



UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA

Dipartimento di Informatica, Sistemistica e Comunicazione

Corso di Laurea Magistrale in Data Science

ADVANCED ANALYTICS PER LA GESTIONE DEI PROCESSI DI DISTRIBUZIONE GAS

Relatore: Prof. Silvio Gerli

Correlatori: Dott. Claudio Calamita, Dott. Giovanni Ingrande

Tesi di Laurea Magistrale di:

Emanuela Elli

Matricola 892901

Anno Accademico 2022-2023

testo a destra

Abstract

Il presente studio si propone di esplorare l'applicabilità dei modelli generativi di intelligenza artificiale nel contesto aziendale nel settore energetico, con particolare attenzione all'estrazione di valori di specifici KPI presenti all'interno di *report ESG* (*Enviroment, Social* e di *Governance*) pubblicati dalle aziende. L'obiettivo è valutare se tali modelli possano essere utilizzati con successo per condurre delle analisi efficaci, fornendo alle aziende uno strumento utile per migliorare la propria competitività sul mercato rispetto ai concorrenti. Per raggiungere tale obiettivo, sono stati utilizzati quattro modelli di *generative IA* differenti: GPT3.5, DollyV2, Llama2 e Gemini. Ad ogni modello sono stati sottoposti i medesimi *report ESG*, dopodiché è stato richiesto di estrarre i valori di una lista di KPI ed infine di inserire i valori trovati all'interno di dizionari JSON. Successivamente, tali estrazioni sono state confrontate con un *ground truth*, ottenuto manualmente, al fine di valutarne le prestazioni.

I risultati della ricerca hanno evidenziato che il modello **Gemini** e **Llama2** presentano delle peculiarità complementari: mentre **Gemini** è caratterizzato da risposte prolixe ma accurate, **Llama2** risulta essere più rigoroso nel fornire *output* correttamente strutturati ma meno preciso nella comprensione del testo. Di conseguenza, è stato implementato un approccio che combina gli *output* di entrambi i modelli, consentendo a **Gemini** di estrarre solamente la porzione di testo relativa al KPI esaminato ed a **Llama2** di fornire il dizionario JSON, per il KPI richiesto, all'interno del testo estratto.

Al termine della ricerca, si è constatato che il modello GPT-3.5 ha dimostrato maggiore efficacia nel bilanciare la precisione dei valori estratti e gli errori commessi. Tuttavia, si intravedono prospettive incoraggianti per migliorare le prestazioni attraverso l'integrazione di più modelli. Lo sviluppo di ulteriori studi ed adattamenti del modello potrebbero portare a risultati ancora più soddisfacenti. Pertanto, questo studio rappresenta un piccolo avanzamento verso l'introduzione di tecnologie innovative nel settore energetico, con l'obiettivo di migliorare le *performance* aziendali. Considerando le necessarie precauzioni in termini di sicurezza informatica ed adattabilità al contesto specifico, i modelli di intelligenza artificiale generativa hanno mostrato un notevole potenziale che può essere sfruttato dalle aziende per distinguersi nel contesto della finanza sostenibile, apportando benefici sia dal punto di vista economico che temporale.

Indice

Introduzione	1
1 Contesto e stato dell'arte	5
1.1 Mercato del gas	6
1.1.1 Approvvigionamento	6
1.1.2 Infrastruttura	7
1.1.3 Vendita	10
1.1.4 Decreto Legislativo n. 164/00	11
1.1.5 Situazione dei consumi nel 2022 (ARERA)	13
1.2 Environmental, Social and Governance (ESG)	16
1.2.1 Contesto storico legato a ESG	19
1.2.2 Rischi ESG	20
1.2.3 Rating ESG	21
1.2.4 AI legata al calcolo della ESG	25
1.2.5 Adeguamento ai fattori ESG delle imprese italiane	29
1.2.6 IA sostenibile	33
1.3 Proposta europea per un quadro giuridico sull'IA (AI Act)	36
1.3.1 Quadro teorico	37
1.3.2 AI Act: contesto globale	39
2 Descrizione del progetto	41
2.1 Fondamenti Tecnologici: Large Language Models	42
2.1.1 Storia dei Large Language Model	44
2.1.2 Architettura generale dei Large Language Models	48
2.1.3 Tecniche di applicazione degli LLM	55
2.1.4 Applicazioni dei LLM	60
2.1.5 Rischi correlati all'utilizzo di LLM	62
2.2 Obiettivo del progetto	63
3 Modelli e strumenti utilizzati	66
3.1 Architettura	66
3.2 Modelli utilizzati	69
3.2.1 GPT	70
3.2.2 Dolly	76
3.2.3 LLAMA	79
3.2.4 Gemini	83
4 Implementazione	86
4.1 Data Acquisition	87
4.2 Data Preparation	91
4.3 Model and Prompt Engineering	91
4.3.1 GPT3.5 Turbo	95
4.3.2 DollyV2-7b	96

4.3.3	LLAMA2-7b	97
4.3.4	Gemini	98
4.4	Post Processing	100
4.5	Data Storage	100
4.6	Data Comparison (Ground Truth)	101
5	Considerazioni e visualizzazione dei risultati	103
5.1	Considerazioni	104
5.2	Visualizzazione dei risultati	107
6	Conclusioni e sviluppi futuri	119
	Glossario	124
	Bibliografia e sitografia	134

Introduzione

Negli ultimi anni, la finanza sostenibile ha assunto un ruolo centrale nell’ambito delle dinamiche economiche e degli investimenti, influenzando sia le strategie aziendali che le scelte degli investitori. Questo approccio finanziario va oltre la mera considerazione degli aspetti economici e finanziari, includendo anche quelli ambientali e sociali e rivestendo un’importanza cruciale nella creazione di valore a lungo termine, nella gestione dei rischi e nel promuovere un impatto positivo sulla società e sull’ambiente. L’interesse per la finanza sostenibile si è diffuso al di là del settore finanziario, coinvolgendo una vasta gamma di attori, tra cui governi, aziende, organizzazioni della società civile e la popolazione stessa. Questo interesse è stato catalizzato dalle sfide globali urgenti come il cambiamento climatico, la disuguaglianza sociale e l’esaurimento delle risorse naturali. Le istituzioni finanziarie e gli investitori rivestono un ruolo chiave nel promuovere la sostenibilità e nell’indirizzare la transizione verso un’economia più inclusiva, con basse emissioni di carbonio e più resistente ai cambiamenti climatici. La crescente attenzione sulla finanza sostenibile al di fuori del settore finanziario mette in luce l’interconnessione tra finanza, economia e benessere della società e dell’ambiente. Questo sottolinea l’importanza di allineare le pratiche finanziarie agli obiettivi di sviluppo sostenibile, favorendo investimenti responsabili per contribuire ad un futuro più equo e sostenibile. I risparmiatori, in questo contesto, possono optare per investire in imprese che non solo generano un rendimento economico ma hanno anche un impatto positivo sull’ambiente o sulla società. Ad esempio, possono scegliere di investire in aziende che adottano pratiche responsabili nell’uso delle risorse naturali e nella tutela degli ecosistemi, che garantiscono condizioni di lavoro sicure e rispettose dei diritti dei lavoratori, o che operano in conformità con principi etici e buone pratiche di *governance* aziendale.

Nel contesto degli investimenti, il concetto di “sostenibilità” viene valutato attraverso gli indicatori chiamati *rating ESG* (*Environment, Social, Governance*), che forniscono un giudizio sintetico sul livello di sostenibilità ambientale, sociale e di governo societario di emittenti come imprese, Stati o organizzazioni sovranazionali, così come di titoli o strumenti di investimento collettivo, come OICR (Organismi di Investimento Collettivo del Risparmio) e gli ETF (Exchange-Traded Fund). Tali *rating* vengono assegnati da agenzie specializzate, che li elaborano basandosi su analisi condotte utilizzando informazioni non finanziarie pubblicate dalle imprese stesse (nota come “dichiarazione non finanziaria”) ed altre fonti come questionari, banche dati e notizie. Tali informazioni riguardano i criteri di sostenibilità adottati nelle pratiche gestionali e nei progetti di investimento delle aziende. Oltre ai *rating ESG*, che rappresentano un’indicazione sintetica della sostenibilità complessiva, le agenzie possono fornire dati dettagliati su singoli aspetti della sostenibilità aziendale, come le emissioni di carbonio o il consumo di acqua. Tuttavia, mancano ancora standard internazionali condivisi per la valutazione della sostenibilità. Quindi attualmente esistono diversi concetti e misure utilizzati per definire cosa sia “sostenibile” nell’ambito dell’attività economica. Nonostante questa mancanza di standardizzazione, i punteggi ESG sono ampiamente utilizzati nel settore finanziario per selezionare strumenti di investimento, costruire portafogli e creare indici di mercato che sono definiti “sostenibili” o sono etichettati come “ESG”. Per le imprese, rappresentare in modo efficace il

proprio approccio alla sostenibilità e delineare il processo di transizione può essere una leva strategica importante, consentendo loro di comunicare agli *stakeholder* come intendono affrontare gli impatti e le opportunità legate alla sostenibilità.

A tale scopo, il 14 dicembre 2022 è stata introdotta la Direttiva (UE) 2022/2464 [1], nota come Corporate Sustainability Reporting Directive (CSRD). La CSRD ha sostituito il termine “dichiarazione di carattere non finanziario” (DNF), utilizzato nella normativa precedente (Non Financial Reporting Directive - NFRD), con “informazioni sulla sostenibilità”. Questo cambiamento va oltre un semplice cambiamento di terminologia, ma riflette l’idea che le informazioni sulla sostenibilità influenzano direttamente la situazione finanziaria dell’azienda. Di conseguenza, la relazione sulla sostenibilità è diventata parte integrante del *report* finanziario annuale, con un allineamento dei processi di produzione dell’informatica ESG e di quella finanziaria.

Un aspetto cruciale da considerare è che la produzione e l’uso di energia, fondamentali per tutte le attività umane, sono le principali fonti di emissioni di carbonio nell’atmosfera a livello globale, soprattutto a causa dell’utilizzo di fonti di energia fossile come carbone, gas e petrolio. L’eccessiva emissione di CO_2 , riconosciuta dalla comunità scientifica internazionale, è la principale causa del cambiamento climatico, che mette in pericolo l’equilibrio del pianeta così com’è conosciuto oggi. Questa consapevolezza è sempre più diffusa anche tra il pubblico ed i consumatori, che richiedono un impegno maggiore verso la sostenibilità da parte delle aziende. Questa crescente consapevolezza spiega la crescente domanda di dati ESG, tempestivi e trasparenti, specialmente da parte degli investitori che operano nel settore dell’energia e dei servizi pubblici. Risulta quindi di estrema importanza per le aziende del settore essere consapevoli del proprio *rating* ESG rispetto ai concorrenti al fine di mantenere la propria competitività sul mercato degli acquirenti. Essere a conoscenza del proprio posizionamento, rispetto alle aziende *competitor*, può consentire alle aziende di identificare i propri punti di forza e di debolezza, nonché di individuare opportunità per migliorare le proprie pratiche aziendali ed allinearsi meglio alle aspettative del mercato e dei consumatori in materia di sostenibilità. In un contesto in cui l’attenzione alla sostenibilità è sempre più rilevante, il mantenimento di un *rating* ESG elevato può essere un vantaggio competitivo significativo per le aziende del settore, consentendo loro di attrarre investimenti e clienti, nonché di ridurre i rischi legati alla reputazione ed alle prestazioni finanziarie.

Il presente studio si propone di esaminare l’utilizzo di modelli avanzati di intelligenza artificiale generativa, noti come Generative AI, al fine di ottimizzare l’analisi dei Key Performance Indicators (KPI) legati al *rating* ESG. Questa ricerca rappresenta un fronte innovativo nel campo dell’intelligenza artificiale e coinvolge modelli sviluppati da *leader* del settore tecnologico. Tra i protagonisti di questo studio vi è il modello GPT-3.5 [2], una versione avanzata del noto GPT (Generative Pre-trained Transformer) sviluppato da OpenAI. La sua complessa architettura neurale consente la generazione di testi coerenti e di alta qualità su svariati argomenti, rendendolo particolarmente adatto all’analisi di *report* complessi. In aggiunta è presente DollyV2 [3], sviluppato da Databricks, emerge come un modello rilevante in quanto il primo Large Language Model (LLM) completamente *open-source* addestrato con un costo contenuto ma con capacità di interazione

comparabili a quelle di ChatGPT. Questa innovazione rappresenta un notevole progresso nell’ambito dell’intelligenza artificiale accessibile e flessibile. Un ulteriore modello è LLAMA2 [4], sviluppato da Meta, che offre un approccio unico poiché è addestrato su set di dati personalizzati, come base di dati di ricerca e documentazione *software*, rendendolo particolarmente adatto per progetti che richiedono una conoscenza specialistica. Infine, Gemini [5], creato da Google, rappresenta un’ulteriore evoluzione nel campo dei modelli di intelligenza artificiale, poiché è addestrato non solo su testi ma anche su immagini, video e audio, conferendogli una potenza multidimensionale ed una capacità più poliedrica. L’obiettivo principale di questo studio è valutare le prestazioni di tali modelli per capacità di estrarre i valori dei KPI dai *report* ESG, al fine di fornire alle aziende uno strumento efficace per migliorare la propria analisi e gestione dei KPI relativi alla sostenibilità.

Il presente lavoro è stato reso possibile grazie allo svolgimento di un tirocinio curricolare presso la società Power Reply Srl [6], un’azienda specializzata nel settore “Energy & Utilities” facente parte del gruppo Reply. Power Reply si distingue per la sua profonda conoscenza del mercato e dei processi aziendali, insieme alla competenza nell’implementazione e gestione di soluzioni informatiche per supportare le attività chiave dei clienti. Il suo portafoglio di servizi include consulenza aziendale, consulenza IT e strategia digitale, mirando a migliorare gli investimenti nella tecnologia dell’informazione e ad aderire ai percorsi di transizione energetica e decarbonizzazione. La società conta su un *team* di oltre 200 professionisti impegnati in progetti che coprono l’intera catena del valore del settore, tra cui analisi e strategie di *business* e IT, progetti di *digital strategy*, pianificazione e progettazione di programmi di trasformazione, gestione dei dati ed integrazione dei sistemi. Durante il tirocinio, è stato possibile prender parte ad un progetto finalizzato a soddisfare le esigenze di una rinomata società operante nel settore della distribuzione del gas. In particolare, data l’importanza crescente e significativa dei *report* e del *rating* ESG, è emersa l’esigenza di raccogliere automaticamente i valori di specifiche informazioni, presenti in tali documenti e pubblicati dalle società concorrenti, utilizzando le nuove tecnologie sviluppate nel campo della Generative AI.

L’obiettivo è, quindi, investigare l’utilizzo di strumenti di Generative AI all’interno del contesto ESG, al fine di migliorare l’efficienza e la tempestività nella raccolta di informazioni utili per comprendere, studiare e migliorare maggiormente la posizione dell’azienda, in relazione a tematiche cruciali, rispetto ai *competitor* presenti nel mercato. In particolare, il presente studio mira a definire un concetto fondamentale noto come “*benchmark*”. Questo termine delinea una misura di confronto, un punto di riferimento utilizzato per valutare le prestazioni di una determinata entità, come un’azienda o un processo, rispetto ad altre entità simili o al mercato nel suo complesso. Nel contesto della presente ricerca, il *benchmark* viene creato utilizzando i dati estratti dalle aziende concorrenti. Questo consentirà di confrontare i valori dei KPI interni, dell’azienda cliente, con le informazioni ottenute dalle società esterne, consentendo un’analisi comparativa delle prestazioni. Inoltre, verrà utilizzato il *benchmark* per condurre simulazioni di scenario, ad esempio esaminando l’impatto di una riduzione percentuale su uno specifico KPI, come le emissioni di CO_2 , o in relazione ad altri KPI correlati. Tali simulazioni sono fondamentali per comprendere meglio come le azioni influenzino le prestazioni dell’azienda stessa rispetto al mercato del settore.

