SyntheMol consente di accelerare in modo significativo le fasi iniziali della ricerca preclinica, riducendo il numero di esperimenti necessari e orientando i ricercatori verso soluzioni chimiche più promettenti. Nel campo dell'astrofisica, invece, modelli come AstroMLab-4[17] stanno ridefinendo l'approccio all'analisi dei dati astronomici. Si tratta di un LLM, specificamente addestrato su letteratura scientifica e dati provenienti da osservazioni astronomiche, che ha dimostrato una notevole capacità di interpretare e generare contenuti pertinenti al dominio.

Per illustrare in modo sintetico l'ampiezza e la varietà di applicazioni dei modelli di IAG nei diversi settori, la Tab. 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 di minori dimensioni in termini di parametri rispetto ai modelli generali e che la loro natura sia prevalentemente *open source*.

[16] *Generative AI develops potential new drugs for antibiotic-resistant bacteria* (2025, July 1), «News Center», https://med.stanford.edu/news/all-news/2024/03/ai-drug-development.html
[17] D.H. Tijmen, Y. Ting, T. Ghosal, T.D. Nguyen, A. Accomazzi, E. Herron, V. Lama, R. Pan, A. Wells, e N. Ramachandra (2025, May 23), *AstroMLab 4: Benchmark-Topping Performance in Astronomy Q&A with a 70B-Parameter Domain-Specialized Reasoning Model*, arXiv.org. https://arxiv.org/abs/2505.17592.

Tab. 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 trasformando significativamente il panorama tecnologico e industriale, ridefinendo la competizione nei settori esistenti e favorendo la nascita di nuovi mercati. 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.