La struttura di questo lavoro di tesi si articola in sei capitoli distinti. Il primo capitolo si propone di fornire una panoramica approfondita del contesto e dello stato attuale del mercato del gas. In particolare, vengono esaminate le dinamiche legate all'approvvigionamento, all'infrastruttura, alla vendita ed alla cornice normativa vigente al fine di offrire una comprensione completa del contesto di ricerca. Inoltre, vengono esplorati il ruolo dell'Environmental, Social and Governance (ESG) nel settore, attraverso un'analisi del contesto storico, dei rischi associati, dei *rating* ESG e degli sforzi di adeguamento da parte delle imprese italiane. Il capitolo si conclude con un'analisi della proposta europea per un quadro normativo sull'Intelligenza Artificiale (AI Act), fornendo così una panoramica dello stato attuale della regolamentazione delle tecnologie legate all'intelligenza artificiale. Il secondo capitolo si concentra sulla descrizione del progetto, offrendo una trattazione dei fondamenti tecnologici che ne costituiscono la base, ovvero i Large Language Models (LLM). Si esamina la loro storia, l'architettura generale, le tecniche di applicazione e i rischi associati al loro utilizzo. Il capitolo si conclude delineando l'obiettivo principale della tesi e specificando il ruolo che tale tecnologia assume all'interno del contesto precedentemente introdotto. Il terzo capitolo espone i modelli e gli strumenti adoperati nel corso del progetto, fornendo una dettagliata analisi dell'architettura e dei modelli specifici impiegati, tra cui GPT, Dolly, LLAMA e Gemini. Il quarto capitolo affronta l'implementazione pratica del progetto, esaminando le diverse fasi operative, quali l'acquisizione e la preparazione dei dati, l'ingegneria dei modelli e la gestione post-elaborazione, che comprende anche la creazione di un *dataset* ad hoc per la comparazione dei modelli. Infine, il quinto capitolo presenta le considerazioni e le visualizzazioni dei risultati ottenuti, seguite dalle conclusioni e dalle prospettive future per il lavoro svolto.

1. Contesto e stato dell'arte

Nel presente capitolo, si procederà con una dettagliata descrizione del contesto legato alla ricerca in oggetto. In particolare, vengono esaminate le diverse fasi che costituiscono l'intera filiera del gas naturale: dalla sua estrazione fino alla sua distribuzione ai consumatori finali. Successivamente, verrà posta una particolare attenzione al decreto legislativo che ha introdotto la liberalizzazione dei mercati energetici in Italia. Questo aspetto riveste un'importanza fondamentale, in quanto ha avuto un impatto significativo sulle dinamiche del settore del gas all'interno del paese. Pertanto, tale decreto sarà esaminato al fine di comprendere come sia stata guidata la liberalizzazione del mercato e come abbia influenzato il funzionamento delle imprese del settore. Al fine di ottenere una visione esaustiva del quadro generale, sarà fornita una panoramica dei consumi di gas a livello mondiale, con particolare riferimento all'Italia. Ciò includerà una valutazione delle tendenze del consumo, nonché una considerazione dei fattori economici, ambientali e geopolitici che esercitano un'influenza sulla domanda del gas.

In seguito, si procederà con un'analisi approfondita dei tre pilastri fondamentali che costituiscono la struttura portante dell'approccio alla valutazione e promozione della sostenibilità nelle organizzazioni aziendali, comunemente noti con l'acronimo ESG, che rappresentano le tre dimensioni cruciali: *Environment* (ambiente), *Social* (sociale) e *Governance* (temi economici). Saranno, quindi, esplorati i fattori che costituiscono tale valutazione ma anche il contesto storico che ha portato alla concettualizzazione dell'approccio ESG, insieme ai rischi ed ai vantaggi che vi sono connessi. Inoltre, verrà dedicata particolare attenzione all'importanza e all'incidenza dell'intelligenza artificiale nel processo di calcolo e valutazione di tali fattori. L'introduzione di soluzioni basate sull'intelligenza artificiale ha rivoluzionato la capacità di monitorare e misurare l'impatto delle attività aziendali sulla sostenibilità, consentendo una valutazione più precisa e tempestiva. Tuttavia, tale innovazione comporta, altresì, nuove sfide e questioni che verranno esaminate nel contesto attuale dell'industria.

In conclusione, sarà condotta un'analisi del processo legislativo attualmente in corso presso il Parlamento Europeo, il quale mira a promuovere l'emanazione del "Artificial Intelligence Act" (AI Act). Tale atto legislativo, in fase di adozione, assume un ruolo di rilievo nell'ambito della regolamentazione delle tecnologie legate all'intelligenza artificiale all'interno dell'Unione Europea. L'analisi si focalizza sulla delineazione del quadro concettuale e sulla valutazione del contesto globale in cui tali regolamentazioni trovano collocazione, prestando particolare attenzione alla conseguente reazione degli Stati Uniti d'America. In tale contesto, saranno esplorate le implicazioni, le sfide e le opportunità che emergono dalla creazione di un apparato giuridico specificamente dedicato all'intelligenza artificiale.

1.1 Mercato del gas

Negli ultimi decenni, il gas naturale è emerso come la principale fonte di combustibile fossile impiegata per la produzione di energia elettrica ed energia termica, con una crescita significativa a partire dagli anni '80. In particolare, fino a poco più di un decennio fa, l'industria del gas naturale si contraddistingueva per una stretta integrazione tra le aziende, lungo l'intera catena di produzione e distribuzione. Inoltre, l'espansione delle infrastrutture avveniva in concomitanza con la stipula di contratti di vendita, dando luogo così alla creazione di complesse reti di trasporto a lunga distanza tra i paesi produttori e quelli consumatori. Questa strategia ha contribuito notevolmente a promuovere la diffusione del gas naturale all'interno del settore energetico di numerosi paesi.

In origine il gas naturale era principalmente destinato all'uso domestico ma, con il decorrere del tempo, ha esteso gradualmente il proprio impiego per includere anche le necessità del settore industriale e, più recentemente, del comparto dell'energia elettrica. Tuttavia, è importante evidenziare che la distribuzione di questa preziosa risorsa non è uniforme a livello globale. In Italia, ad esempio, soltanto il 7% circa del fabbisogno di gas naturale proviene da giacimenti nazionali, mentre la restante parte deve essere importata da nazioni estere [7]. Di conseguenza, la presenza di efficaci sistemi di trasporto che collegano le aree di estrazione con quelle di consumo diventa di fondamentale importanza.

La filiera del gas naturale comprende l'intero processo operativo, come mostrato in **Figura 1.1**, ovvero inizia con la fase di approvvigionamento, passando attraverso il trasporto tramite gasdotti o navi, per poi continuare con il processo di stoccaggio e concludersi con la vendita all'ingrosso o al dettaglio. Tale ciclo coinvolge tutte le fasi, dall'estrazione del gas, proveniente da giacimenti produttivi, fino al momento in cui il gas viene effettivamente utilizzato.

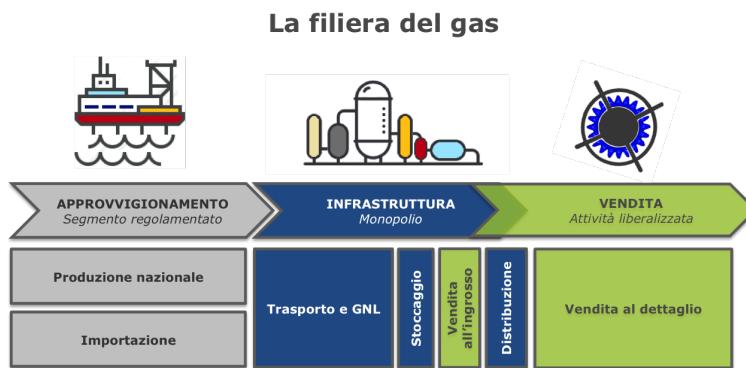


Figura 1.1. Struttura dell'industria del gas articolata su più fasi integrate.

Fonte: [Atena Luce Gas e Servizi](#)

Di seguito vengono descritte nel dettaglio le principali fasi della filiera in questione.

1.1.1 Approvvigionamento

Non considerando ciò che attiene l'esplorazione e la produzione del gas, la fase più a monte della filiera risulta essere l'approvvigionamento. Essa comprende le attività volte al re-

perimento della materia prima necessaria per il soddisfacimento del fabbisogno energetico.

In Italia, l'approvvigionamento di gas naturale al sistema nazionale avviene attraverso due principali canali: la produzione interna e l'importazione. Tuttavia, come precedentemente anticipato nella sezione 1.1, la produzione interna di gas naturale non è sufficiente a coprire l'intera domanda del paese, rappresentando solo il 10% del consumo complessivo. Di conseguenza, il restante 90% deve essere importato da fonti esterne, non solo per soddisfare le esigenze domestiche ma anche, e soprattutto, per le necessità industriali. Fino a circa un decennio fa le forniture di gas utilizzate in Italia provenivano da diverse nazioni tra cui, principalmente, Algeria, Libia e Russia. In aggiunta, il gas naturale può essere importato in forma liquida, conosciuta come *Gas Naturale Liquefatto* (GNL), tramite trasporto marittimo. Successivamente, attraverso l'impiego di una serie di processi chimici, il GNL viene riportato alla sua forma naturale rendendolo idoneo per l'utilizzo nelle reti nazionali di trasporto e distribuzione, come mostrato nella **Figura 1.2**.

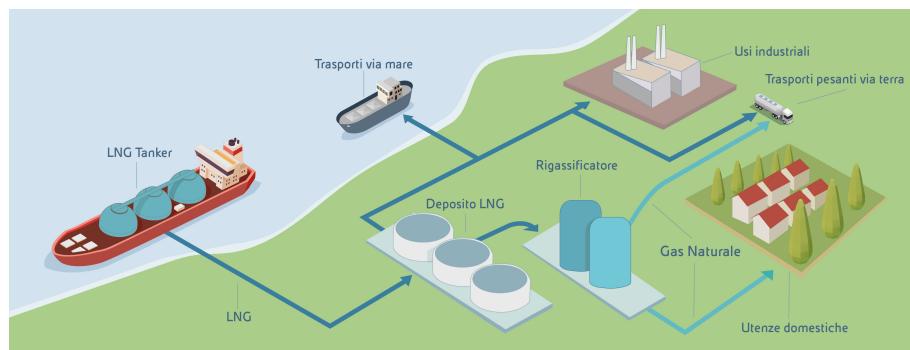


Figura 1.2. Percorso di rigassificazione del *Gas Naturale Liquefatto*.

Fonte: Ente Nazionale Idrocarburi (ENI)

Di conseguenza, l'importanza dell'approvvigionamento di gas naturale e delle sue diverse fonti di origine, nonché dei processi di trasformazione, è fondamentale per garantire una fornitura continua ed efficiente di energia al territorio di interesse.

1.1.2 Infrastruttura

Dopo la fase di estrazione, il gas naturale necessita di un processo di trasporto che implica il trasferimento della risorsa dai siti di produzione, o stoccaggio, ai punti di connessione con la rete a bassa pressione. Il trasporto internazionale del gas richiede la considerazione di numerosi fattori, inclusi fattori di natura economica, geopolitica, tecnologica ed anche ambientale. Non sorprende, infatti, che la pianificazione e la costruzione dei grandi gasdotti abbiano spesso scatenato tensioni tra vari attori internazionali e, in alcuni casi, addirittura innescato conflitti bellici.

Esistono due principali modalità di trasporto del gas:

1. **Gasdotti:** Questa metodologia prevede il trasferimento attraverso una rete di condutture interrate in acciaio all'interno delle quali il gas, in forma gassosa, viene spinto a diverse pressioni (alta, media e bassa) per coprire distanze che possono

estendersi su centinaia di chilometri. I gasdotti possono essere, inoltre, classificati in diverse tipologie in base alla loro copertura geografica ed al raggio d'azione. Le principali tipologie di gasdotti includono:

- (a) *Gasdotti Trasversali*: Essi collegano regioni o paesi differenti attraversando vaste aree geografiche. Sono progettati per il trasporto a lunga distanza per grandi volumi di gas.
- (b) *Gasdotti di Rete Locale*: Si tratta di gasdotti più piccoli che servono principalmente a distribuire il gas naturale all'interno di una specifica area geografica, come una città o una regione metropolitana.
- (c) *Gasdotti Internazionali*: Ovvero gasdotti che attraversano i confini nazionali e collegano paesi diversi, consentendo il trasporto transfrontaliero del gas naturale. Spesso sono parte di reti internazionali di approvvigionamento energetico.
- (d) *Gasdotti di Trasporto Regionale*: Questi gasdotti collegano diverse regioni all'interno di uno stesso paese o di una determinata area geografica. Utilizzati per distribuire il gas naturale su lunghe distanze all'interno di una nazione.
- (e) *Gasdotti di Distribuzione Locale*: Tali gasdotti coprono una piccola area geografica e sono destinati a consegnare il gas naturale agli utenti finali, come case e aziende, all'interno di comunità specifiche.

Di conseguenza, la scelta della tipologia di gasdotto dipende dalle esigenze di trasporto, dalla distanza da coprire sul territorio e dalla quantità di gas da spostare. Ogni tipologia di gasdotto è progettata per scopi specifici e viene dimensionata in base a tali esigenze.

2. **Trasporto via nave**: In questo caso, il gas naturale viene convertito in *Gas Naturale Liquefatto* (GNL) attraverso un processo chimico che lo raffredda a temperature estremamente basse per mantenerlo allo stato liquido, solitamente intorno ai -160°C circa. Successivamente, il GNL viene imbarcato su navi metaniere e, una volta giunto a destinazione, viene riportato allo stato gassoso mediante il processo di rigassificazione. Questo processo consiste nel riscaldare il GNL utilizzando un dispositivo noto come "vaporizzatore". A meno che non siano necessari trattamenti aggiuntivi per rimuovere eventuali impurità, il gas rigassificato viene iniettato nella rete di trasporto nazionale, come mostrato nella **Figura 1.2**.

Il mercato internazionale del gas naturale in Europa ha registrato uno sviluppo significativo dagli anni '70, in risposta alla necessità di spostare grandi quantità di gas dai luoghi di produzione ai luoghi di consumo. Questo ha portato alla realizzazione dei primi gasdotti tra Russia, Germania ed Italia. Successivamente, il sistema interconnesso europeo continua ad espandersi estendendosi dal Mare del Nord e dal Baltico fino al Mediterraneo, dall'Europa orientale e dalla Siberia fino all'Atlantico. Questo ha permesso di sfruttare riserve provenienti da diverse aree di estrazione, di diversificare le vie di trasporto e di facilitare gli scambi internazionali. Il sistema, inoltre, è stato progettato per garantire la fornitura di gas naturale anche in situazioni di approvvigionamento problematico o picchi di consumo elevati.

Come precedentemente esposto, all'interno dei singoli paesi, il gas naturale viene distribuito attraverso reti di trasporto nazionali, che variano a seconda delle specifiche esigenze di ciascun territorio. In Italia uno dei principali operatori di infrastrutture energetiche è la società *Snam Rete Gas* che attualmente gestisce oltre il 90% della rete di trasporto del gas sul territorio (la percentuale esatta potrebbe variare nel tempo a seconda di vari fattori, come acquisizioni, vendite di attività o modifiche normative). Questo “quasi-monopolio” persiste ancora oggi nonostante l'entrata in vigore del decreto legislativo 164/00, il quale mirava alla liberalizzazione del mercato del gas in Italia (argomento trattato in dettaglio nella sezione 1.1.4). La ragione di tale persistenza risiede nell'elevato investimento economico richiesto per la costruzione di una rete di gasdotti. Per questo motivo, spesso, le aziende concorrenti sono scoraggiate dall'ingresso in questo settore, lasciando un notevole margine alle aziende già presenti.

La struttura di distribuzione di Snam si articola in due segmenti distinti: una rete nazionale ed una rete regionale. Una volta che il gas raggiunge il territorio di destinazione, viene incanalato dai gasdotti principali nella rete regionale, con particolare attenzione alle aree destinate allo stoccaggio. Il “processo di stoccaggio” è l'atto di preservare il gas all'interno di formazioni geologiche adeguate, al fine di accoglierlo, mantenerlo e successivamente rilasciarlo. La sua principale funzione consiste nell'archiviare il gas che supera il consumo attuale, consentendo di ritardarne l'uso durante periodi di elevata domanda o di bilanciare le fluttuazioni quotidiane o stagionali (ad esempio durante i picchi di domanda invernale) nella richiesta di gas. Per questo motivo, il sistema di stoccaggio rappresenta una risorsa di importanza cruciale per garantire una fornitura affidabile durante periodi di elevata domanda, oltre ad essere essenziale per gestire situazioni di emergenza all'interno dell'infrastruttura di approvvigionamento del gas naturale.

L'attività di stoccaggio viene principalmente eseguita attraverso l'impiego di strutture geologiche precedentemente utilizzate per l'estrazione di idrocarburi, rappresentando l'unico metodo di stoccaggio attualmente in uso in Italia. In altre nazioni, in cui le condizioni geologiche lo permettono, questo processo avviene in giacimenti salini o in formazioni acquifere, sia lungo le zone costiere che in aree marine remote. Per quanto concerne, invece, l'operazione di immagazzinamento del gas, essa si suddivide fondamentalmente in due fasi distinte, come mostrato in **Figura 1.3**, correlate alla variazione stagionale dei consumi:

- la “*fase di iniezione*”, solitamente dal mese di Aprile fino al mese di Ottobre in Italia, consiste nel trasferire il gas naturale dalla rete di trasporto e immagazzinarlo nei giacimenti appositamente predisposti per tale scopo;
- la “*fase di erogazione*”, generalmente dal mese di Novembre al mese di Marzo, riguarda il prelievo del gas naturale dai giacimenti per rilasciarlo nel sistema e soddisfare la domanda.

Nei mercati ben strutturati, dove sono state stabilite normative chiare e dove esistono infrastrutture robuste, il sistema di stoccaggio del gas naturale assume un ruolo cruciale nella gestione efficiente e trasparente del settore. Questo sistema non solo contribuisce a bilanciare la domanda e l'offerta, poiché il gas naturale è spesso richiesto in quantità

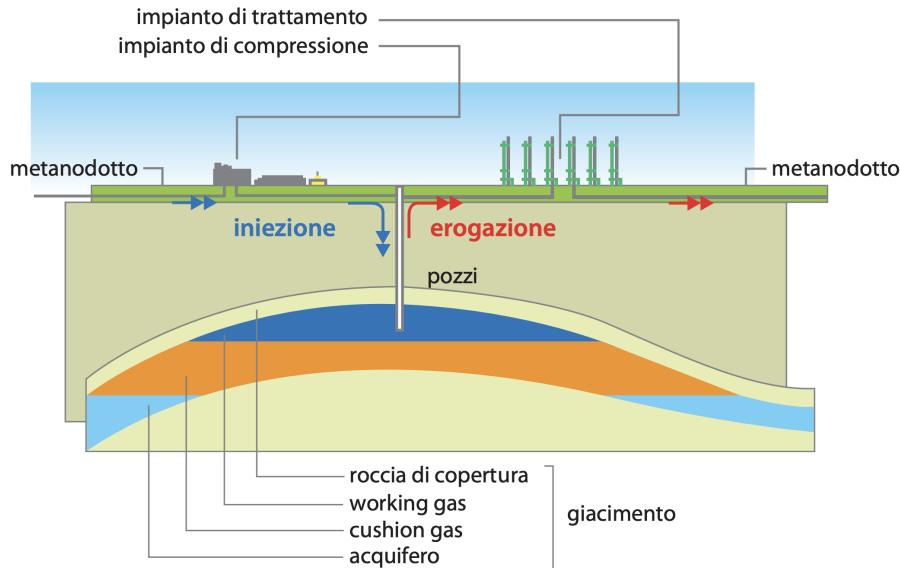


Figura 1.3. Schema di un giacimento di stoccaggio.

Fonte: Assolombarda Gruppo Energia

variabili a seconda delle stagioni e delle esigenze dei consumatori, ma svolge anche una funzione strategica nell'ottimizzazione delle opportunità di breve e medio termine legate alle dinamiche dei prezzi. La necessità di stoccaggio deriva dalla natura variabile della richiesta di gas naturale, influenzata da molteplici fattori quali condizioni meteorologiche, eventi geopolitici e fluttuazioni nel mercato.

I sistemi di stoccaggio consentono, per cui, di garantire una fornitura stabile e costante di gas naturale, riducendo le fluttuazioni dei prezzi e garantendo disponibilità quando richiesta. Allo stesso modo, offrono anche la possibilità di capitalizzare su opportunità di profitto attraverso differenziali di prezzo tra stagioni diverse. Ad esempio, durante periodi di bassa domanda, come accade nella stagione estiva, il gas può essere immagazzinato a costi inferiori per poi essere successivamente venduto a prezzi più alti durante l'inverno, quando la domanda è più elevata. Per tale motivo, il sistema di stoccaggio del gas naturale risulta essere un elemento fondamentale, rivestendo, anche, un'importanza strategica: non solo garantisce un approvvigionamento stabile ma consente anche di massimizzare i profitti sfruttando le fluttuazioni stagionali dei prezzi e le mutevoli esigenze di mercato.

1.1.3 Vendita

Le imprese di distribuzione svolgono un ruolo fondamentale nell'indirizzare il gas naturale verso gli utenti finali, mentre i consumatori interessati all'utilizzo di questo combustibile devono fare affidamento sulle aziende di fornitura per ottenerlo.

Nel contesto italiano, la completa liberalizzazione della vendita di gas naturale è stata implementata a partire dal 1° gennaio 2003, in ottemperanza alle disposizioni del Decreto Legislativo n. 164/00 (discusso nella sezione 1.1.4). Tale disposizione ha contri-

buito all'apparizione di numerose imprese di fornitura, conosciute come "fornitori", che propongono diverse tariffe ai loro clienti. Tuttavia, è fondamentale comprendere che tali tariffe rientrano sempre in un intervallo di valori definito, il quale è influenzato sia dal mercato internazionale che dalle direttive emanate dall'Autorità di Regolazione per l'Energia, Reti e Ambiente (ARERA). A loro volta, i fornitori di gas naturale acquistano questo combustibile dai cosiddetti "shipper", cioè aziende specializzate nella vendita all'ingrosso, a prezzi che variano in base a periodi e costi di mercato. È cruciale notare che i fornitori di gas garantiscono l'acquisto solo per i clienti domestici, gli uffici o i piccoli esercizi commerciali, mentre le grandi industrie e le centrali termoelettriche conducono direttamente le trattative di approvvigionamento con i venditori all'ingrosso.

1.1.4 Decreto Legislativo n. 164/00

Il decreto legislativo numero 164 del 23 maggio 2000 [8], comunemente noto come decreto Letta, deve il suo nome all'allora Ministro dell'Industria, del Commercio e dell'Artigianato, Enrico Letta, che ricopriva questa carica durante i governi presieduti da Massimo D'Alema e Giuliano Amato. Questo decreto rappresenta un significativo avanzamento nel processo di liberalizzazione del mercato del gas in Italia, contribuendo in maniera sostanziale alla trasformazione di questo settore.

Il decreto Letta prosegue l'opera di liberalizzazione che era stata avviata alla fine degli anni '90, in particolare attraverso il Decreto Bersani, il quale aveva permesso l'apertura del mercato energetico. A differenza del Decreto Bersani, Enrico Letta ha concentrato la sua attenzione principalmente sulla liberalizzazione del mercato del gas, abolendo il monopolio esistente e riorganizzando le diverse fasi di cui è costituito. Il processo di liberalizzazione in Italia è stato avviato in primo luogo con le imprese industriali e successivamente si è esteso alle utenze domestiche a partire da gennaio 2003.

Il decreto è stato emanato in risposta alla direttiva comunitaria n. 98/30/CE datata 28 giugno 1998, la quale mirava a riformare la regolamentazione del mercato del gas naturale nell'Unione Europea. Tale direttiva aveva l'obiettivo di stabilire un quadro comune per la deregolamentazione e, quindi, la liberalizzazione dei mercati del gas naturale nei paesi membri dell'Unione Europea. In particolare, essa conteneva disposizioni che richiedevano agli Stati membri di aprire i loro mercati del gas alla concorrenza, di separare le attività di produzione, trasmissione e distribuzione, e di garantire l'accesso non discriminatorio alle reti di trasporto del gas. L'entrata in vigore di questa direttiva è stata, in questo modo, un passo importante verso lo sviluppo di un mercato interno dell'energia in Europa, consentendo la creazione di un ambiente più competitivo e più favorevole alla diversificazione delle fonti di approvvigionamento energetico, garantendo, anche, prezzi più vantaggiosi per i consumatori.

In modo simile al Decreto Bersani, il decreto Letta ha riformato le diverse fasi del mercato, agevolando l'ingresso di nuovi partecipanti e rendendo più flessibili i processi operativi che, in passato, erano soggetti a rigide normative e schemi predefiniti. Tra le principali innovazioni proposte dal Decreto Letta, emerge la suddivisione delle funzioni all'interno

del mercato del gas, che ha portato a una chiara distinzione di ruoli e responsabilità tra le seguenti figure di primaria importanza:

1. **Trasportatore**: incaricato della gestione delle strutture di raccolta del gas e del suo inserimento nel sistema di trasporto;
2. **Grossista**: deterrente la proprietà dei gasdotti gestiti dal trasportatore, ruolo fondamentale nella distribuzione del gas sul mercato;
3. **Distributore**: responsabile della proprietà o della gestione delle reti locali di distribuzione del gas, assicurando un ruolo cruciale nella fornitura a livello locale;
4. **Società di vendita**: fornitore che si occupa direttamente della vendita di gas ai clienti finali, garantendo il collegamento tra il mercato e gli utenti domestici e commerciali.

La chiara separazione dei ruoli, delineata in precedenza, ha apportato un significativo contributo alla formazione di un apparato organizzativo caratterizzato da una maggiore efficienza e trasparenza. Questa separazione, in particolare, ha stimolato la competizione e la diversificazione all'interno del settore energetico. Un esempio evidente di questo principio è rappresentato dalla gestione del gas, in cui il governo nazionale assegna licenze di stoccaggio con una durata ventennale mentre l'immagazzinamento è regolamentato dall'azienda *Stogit*, la principale entità operante nel territorio nazionale nell'ambito del deposito di gas naturale e facente parte dell'ampio conglomerato industriale Snam. Questa distinzione operativa è stata istituita con l'intento di favorire l'apertura al mercato e la promozione di una sana concorrenza, garantendo, al contempo, una chiara regolamentazione e supervisione delle attività cruciali nell'ambito del settore energetico, come lo stoccaggio del gas. La concessione di licenze a lungo termine da parte dello Stato e il controllo affidato ad operatori qualificati, quali *Stogit* e il gruppo Snam, sono elementi fondamentali di questa strategia di separazione.

Distribuzione e vendita

Un ulteriore elemento cardine, introdotto dal decreto Letta, consiste nella separazione delle operazioni di distribuzione e di vendita. In precedenza, queste due attività venivano svolte all'interno di una medesima entità ma, grazie all'emanazione del decreto, si è proceduto ad una chiara distinzione dei rispettivi ruoli. In altre parole, il compito del distributore è ora focalizzato sulla fornitura di servizi a livello locale, cosicché si è trasferita la responsabilità della vendita ai fornitori energetici. La vendita vera e propria è, infatti, consentita solo agli operatori che dispongono di accesso ai sistemi di stoccaggio, ciò ha comportato, pertanto, una revisione delle tariffe che sono state suddivise tra i costi sostenuti dai fornitori e i servizi erogati dai distributori. È necessario sottolineare, tuttavia, che queste tariffe sono soggette a regolamentazione da parte di ARERA, allo scopo di garantire la corretta concorrenza tra gli operatori attivi nel settore (argomento approfondito nella sezione 1.1.5).

Gli attori del mercato

Precedentemente all’emanazione del decreto Letta, come già esposto, il settore dell’approvvigionamento del gas in Italia era in gran parte sotto il controllo della società Snam che deteneva una posizione di quasi totale monopolio nell’ambito della produzione, importazione e stoccaggio del gas. Tuttavia, con l’introduzione delle politiche di liberalizzazione in vigore, si è assistito ad un cambiamento significativo di questo panorama, consentendo ad una gamma più ampia di operatori di entrare nel mercato e prendere parte alla produzione, importazione e commercializzazione del gas. Questa apertura ha portato con sé un aumento della concorrenza ed una maggiore diversificazione delle opzioni a disposizione dei consumatori finali.

È importante notare che, nonostante tali trasformazioni, il controllo della rete di distribuzione del gas è in gran parte rimasto in mano alla società Snam, nonostante alcune modifiche nell’assetto societario del settore. In sintesi, il decreto Letta ha contribuito ad aprire il mercato del gas in Italia a nuovi partecipanti, incrementando la competizione ed ampliando le possibilità di scelta per i consumatori finali. Questo cambiamento ha semplificato, inoltre, il processo di cambio di fornitore senza richiedere la sostituzione del contatore o causare interruzioni nell’approvvigionamento e, per garantire un equo e competitivo mercato, tali innovazioni sono soggette a una regolamentazione apposita.

1.1.5 Situazione dei consumi nel 2022 (ARERA)

L’Autorità di Regolazione per Energia Reti e Ambiente (ARERA) rappresenta un ente amministrativo indipendente in Italia, istituito con l’obiettivo di agevolare la competitività nei settori dell’energia elettrica, del gas naturale, dell’acqua potabile, del teleriscaldamento/teleraffrescamento e della gestione dei rifiuti. Per realizzare tale scopo, ARERA si avvale principalmente di regolamentazioni tariffarie, facilitando l’accesso alle reti, garantendo il rispetto degli standard qualitativi nei servizi, supervisionando l’andamento dei mercati e proteggendo i diritti e gli interessi dei clienti e degli utenti finali.

Il ruolo centrale di ARERA è di fondamentale importanza nel contesto delle infrastrutture e dei servizi pubblici, poiché contribuisce a garantire un ambiente di mercato equo e competitivo, promuovendo l’efficienza e l’innovazione nei settori chiave dell’energia, dell’acqua e dello smaltimento dei rifiuti. La sua azione si traduce in benefici sia per le imprese che operano in questi settori grazie sia ad una concorrenza più aperta e alla possibilità di competere in modo equo, sia per i consumatori finali, che possono godere di servizi migliori e tariffe più convenienti. Inoltre, ARERA svolge un ruolo cruciale nella supervisione e nel monitoraggio costante delle dinamiche di questi mercati, contribuendo a mantenere la trasparenza, l’integrità e la stabilità dei servizi essenziali offerti alla popolazione.

Concentrazione globale dei consumi

Nell’analisi della situazione globale dei consumi di gas, riportata all’interno della “Relazione annuale ARERA 2023 sullo Stato dei servizi e sull’Attività svolta nel 2022” [9], si è riscontrata una contrazione del 1,5% circa dei consumi mondiali rispetto all’anno 2021. L’Europa ha sperimentato la riduzione più significativa, con un decremento percentuale

del 14,0%. Allo stesso tempo, in Asia Pacifico (l'insieme delle nazioni asiatiche ed oceانية le cui coste sono bagnate dall'Oceano Pacifico) e in Cina, la diminuzione è stata rispettivamente del 1,6% e dello 0,8%. Da notare che quest'ultimo risulta essere il primo calo della domanda dopo due decenni. In contrasto, gli Stati Uniti hanno registrato un notevole aumento della domanda, pari al 5,4%, principalmente dovuto all'uso crescente del gas nel settore termoelettrico a causa dell'aumento del prezzo del carbone rispetto al gas statunitense. Tra i principali mercati dell'Unione Europea, tra cui Germania, Italia, Francia, Olanda e Spagna, invece, le riduzioni dei consumi variano dal 3,8% della Spagna al 22% dei Paesi Bassi, con l'Italia a -9,9% e la Germania a -15,3%. Infine, si registra una diminuzione della domanda anche nel Regno Unito pari al 7%.

Nel contesto della produzione globale di gas, si è osservato un mantenimento sostanziale della quantità prodotta ma si è assistito ad un considerevole aumento nella produzione di gas non convenzionale. Questa tipologia di produzione è relativa all'estrazione di gas naturale da risorse geologiche particolari. Le principali risorse coinvolte comprendono il gas di scisto (*shale gas*), ovvero il gas naturale intrappolato all'interno delle rocce di scisto, la cui estrazione richiede l'impiego di una tecnica chiamata "fratturazione idraulica" o "fracking" che implica l'iniezione di acqua, sabbia e sostanze chimiche sotto pressione nelle rocce per creare fratture e consentire al gas di fuoriuscire; il gas di carbone (*coalbed methane*), associato ai depositi di carbone sotterranei, in questo caso il gas è intrappolato nelle fessure di tali giacimenti e spesso richiede la rimozione dell'acqua presente nella riserva mediante pompe e sistemi di drenaggio per estrazione; infine, vi è il gas proveniente da formazioni sabbiose (*tight gas*), racchiuso in rocce sabbiose o argillose con bassa permeabilità, la cui estrazione richiede la perforazione orizzontale e la fratturazione idraulica della roccia stessa. Questa categoria di produzione è aumentata dal 25% del totale nel 2021 al 31% nel periodo corrente.

All'interno del contesto europeo, malgrado un aumento complessivo della produzione pari al 3,6%, prevalentemente grazie ai contributi di Norvegia e Regno Unito, si è riscontrata una contrazione del 7% nella produzione complessiva. Questa diminuzione è sostanzialmente imputabile alla riduzione pianificata della produzione di gas nel giacimento di Groningen nei Paesi Bassi, che costituisce il giacimento più esteso d'Europa ed è stato operativo fin dal lontano 1963. Nel corso degli ultimi anni, questo giacimento ha rilasciato una quantità annuale di gas pari a circa 40 miliardi di metri cubi, contribuendo al 10% del consumo europeo. Tuttavia, benché restino circa 450 miliardi di metri cubi di gas nel giacimento, dopo quasi sessant'anni di estrazione, il governo dei Paesi Bassi ha annunciato la graduale riduzione dell'estrazione, con l'obiettivo di terminarla completamente entro la fine del 2024. La motivazione principale di questa decisione è stata l'aumento della frequenza di terremoti associati alle attività di estrazione nel vasto giacimento, con l'ultimo sisma registrato poche ore prima della cessazione delle operazioni nell'impianto. Durante gli ultimi quarant'anni, le autorità hanno documentato oltre mille terremoti nell'area di interesse. Anche se tali scosse non hanno mai superato i 3,6 gradi sulla scala Richter, la loro costante frequenza rappresenta una preoccupante anomalia, specialmente in una regione che non era precedentemente nota per l'attività sismica, almeno fino agli anni '60, quando l'estrazione di gas di Groningen ha cominciato a rifornire le abitazioni olandesi e quelle dei paesi confinanti, come il Belgio. Tuttavia, gli effetti

negativi non sono limitati alla sfera materiale. Con riferimento a recenti ricerche [10], si è riscontrato che questi terremoti hanno causato disturbi fisici e psicologici nella popolazione locale, tra cui mal di testa, insonnia, palpitzazioni e problemi gastrointestinali, oltre a stati di ansia e depressione. Questi disturbi, a loro volta, hanno contribuito ad aumentare il rischio di decessi prematuri, ad esempio a causa di malattie cardiovascolari o di suicidio, con una stima di 16 decessi prematuri all'anno.

Questa dinamica dei consumi e della produzione di gas a livello globale e regionale è influenzata da una serie di fattori, tra cui le condizioni economiche, la domanda energetica, la disponibilità di risorse e le politiche energetiche in corso. Le differenze nei tassi di declino dei consumi e le variazioni nella produzione riflettono le sfide e le opportunità che l'industria del gas sta affrontando a livello mondiale.

Scoppio della guerra in Ucraina

A causa dei recenti eventi legati al conflitto in Ucraina, si è verificata una ristrutturazione del sistema di approvvigionamento energetico europeo. Questa ristrutturazione ha comportato un incremento dell'utilizzo di GNL disponibile sui mercati internazionali, nonché la costruzione di nuovi terminali di rigassificazione. In aggiunta, si è lavorato per incrementare le importazioni di gas tramite gasdotti alternativi a quelli provenienti dalla Russia. Nel corso del 2021, l'Unione Europea aveva importato una quantità di gas pari a circa 375 miliardi di metri cubi, al netto delle re-esportazioni. Di questa quantità, l'80% proveniva da gasdotti e il restante 20% da GNL. Nel 2022, il totale delle importazioni è diminuito a circa 360 miliardi di metri cubi, rappresentando una diminuzione del 3,6%. Di questa quantità, il 64% proveniva da gasdotti, mentre il 36% proveniva da GNL. Quanto esposto evidenzia la drastica riduzione della quantità di gas importata tramite gasdotti (circa il 21%, equivalente a 63 miliardi di metri cubi), principalmente a causa della decisione dell'UE di ridurre le importazioni di gas dalla Russia. Infatti, nel 2021 la Russia costituiva circa il 50% delle importazioni europee via gasdotto, mentre nel 2022 la sua quota è scesa ad una quota pari al 28%. Per compensare tale riduzione, tuttavia, l'Unione Europea ha incrementato le importazioni di GNL nel 2022, raggiungendo una crescita del 63% rispetto al 2021.

Tra i paesi di provenienza del GNL importato nell'Unione Europea, il 46% proviene dalle Americhe, principalmente dagli Stati Uniti, il 21% dall'Africa, il 15% dal Medio Oriente e un ulteriore 15% dalla Russia. Questo ha comportato un incremento complessivo del 35% rispetto all'anno precedente. Inoltre, è stato osservato un aumento delle forniture di gas tramite gasdotti provenienti dalla Norvegia e dall'Azerbaijan, mentre l'Algeria ha spostato i flussi di GNL dalla Spagna all'Italia.

Nel contesto dei prezzi energetici, i costi del gas naturale hanno subito un significativo aumento a causa degli avvenimenti internazionali. In particolare, con lo scoppio del conflitto in Ucraina, i prezzi di mercato spot (ovvero il costo corrente del gas naturale sul mercato in un dato momento) hanno raggiunto livelli senza precedenti, con punte vicine ai 200 €/MWh, rappresentando un notevole incremento rispetto al passato. Questi prezzi elevati sono stati influenzati dalla diminuzione delle forniture di gas russo e dalla

necessità di rifornire rapidamente le riserve, insieme ad altri fattori contingenti. Anche durante l'estate successiva, i prezzi sono rimasti elevati, mantenendo un mercato squilibrio tra domanda e offerta. Alla fine del 2022, i prezzi del mercato all'ingrosso italiano hanno raggiunto i 140 €/MWh, evidenziando un aumento del 55% rispetto alla media di novembre.

Situazione in Italia nel 2022

Nel corso dell'anno 2022, l'Italia ha sperimentato una notevole contrazione dei consumi di gas naturale, registrando una significativa riduzione del 10% rispetto all'anno precedente. Tale contrazione è stata osservata in diversi comparti chiave dell'economia, tra cui l'industria, la generazione di energia termoelettrica, il settore del commercio e dei servizi, i trasporti e le residenze private. Questo andamento è stato ulteriormente accentuato da una riduzione del 2,7% nella produzione nazionale di gas, che ha determinato una crescente interdipendenza dell'Italia dalle forniture estere, rappresentanti ora il 99% del gas importato.

Un elemento di particolare rilevanza è stato il calo significativo delle importazioni dalla Russia, mentre l'Algeria è emersa come il principale fornitore di gas per l'Italia, seguita dalla Russia stessa e dall'Azerbaigian. È interessante notare come le importazioni via mare abbiano guadagnato un ruolo crescente, con flussi provenienti da paesi come Spagna, Egitto e Nigeria.

In questo contesto, l'azienda Ente Nazionale Idrocarburi (Eni), nonostante mantenga la sua posizione di principale importatore di gas, ha visto una contrazione della sua quota di mercato. Tuttavia, è da notare che il numero di esportatori di gas dall'Italia è cresciuto in modo significativo, contribuendo in maniera tangibile all'aumento dell'indipendenza energetica del paese dalle importazioni russe.

Un'altra dinamica degna di attenzione, è l'incremento sostanziale delle riserve di gas immagazzinate. Questo aumento si è verificato principalmente come risposta alla ridotta produzione idroelettrica, che è stata influenzata da condizioni di siccità e problemi nelle centrali elettriche del paese. Tale accumulo di risorse energetiche riveste un'importanza cruciale nella garanzia per la continuità dell'approvvigionamento energetico nazionale.

1.2 Environmental, Social and Governance (ESG)

La Responsabilità Sociale delle Imprese (RSI) ha acquisito una rilevanza sempre maggiore nel contesto aziendale contemporaneo, esercitando un'influenza significativa sulla percezione pubblica delle aziende [11]. Studi recenti [12] hanno evidenziato che comportamenti eticamente e socialmente responsabili da parte delle imprese sono strettamente collegati ad una migliore reputazione aziendale, alla soddisfazione dei clienti ed alla fiducia che questi ultimi ripongono nell'azienda. La necessità di assumersi la responsabilità sociale delle azioni aziendali è sempre più riconosciuta anche dai decisori politici. In particolare, l'Unione Europea adotta regolamenti che obbligano le aziende a migliorare il loro impatto ambientale e sociale.

Motivate dall'incremento di interesse riguardo la RSI, le organizzazioni si sono impegnate ad incorporare i criteri sociali e ambientali nelle proprie pratiche aziendali ed a diffondere relazioni aziendali dedicate alla sostenibilità. Secondo quanto riportato dal Governance & Accountability Institute, un'organizzazione di consulenza e ricerca con sede negli Stati Uniti focalizzata su questioni legate alla sostenibilità, nel 2018 l'86% delle aziende elencate nell'indice S&P 500 ha pubblicato relazioni sulla sostenibilità aziendale o sulla responsabilità d'impresa. L'indice S&P 500 rappresenta le *performance* delle 500 principali società quotate in borsa negli Stati Uniti, selezionate da Standard & Poor's, un'agenzia di *rating* finanziario e di analisi. Questa selezione si basa su criteri diversi, tra cui la capitalizzazione di mercato, la liquidità e la rappresentanza di vari settori dell'economia. Questo aumento significativo rispetto al 20% del 2011 [13] evidenzia un crescente interesse ed impegno delle aziende nei confronti della sostenibilità e della responsabilità sociale.

Al fine di valutare l'impatto sociale ed ambientale di un'organizzazione, è stato formulato il concetto di “Environmental, Social and Governance” (ESG). A differenza dell'approccio che valuta un'impresa esclusivamente da un punto di vista commerciale e finanziario, l'analisi ESG considera l'impatto dell'azienda sull'ambiente e sulla società, valutando, inoltre, come tale impatto sia gestito. Pertanto, questo tipo di analisi contempla il modo in cui un'azienda influisce sull'ambiente e sulla società, basandosi su tre criteri fondamentali che le aziende devono considerare per garantire la propria sostenibilità a lungo termine. Nel dettaglio:

- Il criterio “**E**” si riferisce all’ambiente e si concentra sulla gestione degli impatti ambientali, sia diretti che indiretti, sul rispetto dell’ambiente e della biodiversità, sulla contribuzione alla riduzione delle emissioni di gas serra e sul supporto all’economia decarbonizzata (un’economia che mira a bilanciare l’attività economica con la necessità di preservare l’ambiente e mitigare il riscaldamento globale, riducendo l’emissione di gas serra, in particolare il biossido di carbonio, ovvero CO_2);
- Il criterio “**S**”, invece, riguarda gli aspetti sociali e si occupa della gestione delle risorse umane, dei rapporti con i fornitori e con le comunità, promuovendo politiche di lavoro dignitoso, uguaglianza di opportunità e la salvaguardia dei diritti umani;
- Infine, il criterio “**G**” verte sulla *governance* e si concentra sull’adozione di pratiche di gestione ottimali, sull’aderenza a codici etici, sulla trasparenza e sull’impegno nella lotta alla corruzione da parte dei dirigenti aziendali e del consiglio di amministrazione.

All'interno della **Figura 1.4**, vengono mostrati i *driver* più importanti per ogni fattore ESG. In particolare [14]:

- **Fattore Ambiente**

- *Cambiamento climatico*. Si procede alla valutazione delle emissioni di CO_2 prodotte dalla società, sia in valore assoluto che in relazione ai concorrenti.



Figura 1.4. Schema riassuntivo dei fondamenti alla base del concetto di ESG.
Fonte: Innolva

- *Risorse Naturali.* Viene effettuata un'analisi della quantità di *input* utilizzati per la produzione di beni e servizi specifici, inclusi i consumi d'acqua e gli sprechi associati.
- *Inquinamento e sprechi.* Si esaminano gli sprechi pericolosi e l'impatto ambientale generato dalla società, includendo l'inquinamento dell'aria, dell'acqua e del suolo, nonché l'adozione di pratiche di imballaggio sostenibile.
- *Utilizzo ambientale.* Si valutano gli effetti degli stabilimenti presenti e futuri sulla fauna e la flora locali, compresa la questione della deforestazione.

• Fattore Sociale

- *Gestione delle risorse umane.* Si effettua un'analisi della diversità di genere, religione ed etnia all'interno dell'azienda, insieme ai programmi di attrazione dei talenti e agli standard lavorativi adottati.
- *Prassi di sicurezza.* Vengono valutate le misure aziendali per garantire la sicurezza dei lavoratori, insieme al numero di incidenti sul lavoro ed ai resi dei prodotti da parte dei clienti a causa di preoccupazioni sulla sicurezza.
- *Gestione dei clienti.* Si esaminano le statistiche di soddisfazione della clientela e gli standard di protezione dei dati dei clienti.
- *Impatti sulla comunità.* Si valutano le iniziative di coinvolgimento delle comunità in cui le aziende operano, insieme alle esposizioni al rischio di guerra e terrorismo.

• Fattore Governance

- *Struttura gerarchica.* Si analizzano la composizione e la diversità del Consiglio di Amministrazione (CdA), il numero di presenze dei membri, la remunerazione e gli incentivi dei dirigenti e l'organizzazione aziendale.

- *Codici di condotta e valori.* Si valuta l’effettiva applicazione dei codici comportamentali, la coerenza tra le dichiarazioni pubbliche sui valori etici dell’azienda e le pratiche interne, nonché la presenza di funzioni aziendali dedicate al monitoraggio (*compliance*) del rispetto delle regole.
- *Trasparenza e reporting.* Si verifica il grado di divulgazione degli indicatori ESG, le pubblicazioni degli annuali bilanci di sostenibilità e la trasparenza in merito ai livelli di tassazione aziendale.
- *Rischio informatico.* Si valuta la presenza di procedure di sicurezza informatica, insieme ai piani di prevenzione ed ai piani di contingenza in caso di attacchi informatici (ovvero insiemi di procedure e strategie pianificate in anticipo per affrontare e rispondere efficacemente a situazioni di emergenza causate da violazioni della sicurezza informatica).

Generalmente, questo comporta un approccio rigoroso al fine di valutare le *performance* aziendali, stabilire obiettivi e priorità, valutare i potenziali rischi ed individuare opportunità, elaborare piani per ridurre tali rischi e definire strategie di transizione, gestire l’allocazione delle risorse e svolgere altre attività volte a garantire una crescita stabile. Di conseguenza, l’aggregazione di dati indicatori all’interno di ciascuno dei criteri legati all’ambiente, alla responsabilità sociale ed alla *governance* (ESG) permette di creare un *report* ESG che raccoglie tutte le informazioni non finanziarie relative all’azienda.

1.2.1 Contesto storico legato a ESG

Un punto di partenza significativo per ciò che può essere definito come un “percorso di consapevolezza” che ha portato alla concettualizzazione dell’approccio ESG può essere ricondotto agli anni ‘70. In quel periodo, The Club Of Rome, un’associazione non governativa senza scopo di lucro composta da scienziati, economisti, uomini e donne d’affari, attivisti dei diritti civili, alti dirigenti pubblici internazionali e capi di Stato provenienti da tutto il mondo, lavorò in collaborazione con il MIT (Massachusetts Institute of Technology) portando alla creazione del rapporto ben noto intitolato “The Limits to Growth” [15]. La rilevanza di questo rapporto emerge con maggior chiarezza oggi, poiché sottolinea in modo inequivocabile che la prospettiva di una crescita infinita in un contesto di risorse naturali non rinnovabili risulta irrazionale e impraticabile. Nel medesimo contesto storico, precisamente nel 1972, si è tenuta la prima conferenza sul tema ambientale delle Nazioni Unite, da cui è scaturita la Dichiarazione di Stoccolma [16]. In tale dichiarazione, è stato sancito il “diritto universale di tutti gli esseri umani ad avere accesso a condizioni di vita soddisfacenti in un ambiente che permetta una vita dignitosa e prospera”.

Questo percorso ha condotto a sviluppi significativi relativi ai concetti di sviluppo sostenibile, che sono emersi in modo ancora più evidente nel 1987. In quell’anno, la Commissione Mondiale per l’Ambiente e lo Sviluppo delle Nazioni Unite ha pubblicato il rapporto intitolato “Our Common Future” [17]. Questo documento si è proposto di conciliare lo sviluppo economico con la sostenibilità, introducendo il concetto di “sviluppo sostenibile”, che si avvicina ai principi fondamentali alla base dell’ESG. Va notato che, in quegli anni, l’attenzione era prevalentemente concentrata sul fattore “E”, ovvero sull’impatto ambientale dello sviluppo economico. Tuttavia, come precedentemente sottolineato, i

principi dell’ESG abbracciano anche questioni relative all’impatto sociale, all’inclusività, alla responsabilità verso le persone e le comunità, nonché alle nuove forme di *governance* responsabile. Nel corso del tempo, questi principi hanno contribuito a definire l’approccio globale dell’ESG.

Nonostante la finanza sostenibile costituisca una pietra miliare nella storia dell’investimento responsabile, i principi dell’Environmental, Social and Governance (ESG), applicati alla finanza moderna, possono essere tracciati con precisione al 2004. In quell’anno, Kofi Annan, all’epoca Segretario Generale delle Nazioni Unite, invitò una selezione di circa 50 CEO delle principali istituzioni finanziarie globali ad unirsi in un sforzo congiunto per introdurre i valori del *framework* ESG nei mercati dei capitali. L’acronimo ESG venne coniato, poi, nel 2005 durante la conferenza “Who Cares Wins” [18]. Questo straordinario evento ha riunito un’ampia gamma di professionisti, tra cui investitori istituzionali, analisti di ricerca *buy-side* e *sell-side*, gestori patrimoniali, consulenti globali ed enti governativi. Durante questa conferenza, si è esaminato in dettaglio come i fattori ESG potessero influenzare la ricerca finanziaria e la gestione del patrimonio. Successivamente, nel marzo del 2018, la Commissione Europea ha presentato un “Piano d’Azione per la finanza sostenibile” [19], delineando una strategia chiara per la creazione di un sistema finanziario che promuova lo sviluppo sostenibile sotto gli aspetti economici, sociali ed ambientali. Questo importante passo in avanti è stato un contributo cruciale per l’effettiva attuazione dell’Accordo di Parigi [20] sul cambiamento climatico e per l’Agenda 2030 [21] delle Nazioni Unite, che promuove il concetto di sviluppo sostenibile.

La finanza sostenibile rappresenta un obiettivo di primaria importanza, poiché mira ad instaurare un equilibrio virtuoso tra il perseguimento del profitto finanziario e l’impatto positivo sulla società e sull’ambiente. Questa prospettiva a lungo termine spinge i capitali verso attività economiche che non solo generano rendimenti positivi, ma contribuiscono in modo concreto al benessere della società ed alla salvaguardia dell’ambiente. In tal senso, la finanza sostenibile abbraccia un approccio più responsabile ed etico agli investimenti, rafforzando l’idea che il successo finanziario debba essere strettamente connesso al bene comune. Elemento fondamentale nella promozione di questa filosofia sono, appunto, i criteri ESG, che forniscono una base oggettiva e standardizzata per valutare le prestazioni delle aziende nei ambiti chiave. In passato, le iniziative sociali ed ambientali, così come le pratiche di buona *governance*, erano considerate decisioni soggettive a discrezione delle aziende, rendendo complesso il confronto tra realtà aziendali diverse. I criteri ESG, invece, introducono un insieme di parametri misurabili e comparabili che consentono di valutare in modo oggettivo l’impegno di un’azienda nell’ambito ambientale, sociale e di *governance*. Questi criteri giocano un ruolo fondamentale nell’orientare gli investimenti verso imprese che abbracciano i principi della finanza sostenibile, promuovendo una crescita economica più equa e responsabile.

1.2.2 Rischi ESG

I rischi ESG, comunemente noti come “rischi di sostenibilità” o “rischi non finanziari”, abbracciano una vasta gamma di tematiche, tra cui il cambiamento climatico, i diritti umani, le dinamiche aziendali, la tassazione etica e la *governance*. La gestione inadeguata

di tali rischi può comportare gravi conseguenze sia a livello economico che reputazionale. La crescente attenzione da parte degli *stakeholder* verso la responsabilità sociale delle imprese sottolinea l'importanza cruciale di affrontare efficacemente questi aspetti.

I rischi connessi all'ambiente nella sfera aziendale comprendono una serie di questioni di notevole rilevanza. Queste problematiche riguardano principalmente la gestione dei rifiuti, la misurazione dell'impronta di carbonio, le sfide correlate al cambiamento climatico e l'inquinamento ambientale. D'altra parte, i rischi di natura sociale si concentrano sulla gestione del capitale umano, con un'attenzione particolare alla parità nel trattamento e nelle opportunità offerte ai dipendenti, nonché alla promozione della diversità ed un'attenzione sui rischi professionali nel contesto aziendale. Inoltre, si affrontano le sfide legate alle relazioni con il personale. Infine, i rischi di *governance* includono la corruzione, la concussione, la remunerazione dei dirigenti, la gestione fiscale e la promozione dell'equità all'interno dei consigli di amministrazione. È importante sottolineare che tali rischi non sono distinti, ma piuttosto interconnessi tra loro, richiedendo, pertanto, una gestione oculata e sinergica. L'integrazione dei criteri ESG offre, però, anche un'ampia gamma di vantaggi significativi per le aziende. Tra questi vantaggi rientrano la creazione di un solido vantaggio competitivo, la mitigazione dei rischi, l'attrazione di talenti di alto livello (poiché i lavoratori prediligono le aziende impegnate nei confronti della società e che hanno uno scopo definito, pertanto l'integrazione dei criteri ESG funge da incentivo per lavorare per un'azienda) e l'accesso a nuove opportunità di investimento. Ad esempio, negli ultimi anni e a seguito della pandemia, gli investitori hanno dimostrato di esigere un maggiore impegno da parte delle aziende per frenare e mitigare i cambiamenti climatici, guidare la transizione ad un'economia decarbonizzata, contribuire a far fronte alle sfide sociali ed alla riduzione delle disuguaglianze. In sintesi, i criteri ESG costituiscono un pilastro fondamentale per assicurare la sostenibilità a lungo termine delle aziende, fornendo gli strumenti essenziali per affrontare e superare le sfide relative all'ambiente, alla sfera sociale ed alla *governance*.

1.2.3 Rating ESG

Il *rating* ESG rappresenta l'analisi delle *performance* di un'azienda in termini di sostenibilità ambientale, impatto sociale e procedure di *governance*. Questa valutazione offre un'indicazione della capacità aziendale nel gestire rischi ed opportunità in queste tre fondamentali dimensioni. La stima di tale valutazione si basa su diversi fattori, tra cui:

- La trasparenza delle informazioni aziendali;
- La gestione del rischio ambientale;
- La conformità alle normative ambientali;
- La tutela dei diritti umani, la promozione della diversità ed inclusione;
- La gestione delle controversie;
- E la struttura di *governance*.

Di conseguenza, la valutazione può riguardare il livello di allineamento e *compliance* di un’azienda rispetto alle strategie ed alle direttive internazionali in materia di sostenibilità stabilite da istituzioni quali l’Unione Europea, le Nazioni Unite e l’Organizzazione per la Cooperazione e lo Sviluppo Economico (OCSE). Quest’ultima svolge principalmente un ruolo di assemblea consultiva, offrendo un’occasione di confronto sulle esperienze politiche, la risoluzione di problematiche comuni, l’identificazione di pratiche commerciali e il coordinamento delle politiche locali ed internazionali dei Paesi membri. Questi Paesi, caratterizzati da un’economia di mercato, condividono sviluppi comuni nel contesto economico globale.

Le imprese possono conseguire un *rating* ESG avvalendosi della consulenza di un *rating advisor*. Tale figura professionale è un esperto nelle valutazioni ESG, in grado di offrire un’analisi approfondita delle *performance* aziendali in termini di sostenibilità, suggerendo azioni specifiche per migliorare la valutazione complessiva. Il *rating advisor*, inoltre, può supportare l’azienda nella preparazione delle informazioni necessarie per la valutazione stessa, offrendo indicazioni relative a quali dati fornire e come presentarli in maniera efficace. Tale figura può supportare le aziende portando diversi vantaggi. Obiettivo primo delle aziende è avere un elevato punteggio per la valutazione ESG in modo tale che possa accrescere la credibilità e l’immagine aziendale, evidenziando l’effettivo impegno verso la sostenibilità e la responsabilità sociale. In secondo luogo, un valutazione ESG molto positiva può migliorare l’accesso al credito bancario, poiché le istituzioni finanziarie prestano sempre maggiore attenzione alle questioni ESG concedendo finanziamenti a condizioni più favorevoli alle imprese con un *rating* elevato.

È importante evidenziare, tuttavia, che non esiste uno standard internazionale consiviso tra le varie agenzie di *rating* per il calcolo del *rating* ESG. Sebbene, quindi, non sia possibile delineare un metodo di calcolo univoco, è possibile distinguere due approcci per il reperimento dei dati [22]:

- Un **approccio quantitativo**, basato sulle informazioni rese pubbliche dall’azienda conformemente agli standard internazionali;
- Un **approccio qualitativo**, fondato su dati ottenuti attraverso questionari, anche esterni, successivamente valutati secondo diverse metodologie.

Optare per il primo approccio, ovvero misurare le *performance* ESG esclusivamente attraverso le informazioni autodichiarate dall’azienda, comporta il rischio di una valutazione incompleta a causa della parzialità dei dati considerati e della modalità di *reporting* scelta dalla società. In contrasto, il secondo approccio, che incorpora dati alternativi esterni generati dagli *stakeholder* dell’azienda, può offrire una visione autenticamente imparziale e veritiera dei parametri di sostenibilità.

Nonostante non esista una metodologia univoca, è possibile evidenziare delle prassi comuni a tutte le agenzie di *rating*. Tra i più utilizzati vi sono:

- MSCI;
- FTSE;

- Refinitiv;
- Ecovadis;
- Standard Ethics.

Il modello MSCI ESG rating [23] costituisce un approccio fondato sull’analisi dell’esposizione ai rischi ed alle opportunità legate agli aspetti ESG di un’azienda, con particolare riguardo agli elementi rilevanti per il settore di competenza. Questa valutazione comprende non solo la misurazione delle capacità di gestione dei rischi e delle opportunità in confronto ai *competitor*, ma anche la fornitura di strumenti e servizi di supporto decisionale per la comunità di investimento internazionale, posizionando MSCI come un *leader* in questo settore. Attraverso la scala di valutazione da “AAA” a “CCC”, il MSCI ESG Ratings mira a quantificare l’esposizione di un’azienda ai rischi e alle opportunità ESG significativi per il suo settore, valutando l’efficacia della gestione di tali elementi, sia in termini generali che in confronto ai *competitor*. La metodologia considera in modo approfondito sia i rischi che le opportunità ESG materiali, derivanti da tendenze su larga scala come il cambiamento climatico, la scarsità di risorse ed i cambiamenti demografici, nonché dalla natura specifica delle attività aziendali. In sintesi, la classificazione proposta da MSCI offre una visione complessiva della posizione di un’azienda nei confronti dei fattori ESG, contribuendo così a orientare le decisioni di investimento nella comunità internazionale.

Risulta essere alquanto diverso il modello FTSE [24]. In questo caso il *rating* ESG si calcola secondo 12 macro criteri relativi ad ambiente, impatto sociale e *governance*. Tra questi macro criteri ci sono biodiversità, cambiamento climatico, lotta alla corruzione, *risk management* (l’insieme di azioni intraprese dalle aziende nel tentativo di controllare il livello di rischio associato alle linee di *business* e, in generale, all’impresa nel suo complesso) e *customer responsibility* (impegno degli individui come singoli, o come parte di una società, nel contribuire alla gestione delle problematiche di impatto sociale ed etico all’interno della propria area di attività).

Il modello di EcoVadis [25], invece, lanciata nel 2007 a Parigi e che vanta ora 600 dipendenti e sedi in tutto il mondo, adotta una metodologia costruita su standard internazionali di sostenibilità che includono:

- Il *Global Reporting Initiative*, ente internazionale, senza scopo di lucro, nato per definire gli standard di rendicontazione della *performance* sostenibile di aziende e organizzazioni di qualunque dimensione, appartenenti a qualsiasi settore e paese del mondo;
- Il *Global Compact* delle Nazioni Unite, patto non vincolante nato per incoraggiare le imprese di tutto il mondo ad adottare politiche sostenibili ed osservanti la responsabilità sociale d’impresa e per rendere pubblici i risultati delle azioni intraprese;
- E l’ISO 26000, “Guida alla responsabilità sociale” è una guida per tutti i tipi di organizzazioni, indipendentemente dalle dimensioni e localizzazioni. Fornisce indicazioni su concetti, termini e definizioni relativi alla responsabilità sociale.

Il modello Refinitiv [26], invece, azienda del LSEG (*London Stock Exchange Group*), importante gestore di Borse finanziarie, misura le *performance* della società sulla base di dati di dominio pubblico. Essa cattura e calcola oltre 630 misurazioni di natura ESG sulla base di cui un sottoinsieme di 186 società tra le più comparabili e rilevanti per settore, determina la valutazione complessiva dell’azienda ed il processo di assegnazione del punteggio.

Infine, nel caso del *rating* proposto da *Standard Ethics* [27], il quale offre valutazioni sollecitate e non finanziarie a entità che desiderano sottoporsi ad esame, è opportuno sottolineare che tale organizzazione si astiene dal formulare raccomandazioni agli investitori ed impartire consulenze finanziarie. Il procedimento di valutazione proposto da *Standard Ethics* è caratterizzato da un processo guidato dagli analisti con una durata media compresa tra sei e otto settimane. Le valutazioni sono espresse su una scala di nove gradi, in cui il massimo punteggio assegnabile è “EEE”, mentre il minimo è indicato con la lettera “F”.

La valutazione del rischio legato ai fattori ESG assume un ruolo cruciale anche nella determinazione del valore economico di un’azienda. In particolare, si può quantificare l’entità del rischio economico che un’azienda potrebbe affrontare a causa di tali fattori, focalizzandosi sul livello di rischio ESG non gestito. Tale valutazione si traduce in punteggi inferiori al diminuire del rischio non gestito, indicando una migliore *performance* in ambito ESG, come evidenziato nel sistema di *rating* proposto da *Sustainalytics*. *Sustainalytics*, società affiliata a *Morningstar* (società *leader* nella fornitura di ricerche finanziarie indipendenti), si configura come una delle principali entità indipendenti specializzate nella ricerca, valutazione e analisi ESG e di *corporate governance*. Il suo ruolo fondamentale consiste nel fornire supporto agli investitori nello sviluppo e nell’attuazione di strategie di investimento responsabili. Ciò avviene attraverso l’integrazione di informazioni dettagliate e valutazioni sulla sostenibilità nei processi decisionali legati agli investimenti. Un rischio viene considerato significativo per un settore quando esiste la probabilità che le aziende in esso operanti debbano affrontare costi sostanziali associati a tale rischio. Ad esempio, potrebbe verificarsi un divieto normativo su un *input* chimico, richiedendo alle aziende la riformulazione di un prodotto. Inversamente, un’opportunità è rilevante per un settore quando esiste la probabilità che le aziende possano sfruttarla per ottenere benefici economici. Questo potrebbe manifestarsi, ad esempio, nel caso in cui le aziende del settore possano capitalizzare una tecnologia innovativa e sostenibile, come quella orientata alla *green economy*.

Nel complesso, la qualità di un *rating* ESG è influenzata da diversi fattori, tra cui la chiarezza e la solidità della metodologia utilizzata, l’attenzione alle questioni significative, la credibilità delle fonti di dati impiegate, l’esperienza e la competenza del *team* di ricerca, l’effettiva partecipazione dell’azienda sottoposta a valutazione e degli *stakeholder* nel processo di valutazione ed infine l’adozione del *rating* nell’ambito di riferimento.

Alla luce di quanto spiegato, è facile intuire perché sempre più investitori considerano le aziende con un elevato *rating* ESG meno rischiose. L’impegno verso la sostenibilità le rende capaci di gestire al meglio i rischi ambientali, sociali e di *governance*.

1.2.4 AI legata al calcolo della ESG

A causa della crescente proliferazione dei dati a disposizione, si pone una sfida sempre più intricata per gli investitori, le imprese e le istituzioni governative nella formulazione di decisioni basate su una valutazione completa delle tematiche connesse all'ambito ESG. Questi dati, derivanti da una vasta gamma di fonti quali notizie, social media, sensori remoti, *Internet of Things* (IoT) e molto altro, richiedono un impegno significativo, sia in termini di risorse finanziarie che di tempo, per essere adeguatamente elaborati e compresi. In questo contesto, l'applicazione di algoritmi di intelligenza artificiale, tra cui il Machine Learning e l'analisi dei dati, si pone come una soluzione di grande rilevanza. Tali algoritmi consentono di accelerare il processo decisionale e migliorare la comprensione delle informazioni estratte da questa vasta e complessa mole di dati.

Il campo dell'Intelligenza Artificiale (IA) rappresenta una disciplina in costante evoluzione, abbracciando un insieme di approcci numerici volti alla risoluzione di complessi problemi di previsione, ottimizzazione, classificazione e *clustering*. Pur essendo un ambito di ricerca relativamente giovane, l'IA ha suscitato l'interesse a livello mondiale con considerevoli investimenti provenienti sia da governi che da imprese del settore privato. In particolare, i recenti successi ottenuti nel campo del Deep Learning hanno spinto l'IA a raggiungere uno stadio avanzato di sviluppo e ricerca in svariati settori. I metodi di Machine Learning e Deep Learning hanno dimostrato risultati significativi in numerosi contesti, spaziando dalla logistica alla produzione, dalla elaborazione delle immagini alla generazione di testi. È in virtù di questa flessibilità e versatilità che l'IA può essere agevolmente adattata per affrontare le sfide correlate all'ambito ESG.

Environmental

Un campo di applicazione chiave per l'integrazione del Machine Learning nell'ambito dell'ESG è l'ecologia, rappresentando il primo pilastro degli obiettivi di sviluppo sostenibile (SDG). Secondo lo studio “The role of artificial intelligence in achieving the Sustainable Development Goals” [28], i metodi di intelligenza artificiale hanno il potenziale per contribuire agli obiettivi di sviluppo sostenibile in questo settore fino al 93%. Questi obiettivi comprendono iniziative relative all'azione sul clima, il monitoraggio dei disastri, l'adozione delle energie rinnovabili, la riduzione dell'inquinamento, il ripristino delle foreste e la salvaguardia della biodiversità. Molte di queste sfide possono essere affrontate attraverso l'acquisizione di immagini satellitari oppure aeree, seguita dalla loro elaborazione tramite algoritmi di visione artificiale. Queste immagini raccolgono dati sia ottici che multispettrali e possono essere integrati con altre informazioni rilevate a distanza. Inoltre, l'adozione di telecamere di localizzazione, fisse o montate su veicoli aerei senza pilota (UAV), risulta essere fondamentale per l'osservazione della fauna selvatica nei loro habitat naturali. Solitamente, gli algoritmi di visione artificiale si basano su reti neurali convoluzionali, con alcune delle architetture più comuni, tra cui la famiglia “You Only Look Once” (YOLO) [29] e la famiglia “Single-Shot MultiBox Detector” (SSD) [30].

Un esempio concreto di applicazione del Machine Learning in contesti ambientali può essere trovato nel recente studio “Large-scale forecasting of *Heracleum sosnowskyi* habitat suitability under the climate change on publicly available data” [31]. In questo caso

gli autori analizzano la diffusione della pianta invasiva *Heracleum Sosnowskyi* in alcune regioni dell’Europa settentrionale e centrale. Questa pianta è stata introdotta in Europa negli anni ‘70 e si è diffusa rapidamente, rappresentando una minaccia per la flora autoctona e portando al rischio di scottature cutanee dovute alle sostanze che contiene. Utilizzando dati ambientali pubblicamente disponibili, gli autori hanno sviluppato modelli di previsione per stimare la futura distribuzione di questa pianta entro il 2060, tenendo conto dei diversi scenari climatici e delle emissioni di carbonio. I risultati indicano che c’è una significativa probabilità di espansione della pianta in regioni precedentemente non colpite dal suo sviluppo.

È possibile, tuttavia, monitorare e controllare la diffusione di una pianta tramite l’uso di droni UAV che acquisiscono immagini della flora, consentendo alle aziende forestali locali di prendere le opportune misure. Le immagini satellitari, infatti, rappresentano una risorsa preziosa per l’osservazione su vasta scala del territorio, con dati disponibili da varie piattaforme spaziali come ad esempio *Maxar* e *WorldView*. Questi dati possono essere utilizzati per monitorare sia eventi a breve termine come incendi e inondazioni, sia cambiamenti a lungo termine come alterazioni legate al clima ed all’attività umana. Inoltre, è importante notare che molte di queste immagini satellitari di base, come Google Earth, sono accessibili al pubblico, agevolando l’implementazione degli obiettivi di sviluppo sostenibile. Ad esempio, nell’ambito della gestione forestale, l’uso di algoritmi di Machine Learning è estremamente utile per valutare parametri come l’indice di area fogliare (IAF), la struttura della vegetazione, l’umidità e la densità di alberi per unità di superficie. Questi dati forniscono indicazioni importanti sulla salute generale delle foreste e possono essere utilizzati per valutare la situazione a livello regionale. Un obiettivo correlato è la neutralità del carbonio, che implica il bilanciamento costante delle emissioni di carbonio attraverso l’assorbimento da parte della vegetazione. Questo processo coinvolge modelli biogeochimici speciali come Forest-DNDC [32] e la valutazione del potenziale di assorbimento delle piante rispetto alle emissioni di carbonio locali.

In generale, l’ambiente fornisce una vasta gamma di casi di studio che possono essere affrontati efficacemente con l’uso di algoritmi di IA. Le informazioni raccolte possono essere utilizzate per avere una comprensione approfondita degli oggetti di studio o come dati di *input* per scopi di previsione e classificazione. Date le crescenti preoccupazioni ambientali in molte società, è probabile che nel prossimo futuro ci saranno sempre più opportunità per implementare strumenti di IA in questo settore.

Social

All’interno del documento scientifico intitolato “Practical AI Cases for Solving ESG Challenges” [33], emerge l’ipotesi che l’intelligenza artificiale possa rappresentare una risorsa cruciale nel raggiungimento degli obiettivi di sviluppo sostenibile, costituendo addirittura l’82% delle strategie delineate all’interno della figura. Tra questi obiettivi figurano la riduzione della povertà, l’accesso ad un’istruzione di qualità, e la garanzia di servizi igienico-sanitari e di acqua pulita, come indicato nella recente ricerca “The Role of Artificial Intelligence in Achieving the Sustainable Development Goals” [28].

Allo stesso modo, la pandemia di SARS-CoV-2 ha ulteriormente sottolineato l’importanza cruciale dell’intelligenza artificiale nel modellizzare la diffusione di malattie [34]. Questi sviluppi significativi sollevano questioni cruciali riguardo la capacità dell’intelligenza artificiale di impattare positivamente su questioni di vasta portata, tra cui la salute pubblica ed il benessere sociale. Pertanto, questa prospettiva acquisisce una particolare rilevanza nell’ambito delle sfide legate all’ambiente, al sociale ed alla *governance* (ESG).

Nel contesto degli Obiettivi di Sviluppo Sostenibile (SDG), mostrato nella **Figura 1.4**, si cerca di dare un’enfasi particolare al concetto di “città intelligente”. Queste città sono caratterizzate dall’impiego avanzato della tecnologia per la raccolta e l’efficiente utilizzo dei dati digitali al fine di gestire le infrastrutture e le risorse urbane. Di seguito, sono forniti alcuni esempi di applicazioni di intelligenza artificiale in questo contesto:

- Ottimizzazione delle reti elettriche per adattarle alle esigenze dei clienti e sfruttare diverse fonti di generazione, come parchi eolici e pannelli solari;
- Manutenzione predittiva delle infrastrutture critiche al fine di prevenire incidenti e interruzioni;
- Creazione di ambienti orientati all’umanità che promuovono un’interazione trasparente tra i residenti e le autorità;
- Utilizzo di assistenza sanitaria avanzata per migliorare la diagnosi e la pianificazione del trattamento;
- Ottimizzazione della logistica e della pianificazione per aumentare la sostenibilità dei trasporti;
- Adattamento dei semafori alle condizioni del traffico al fine di evitare ingorghi.

Nel contesto delle città intelligenti e delle sfide legate ai cambiamenti climatici, la risoluzione di tali problemi richiede principalmente l’applicazione di algoritmi di ottimizzazione, classificazione e raggruppamento. In particolare, vengono utilizzati approcci avanzati come le foreste casuali, il potenziamento, le reti neurali Deep Learning ed altri metodi. Inoltre, l’impiego diffuso di sistemi di visione artificiale è sempre più evidente nelle città moderne, grazie alla crescente presenza di telecamere nelle aree urbane.

L’ambito dell’intelligenza artificiale trova applicazione anche nella gestione del territorio e nella revisione catastale. Questo coinvolge l’analisi di immagini satellitari per identificare e classificare elementi sulla superficie terrestre, come campi coltivati, edifici, parchi ed impianti industriali. Tali dati possono poi essere confrontati con le informazioni catastali esistenti. Algoritmi avanzati possono, persino, fornire consulenze ai proprietari terrieri su come ottimizzare lo sviluppo delle loro aree, contribuendo così ad una migliore gestione delle risorse. Un ulteriore esempio, riguarda l’uso dell’intelligenza artificiale nella logistica navale nelle regioni marine con presenza di ghiaccio, dove il cambiamento climatico ha aumentato il traffico marittimo nell’Artico. In quest’ultimo, l’IA viene impiegata per prevedere la formazione del ghiaccio marino e ottimizzare le rotte di navigazione.

In generale, le applicazioni dell'intelligenza artificiale nel contesto sociale sono estremamente varie, coprendo una vasta gamma di settori: dalla gestione delle risorse urbane all'ottimizzazione della logistica e alla previsione della diffusione di malattie. Con la crescita della popolazione urbana e l'adozione sempre più ampia delle tecnologie dell'informazione, è probabile che emergano ulteriori casi di utilizzo pratico dell'intelligenza artificiale in futuro.

Governance

Per quanto riguarda la dimensione della *governance*, nell'ambito dell'intelligenza artificiale, è possibile individuare due prospettive distintive [33]: una di carattere esterna ed una interna.

La prospettiva esterna si focalizza sulle relazioni delle imprese con i vari attori interessati, compresi organismi governativi ed investitori. Tuttavia, a partire dal 2023, è emerso un notevole deficit di un *corpus* normativo universalmente valido relativo all'intelligenza artificiale, il che comporta notevoli sfide nell'adozione di un approccio standardizzato per l'attuazione del Machine Learning. La rapida evoluzione tecnologica, inoltre, impedisce la formulazione di un quadro legislativo stabile e duraturo che non risulti (presto) obsoleto. Di conseguenza, risulta necessario promuovere importanti sforzi di ricerca in quest'ambito ma anche permettere l'elaborazione di nuove leggi coinvolgendo esperti nel campo dell'IA al fine di evitare l'implementazione di politiche inefficaci o, peggio ancora, controproducenti.

Nonostante tali sfide, emergono interessanti applicazioni dell'Intelligenza Artificiale nell'ambito della *governance*. A titolo d'esempio, è possibile citare l'impiego dell'IA per il monitoraggio dell'evasione fiscale o del cosiddetto “*greenwashing*”, ovvero la pratica di promuovere prodotti o organizzazioni come ecologici quando in realtà non risultano esserlo. Nel contesto degli investitori, inoltre, si osserva un'applicazione pratica dell'Intelligenza Artificiale relativa all'analisi delle *performance* aziendali basata su dati pubblici come comunicati stampa, rapporti finanziari, rassegne sociali e articoli di stampa. Questo metodo sfrutta algoritmi di elaborazione del linguaggio naturale (*Natural Language Processing*, NLP) per condurre una rapida analisi del testo, condurre analisi del *sentiment* e generare sintesi informative. Un modello specifico chiamato *esgNLP* [35] è stato sviluppato appositamente per identificare i termini correlati ai rischi ambientali, sociali e di *governance* (ESG) e valutarne il contesto, allo scopo di calcolare un punteggio globale per l'azienda, basato sulle valutazioni positive e negative (*sentiment*) presenti in articoli e rapporti. Questi casi dimostrano come l'IA abbia già un impatto positivo nel settore della *governance*, migliorando la capacità di supervisione e valutazione delle pratiche aziendali, contribuendo così ad incrementare la trasparenza e la responsabilità.

Per quanto riguarda la dimensione interna della *governance*, che verte sull'autovalutazione dell'azienda, l'attenzione si concentra sulla valutazione dell'efficacia delle politiche aziendali e dei processi burocratici. In questo contesto, gli strumenti più efficaci sfruttano la tecnologia di Elaborazione del Linguaggio Naturale (NLP) per automatizzare diverse attività, quali la verifica della coerenza dei documenti, la generazione di *report* basati su

modelli predefiniti, l'implementazione di motori di ricerca, la creazione di basi di conoscenza e il supporto di *chatbot* per assistere i dipendenti. Questo approccio si distingue dall'automazione dei processi robotici standard (Robotic Process Automation, RPA [36]), la quale si limita a replicare in sequenza azioni precedentemente compiute dagli utenti.

In sintesi, la principale tecnologia di Intelligenza Artificiale impiegata nell'ambito della *governance* è rappresentata dalla tecnologia NLP, orientata all'analisi del testo. Inoltre, mentre l'aspetto esterno richiede ulteriori ricerche, le applicazioni interne hanno dimostrato l'efficacia dell'IA. Con i recenti progressi nei modelli linguistici avanzati, si assiste all'emergere di nuove e promettenti applicazioni dell'IA in questo settore.

1.2.5 Adeguamento ai fattori ESG delle imprese italiane

Il rapporto ESG Outlook, fornito dalla società Centrale Rischi Finanziari (CRIF), offre un'analisi dettagliata delle questioni legate all'ESG in Italia, concentrando su imprese, individui ed immobili. Questa analisi si basa su una vasta quantità di dati e strumenti di analisi ESG accumulati nel corso degli anni dalla società CRIF. In particolare, il rapporto fornisce una panoramica approfondita sulla valutazione dell'adeguatezza ESG delle aziende italiane, tenendo conto delle loro caratteristiche, come dimensione, settore ed area geografica. Per lo sviluppo della ricerca, CRIF ha proceduto, alla fine del 2022, ad una selezione accurata di circa 150.000 imprese italiane rappresentative. Attraverso un'analisi sulla base del proprio patrimonio informativo, sono stati in grado di fornire un quadro originale riguardo le sfide legate alla sostenibilità. Un elemento fondamentale nell'ambito di questa analisi è rappresentato, appunto, dal punteggio ESG, il quale riassume il grado di aderenza di ciascuna impresa ai principi di sostenibilità tenendo conto del settore di appartenenza e della zona geografica in cui opera. Il punteggio ESG di CRIF sintetizza più di 150 indicatori relativi ai componenti *Environmental* (E), *Social* (S) e *Governance* (G), organizzati successivamente in base alle categorie informative definite dalla normativa come *EBA* (*European Banking Authority*) Factor [37].

I risultati attualmente mostrano, come riportato nella **Figura 1.5**, che l'8% delle aziende ha un basso livello di adeguatezza ESG, il 60% ha un livello medio-basso, mentre oltre il 30% si trova in una fase avanzata del percorso di transizione verso un'economia più sostenibile. In particolare, le aziende con un fatturato superiore a 10 milioni di euro risultano essere più avanzate nel loro impegno verso la sostenibilità, con il 39% di esse che raggiunge livelli alti, o molto alti, di adeguatezza ESG. Le piccole e medie imprese (PMI), con un fatturato inferiore a 10 milioni di euro, costituiscono, appunto, il settore che richiede il maggior sostegno nella sua transizione verso la sostenibilità.

Nell'ambito della valutazione complessiva della sostenibilità delle PMI, uno dei principali aspetti presi in considerazione riguarda i fattori legati all'ambiente, notando che attualmente questo rappresenta un aspetto molto rilevante anche dal punto di vista delle autorità di regolamentazione. CRIF ha effettuato una valutazione dell'adeguatezza delle PMI nella gestione dei rischi ambientali mediante l'impiego di uno strumento denominato "Score Ambientale" (conosciuto come "Score E"). L'analisi condotta ha messo in luce notevoli differenze tra le PMI italiane, sia in termini geografici (con variazioni significative

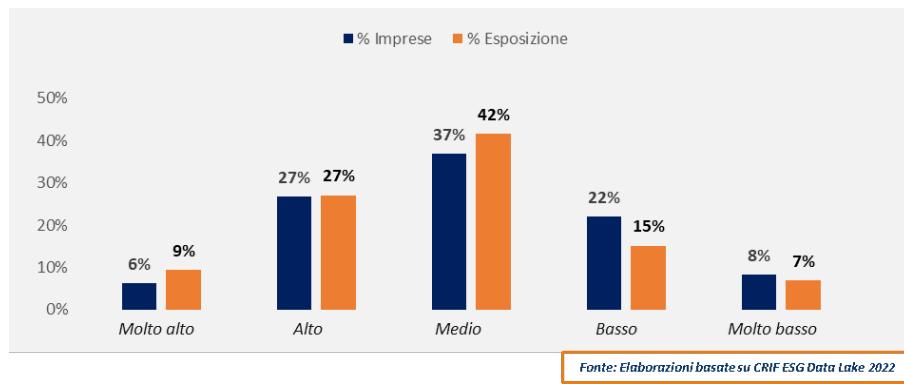


Figura 1.5. Distribuzione dei livelli di adeguatezza ESG per numero di imprese italiane e per esposizione creditizia.

Fonte: ESG Outlook rilasciato da CRIF, giugno 2023

tra le regioni) che settoriali. In particolare, le regioni della Lombardia e del Piemonte hanno registrato i punteggi ambientali più elevati, con oltre il 60% delle aziende che si sono dimostrate altamente adeguate in materia di sostenibilità ambientale. Inoltre, tra i settori economici presi in esame, il comparto immobiliare e le attività legate al tempo libero (“leisure”) hanno dimostrato di essere particolarmente performanti secondo il medesimo *score*.

Un altro aspetto di notevole importanza, esaminato nell’ambito dell’ESG Outlook di CRIF, è la valutazione dell’effetto potenziale dei rischi legati all’ambiente, in particolare quelli derivanti dal cambiamento climatico e dal deterioramento ambientale, sull’aspetto economico e finanziario. Questa valutazione si articola in due categorie principali: i rischi cronici, che sono collegati ai cambiamenti climatici in corso in maniera graduale, ed i rischi acuti, rappresentati da eventi improvvisi come catastrofi naturali. Come mostra la **Figura 1.6**, si è riscontrato che il 5,9% delle piccole e medie imprese presenta un elevato o molto elevato grado di esposizione ai rischi fisici acuti. Per quanto riguarda i rischi fisici cronici (**Figura 1.7**), si è osservato che il 16% delle imprese si trova in una posizione di elevata esposizione, per cui con un grado di rischio definito come alto o molto alto.

L’impatto derivante dalla transizione verso la sostenibilità è particolarmente pronunciato nel settore dell’estrazione mineraria. Nel contesto dell’ESG Outlook, è stato sviluppato un modello proprietario al fine di valutare a lungo termine gli effetti finanziari di tale transizione. Questo modello tiene in considerazione diversi fattori, inclusi i costi, i ricavi e gli investimenti, fornendo una prospettiva chiara sui possibili sviluppi futuri. I risultati ottenuti mettono in luce una notevole variabilità nei costi associati alla transizione verso un’economia sostenibile tra i vari settori industriali. In particolare, i costi connessi a questa transizione, che includono sia quelli diretti, come le tasse sul carbonio, sia gli investimenti, espressi come percentuale del fatturato, mostrano un’ampia gamma di variazioni. I settori ad elevata intensità energetica, tra cui l’estrazione mineraria, i trasporti, la chimica e la lavorazione dei metalli, evidenziano impatti significativi, con un’attesa percentuale di costi che oscilla tra il 3% e l’8% del fatturato annuo, come mostrato nella **Figura 1.8**.

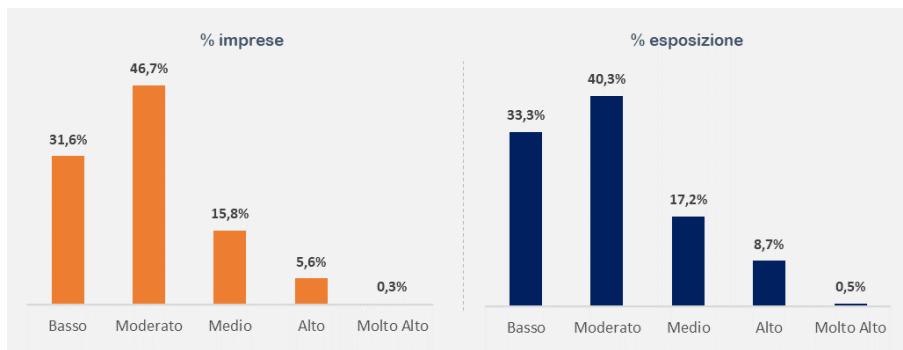


Figura 1.6. Distribuzione della percentuale di PMI italiane per livello di esposizione ai rischi fisici acuti e cronici. Il grafico a sinistra illustra la distribuzione percentuale delle PMI italiane in base al livello di rischio fisico cronico. Le categorie di rischio vanno da “Basso” a “Molto Alto”, con la percentuale più elevata (46,7%) che rientra nella categoria “Moderato”. Le imprese con un livello di rischio “Molto Alto” sono significativamente meno, rappresentando solo lo 0,3%. Sul lato destro, il grafico evidenzia le percentuali di esposizione delle PMI ai rischi fisici cronici, con una distribuzione che privilegia il livello “Medio” con il 43,1% delle imprese. La categoria “Alto” include il 26,5% delle imprese e quelle in “Molto Alto” sono il 7,2%, indicando che un considerevole segmento di imprese affronta rischi significativi.

Fonte: ESG Outlook rilasciato da CRIIF, giugno 2023

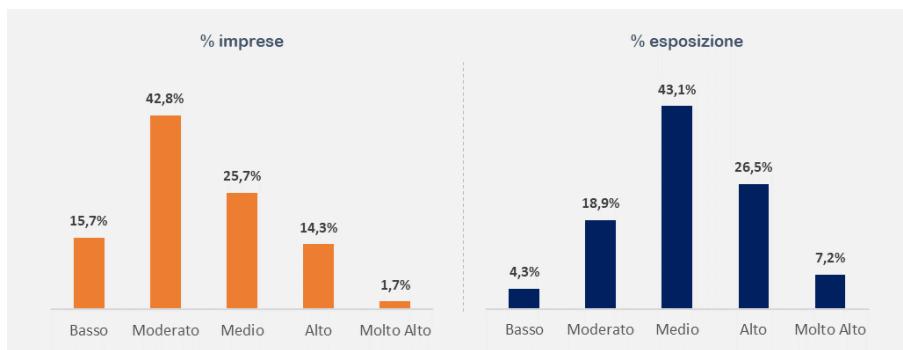


Figura 1.7. Distribuzione percentuale di PMI italiane in base al loro livello di esposizione ai rischi fisici cronici, suddiviso in cinque categorie di rischio: “Basso”, “Moderato”, “Medio”, “Alto” e “Molto Alto”. La percentuale più alta di imprese, pari al 42,8%, si colloca nella categoria di rischio “Moderato”. La categoria “Alto” include il 14,3% delle imprese, mentre quelle classificate con un rischio “Molto Alto” costituiscono il 1,7%. Il grafico a sinistra mostra le percentuali delle imprese in ciascuna categoria di rischio, con un evidente incremento delle imprese che si trovano nelle categorie di rischio “Moderato” e “Medio”. Le categorie di rischio più elevate (“Alto” e “Molto Alto”) rappresentano insieme il 16% delle PMI, segnalando una significativa esposizione a rischi fisici cronici legati a cambiamenti climatici progressivi.

Fonte: ESG Outlook rilasciato da CRIIF, giugno 2023

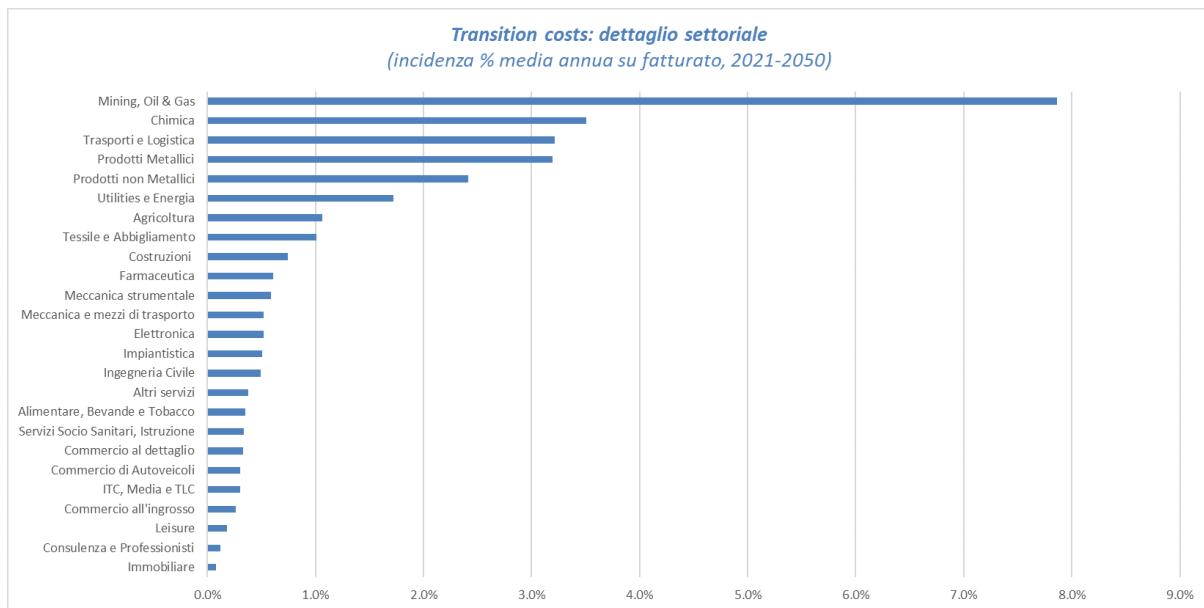


Figura 1.8. Costi di transizione per settore (incidenza della percentuale media annua sul fatturato, 2021-2050).

Fonte: ESG Outlook rilasciato da CRIF, giugno 2023

Settori come la lavorazione di prodotti non metallici, la produzione e la distribuzione di elettricità e gas presentano impatti moderati ma comunque rilevanti, con una percentuale di costi prevista di circa il 2% al 3% del fatturato annuo. Al contrario, i settori correlati ai servizi, alle attività immobiliari ed al commercio mostrano un impatto marginale, con una percentuale inferiore all'0,5% del fatturato annuo. In generale, emerge una stretta correlazione tra il livello attuale di intensità delle emissioni e l'entità dell'impatto derivante dalla transizione.

L'indagine sul fattore *Sociale* concentra la sua attenzione sul benessere sociale all'interno delle organizzazioni aziendali, nonché sull'impatto che queste esercitano sul territorio e sulla comunità circostante. Dal *report* emerge che l'attenzione rivolta alle questioni sociali è strettamente correlata alle dimensioni medie delle aziende, evidenziando una crescente relazione positiva tra il fatturato e l'efficacia delle politiche sociali adottate, e risulta più sviluppata nel Nord Italia e nel Lazio. In particolare, nel settore farmaceutico si registra un punteggio sociale (“Score S”) notevolmente elevato, con il 48% delle imprese che presentano un elevato livello di adeguatezza sociale, mentre complessivamente il 78% delle aziende ottenute punteggi sociali rilevanti. Al contrario, più di un terzo delle aziende agricole evidenzia un ritardo significativo in termini di adeguatezza sociale, probabilmente influenzato dalla dimensione ridotta di molte imprese operanti in questo settore, il che rende più complesso garantire adeguate tutele in termini di sicurezza e stabilità delle condizioni lavorative.

Anche nell'analisi del fattore *Governance* (“Score G”), che esamina aspetti quali la capacità di gestire la diversità all'interno di un'azienda, la sua trasparenza e l'etica delle sue pratiche aziendali, emerge la conferma della correlazione tra il punteggio ottenuto e

le dimensioni dell’azienda, simile a quanto osservato nel contesto del fattore sociale. Le imprese di maggiori dimensioni tendono spesso a strutturarsi in modo da offrire una maggiore divulgazione di informazioni verso l’esterno, attraverso l’adozione di valutazioni di legalità, codici etici e/o la pubblicazione di bilanci certificati su base volontaria. Nel settore farmaceutico si registra una performance superiore, seguito da quello dell’elettronica, meccanica e mezzi di trasporto, mentre le aziende agricole, del settore delle costruzioni, turistiche ed immobiliari presentano percentuali inferiori in termini di punteggi di *governance*.

Il rapporto evidenzia, anche, che circa un quarto degli immobili in Italia è esposto ad un alto o molto alto rischio fisico, con rischio sismico rilevante soprattutto nelle regioni più popolate. Le perdite dovute ai quattro principali rischi fisici, come il vento, le frane, le inondazioni ed i terremoti, sono inferiori allo 0,2% per circa la metà degli immobili residenziali. Infine, il rapporto dedica un capitolo al contributo degli attori finanziari per favorire la transizione verso la sostenibilità e presenta il progetto TranspArEEEnS [38], nato nel 2021 con l’obiettivo di fornire alle piccole e medie imprese uno strumento di valutazione dei rischi ESG basato su dati qualitativi e quantitativi.

1.2.6 IA sostenibile

Nel corso di questa trattazione, si è esaminato il Machine Learning (ML) unicamente come uno strumento destinato a risolvere i problemi legati all’ambiente, alle questioni sociali ed alla *governance* (ESG). Tuttavia, è necessario notare che gli stessi approcci di Intelligenza Artificiale possono svolgere un ruolo centrale all’interno dell’analisi ESG [39]. Tale situazione si verifica quando un’azienda sfrutta ampiamente l’utilizzo di algoritmi di ML, consentendo loro di riorganizzare in modo automatico le risorse e regolare i parametri di produzione senza richiedere intervento umano diretto. Un esempio lo si trova nel contesto della selezione del regime operativo ottimale per una centrale termoelettrica: se tali algoritmi non vengono preventivamente dotati di direttive mirate a mitigare l’impatto ambientale e la quantità di carbonio emesso, essi non terranno mai in considerazione le conseguenze ecologiche delle loro decisioni. In tal caso, è probabile che scelgano un regime di funzionamento economicamente vantaggioso ma dannoso per l’ambiente, privilegiando la riduzione dei costi rispetto alla massimizzazione della produzione di energia elettrica.

Un’altra problematica riguarda l’ispezione etica dei sistemi di intelligenza artificiale. Tale questione deve essere affrontata se si intende garantire che gli algoritmi di Machine Learning prendano decisioni corrette, sicure ed imparziali anche quando addestrati su insiemi di dati non rappresentativi [28]. Nell’articolo “What about investors? ESG analyses as tools for ethics-based AI auditing”[39] vengono elencati più di 173 quadri di controllo per l’IA, tuttavia emerge la mancanza di parametri e standard universalmente accettati. In generale, c’è un consenso sull’applicazione di diversi criteri ai sistemi di IA, tra cui l’etica, la trasparenza, la responsabilità, l’equità e la *privacy* [40]. Inoltre, gli algoritmi devono essere progettati per essere robusti, in modo da gestire possibili incongruenze nei dati di *input*. La sfida principale in questo contesto consiste nel tradurre questi criteri di riferimento generali in misure pratiche, formulando norme rigorose e fornendo descrizioni quantitative in modo che le parti coinvolte possano applicarle nella loro attività quotidiana.

na. Una soluzione pragmatica potrebbe essere quella di adottare uno dei quadri normativi esistenti come base legislativa, con il contributo di esperti di Intelligenza Artificiale.

Attualmente, vi è un dibattito esteso riguardo all'impatto sociale dell'Intelligenza Artificiale, considerata una tecnologia innovativa [41]. Nonostante la creazione di nuove posizioni lavorative nell'ambito dell'informatica, l'applicazione del Machine Learning spesso porta all'automatizzazione di specifici processi aziendali e tecnologici. Di conseguenza, le aziende stanno esaminando la possibilità di eliminare ruoli meno qualificati, il che comporta tagli salariali oppure la perdita di lavoro per alcuni dei loro dipendenti, il cui lavoro è diventato obsoleto [28]. Inoltre, le aziende che non riescono ad implementare prontamente l'impiego dell'Intelligenza Artificiale rischiano di diventare meno competitive sul mercato e di vedere diminuire le opportunità di ottenere finanziamenti dagli investitori.

È importante notare che le imprese private tendono a concentrarsi principalmente su settori che offrono vantaggi economici evidenti, come ottimizzare la produzione in una struttura di assemblaggio, mentre problemi con un impatto meno immediato sul modello di *business* diventano meno attraenti, ad esempio, l'implementazione di un ufficio intelligente in una piattaforma petrolifera. Questo può portare a progressi disomogenei verso determinati obiettivi di sviluppo sostenibile. Questo viene accentuato dalle differenze nelle capacità di finanziamento, ovvero con le grandi aziende che hanno maggiori risorse da investire nell'Intelligenza Artificiale rispetto alle piccole imprese [28]. Probabilmente, le soluzioni più efficaci in questa situazione includono il potenziamento della consapevolezza pubblica sull'Intelligenza Artificiale, la promozione dell'acquisizione di competenze professionali nel campo del Machine Learning da parte dei dipendenti ed un aumento del finanziamento per applicazioni senza scopo di lucro.

Numerose applicazioni del Machine Learning, come il monitoraggio ed il riconoscimento delle persone, sollevano questioni etiche, specialmente quando vengono combinati con la valutazione e la classificazione dei cittadini [42]. Siccome diversi paesi presentano *background* culturali e politici diversi, un'applicazione di algoritmi di ML per il pubblico che ha funzionato positivamente all'interno di un contesto, ad esempio per le raccomandazioni all'interno di un *social network*, può avere conseguenze significativamente diverse in altri scenari [28]. Pertanto, l'uso di tali strumenti richiede una notevole cautela e la necessità di esser verificati preventivamente rispetto al pubblico di riferimento.

Un altro aspetto interessante riguarda le cosiddette *fake news*, ossia la divulgazione di informazioni false o fuorvianti presentate come veritieri. Queste notizie si diffondono regolarmente all'interno dei media, spesso trattando argomenti sensazionali come scandali di celebrità, elezioni truccate o proposte di tagli fiscali [43]. Dopo la diffusione di una *fake news*, risulta estremamente difficile contrastarla poiché il pubblico tende ad essere suscettibile alle emozioni e continua a diffondere dati errati. Questa situazione è ulteriormente complicata per l'avanzamento dell'Intelligenza Artificiale, che fornisce potenti strumenti per la generazione di notizie false, specialmente tramite Generative Adversarial Network (GAN) [44]. Le GAN sono in grado di replicare in modo efficace il contenuto educativo, rendendolo quasi indistinguibile dal materiale autentico, agli osservatori. Queste tecnologie possono facilmente manipolare video e immagini, sostituendo volti e sfondi (creando

i cosiddetti *deepfake*), potenzialmente causando gravi danni economici o danneggiando la reputazione [45]. Finora sono stati proposti diversi approcci pratici per individuare e gestire contenuti grafici falsi [46] ma è necessaria anche una ricerca continua per migliorarne le prestazioni e contrastare questo sviluppo [47].

Un’ulteriore tematica, che richiede un maggior approfondimento, è rappresentata da ChatGPT [48]. Questo strumento basato sull’Intelligenza Artificiale è in grado di replicare in modo efficace un interlocutore conducendo discussioni, scrivendo testi e generando codice di programmazione di base. Indubbiamente, ChatGPT possiede un notevole potenziale. Tuttavia, la grande attenzione mediatica e le aspettative del pubblico stanno ostacolando una valutazione ponderata del suo valore pratico nel settore, specialmente all’inizio del 2023. Per funzionare efficacemente all’interno di un contesto aziendale, ChatGPT richiede un processo di formazione, l’accesso costante ai dati ed alla documentazione interna dell’azienda. Tuttavia, quest’accesso potrebbe entrare in conflitto con i servizi di sicurezza delle informazioni dell’azienda, i quali hanno il compito principale di evitare la divulgazione di informazioni sensibili, come dati personali, dati finanziari o conoscenze procedurali (“know-how”). Pertanto, il primo passo dovrebbe essere l’adeguamento delle aspettative sia da parte dei potenziali utenti di ChatGPT che dell’Information Security Management System (ISMS, ovvero una norma su base volontaria utile per gestire i sistemi informatici e la sicurezza dei dati in modo completo ed efficace [49]).

Recentemente, però, sono stati registrati significativi sviluppi in relazione a questo ambito. Alla fine di Agosto 2023, infatti, la società OpenAI ha introdotto il prodotto denominato ChatGPT Enterprise [50], il quale offre l’opportunità di accedere a GPT-4 ad elevate velocità e senza restrizioni, consentendo l’uso di *input* più estesi e con un contesto composto da ben 32.000 *token*. Questa versione avanzata permetterà, inoltre, l’analisi dei dati in maniera illimitata, offrirà modelli di *chat* condivisibili all’interno dell’organizzazione, una piattaforma di amministrazione dedicata ed il supporto per l’autenticazione Single Sign-On (SSO), che consente agli utenti di accedere ad una sessione utilizzando un unico set di credenziali e di ottenere in modo sicuro l’accesso a diverse applicazioni e servizi correlati, senza dover effettuare ulteriori accessi. L’obiettivo di tale innovazione include, anche, la fornitura di crediti API per la creazione di soluzioni personalizzate e la garanzia che i dati aziendali rimarranno inaccessibili alla piattaforma. Perciò, ChatGPT Enterprise migliorerà notevolmente l’esperienza offerta fino ad oggi da GPT-4, affrontando alcune delle criticità precedentemente attribuite ad OpenAI da parte dei datori di lavoro. Fino allo sviluppo completo di questa tecnologia, però, è necessario riorganizzare l’ambiente IT aziendale in modo da consentire ai sistemi come ChatGPT di esaminare esclusivamente dati generici e innocui, preservando al contempo la loro funzionalità nell’agevolare gli utenti nelle attività pratiche, come la gestione delle riunioni e la creazione di promemoria.

Da notare, infine, che i metodi di Machine Learning stanno diventando sempre più esigenti dal punto di vista energetico. Ad esempio, per addestrare una rete NLP all’avanguardia come GPT-3.5, sono richiesti quasi 936 MWh, una quantità di energia equivalente a quella consumata da 468.000 bollitori elettrici medi (2 kWh) o 12.480 batterie per auto Tesla con batteria da 75 kWh. Si stima che entro il 2030, le tecnologie dell’informazione e della comunicazione, compresa l’Intelligenza Artificiale, consumeranno il 20% dell’elettricità

mondiale, con conseguenze significative sull'impronta di carbonio se questa problematica non verrà affrontata in anticipo [51].

La complessità computazionale di molti problemi legati al Machine Learning sta anche incrementando la domanda di unità di elaborazione grafica, comunemente utilizzate in calcoli intensivi. La produzione di queste unità comporta l'impiego di materiali tossici, con gravi impatti ambientali [52]. In particolare, acidi come l'arsenico, la fosfina, il solfidrico e il fluoridrico sono ampiamente utilizzati in diverse fasi della fabbricazione dei semiconduttori, comportando rischi diretti per la salute umana e aumentando le probabilità di cancro a lungo termine. Queste problematiche potrebbero essere parzialmente mitigate tramite l'implementazione dell'"Intelligenza Artificiale Verde", basata su reti neurali appositamente progettate per un minore consumo energetico e, di conseguenza, un minore impatto ecologico [53].

1.3 Proposta europea per un quadro giuridico sull'IA (AI Act)

All'interno del contesto digitale, l'Unione Europea si propone di regolamentare l'Intelligenza Artificiale al fine di garantire condizioni più favorevoli per lo sviluppo e l'utilizzo di questa innovativa tecnologia. L'IA, in virtù delle sue potenzialità, quali il miglioramento dei servizi sanitari, dei trasporti più sicuri ed ecologici, di una produzione più efficiente e di un'energia più economica e sostenibile, rappresenta un'opportunità considerevole. Per questo motivo, il 14 giugno 2023 il Parlamento Europeo ha approvato una successione di emendamenti proposti nel normativa nota come "Artificial Intelligence Act" (AI Act) [54], mirati a regolamentare le intelligenze artificiali all'interno dell'Unione Europea. I dettagli di tale legislazione saranno oggetto di discussione con gli Stati membri dell'Unione Europea, con l'obiettivo di giungere alla ratifica definitiva dell'AI Act. Questo processo di negoziazione rappresenta un elemento cruciale del sistema legislativo dell'Unione Europea e sarà essenziale per garantire che il regolamento sull'AI Act sia bilanciato ed efficace. Il 9 Dicembre 2023 Commissione, Consiglio e Parlamento europeo hanno approvato l'accordo provvisorio relativo all'AI Act, nelle settimane successive sono stati poi proseguiti i lavori a livello tecnico per definire i dettagli del nuovo regolamento. Al termine di tali lavori, la presidenza presenterà il testo di compromesso al Comitato dei rappresentanti permanenti degli Stati membri (*Coreper*), incaricato di preparare i lavori del Consiglio dell'Unione Europea, per ottenere l'approvazione formale. La validazione del testo completo richiederà la conferma da parte di entrambe le istituzioni coinvolte, seguita da un'accurata revisione giuridico-linguistica. In conformità con l'accordo provvisorio, è previsto che il regolamento sull'IA entri in vigore il 1° Gennaio 2026, dando tempo ai produttori di sistemi di IA fino al 1° Gennaio 2029 per conformarsi alle nuove disposizioni.

L'iter della proposta di regolamento sull'AI Act ha avuto origine il 12 Aprile 2021, quando la Commissione Europea presentò l'iniziativa per promuovere un utilizzo "affidabile" dell'Intelligenza Artificiale all'interno dell'Unione Europea. Da allora, la proposta è stata passata attraverso diverse fasi di revisione e negoziazione fino alla formulazione dell'"Artificial Intelligence Act" nel Giugno 2023.

1.3.1 Quadro teorico

Il quadro teorico proposto si orienta verso un paradigma di valutazione del rischio connesso al potenziale danno che l’Intelligenza Artificiale può causare ai diritti fondamentali degli individui. Tale approccio implica che l’entità dei doveri e delle regolamentazioni sarà proporzionale al grado di rischio associato. Sebbene molti sistemi di Intelligenza Artificiale comportino rischi minimi, è meglio procedere con una valutazione attenta e rigorosa. Pertanto, come mostrato in **Figura 1.9**, si distinguono [55]:

- **Rischio inaccettabile.** Sistemi di Intelligenza Artificiale considerati una minaccia per le persone e, per questo motivo, verranno vietati. Essi includono:
 - Manipolazione cognitivo comportamentale di persone o specifici gruppi vulnerabili, ad esempio giocattoli ad attivazione vocale che incoraggiano comportamenti pericolosi nei bambini;
 - Punteggio sociale, ovvero classificazione delle persone in base al comportamento, allo stato socioeconomico o alle caratteristiche personali;
 - Utilizzo di materiale protetto da *copyright* nell’addestramento di intelligenze artificiali generative;
 - Sistemi di identificazione biometrica in tempo reale e remota, come il riconoscimento facciale, che utilizzano dati sensibili, come genere, razza, etnia, stato di cittadinanza, religione e orientamento politico.

Tuttavia, è necessario notare che durante i negoziati con il Consiglio europeo potrebbero emergere alcune modifiche rispetto ad un divieto totale di identificazione biometrica poiché alcune situazioni possono richiedere l’uso di IA, secondo quanto indicato da molti corpi di polizia nazionali, per scopi di contrasto alla criminalità. Ad esempio, i sistemi di identificazione biometrica remota “post”, in cui l’identificazione avviene dopo un ritardo significativo, saranno autorizzati a perseguire reati gravi ma solo dopo l’approvazione del tribunale.

- **Alto rischio.** Sistemi di Intelligenza Artificiale che influiscono negativamente sulla sicurezza o sui diritti fondamentali. Essi saranno suddivisi in due categorie:
 - Sistemi di IA utilizzati nei prodotti che rientrano nella legislazione sulla sicurezza dei prodotti dell’Unione Europea, questo include giocattoli, aviazione, automobili, dispositivi medici e ascensori;
 - Sistemi di IA rientranti in otto ambiti specifici che dovranno essere registrati in un database UE:
 - * Identificazione biometrica e categorizzazione delle persone fisiche;
 - * Gestione e funzionamento delle infrastrutture critiche;
 - * Istruzione e formazione professionale;
 - * Occupazione, gestione dei lavoratori e accesso al lavoro autonomo;
 - * Accesso e godimento dei servizi privati essenziali e dei servizi e benefici pubblici;
 - * Applicazione della legge;

- * Gestione della migrazione, dell’asilo e del controllo delle frontiere;
- * Assistenza nell’interpretazione giuridica e nell’applicazione della legge.

Ogni sistema di Intelligenza Artificiale ad alto rischio verrà sottoposto a valutazione sia prima della sua commercializzazione che in modo continuativo durante il suo intero ciclo di vita.

- **Rischio limitato.** Sistemi di Intelligenza Artificiale che dovrebbero soddisfare requisiti minimi di trasparenza in modo da permettere agli utenti di prendere decisioni informate. Dopo aver interagito con le applicazioni, come ad esempio i *ChatBot*, l’utente può decidere se desidera continuare ad utilizzarle. Gli utenti dovrebbero ricevere informazioni quando interagiscono con uno strumento di IA, incluso nei casi di sistemi che generano o manipolano contenuti di immagini, audio o video, come ad esempio i *deepfake*. Il *deepfake* è una tecnica di manipolazione multimediale avanzata che utilizza l’apprendimento automatico, in particolare le reti neurali artificiali, per creare contenuti audiovisivi, come video e audio, in cui il volto e la voce di una persona vengono sostituiti in modo realistico da quelli di un’altra, spesso senza il consenso o la conoscenza delle persone coinvolte. Questa tecnologia può essere utilizzata per scopi creativi ma solleva gravi preoccupazioni legate alla manipolazione e alla diffusione di contenuti falsi o fraudolenti.
- **Rischio minimo o nullo.** In questo contesto non sono previsti obblighi e l’uso dell’Intelligenza Artificiale è consentito in modo libero. Vengono classificati come sistemi a rischio minimo o nullo applicazioni come i videogiochi potenziati dall’Intelligenza Artificiale o i filtri *antispam*.

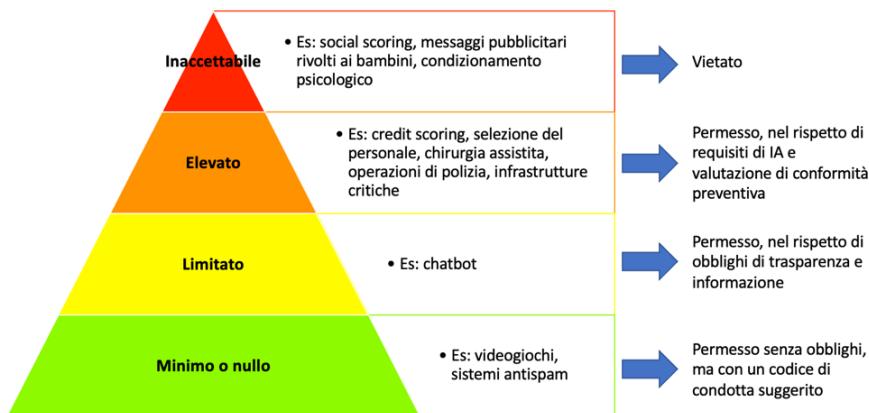


Figura 1.9. Schema piramidale in cui vengono classificati i sistemi di Intelligenza Artificiale in base al loro potenziale rischio per i diritti fondamentali e la sicurezza dei cittadini europei.

Fonte: Istituto per gli Studi di Politica Internazionale (ISPI)

Tali emendamenti introducono, pertanto, restrizioni e divieti riguardo a determinati utilizzi dell’Intelligenza Artificiale, mirando a garantire una maggiore sicurezza e tutela dei

diritti. Ad esempio imponendo il divieto della polizia predittiva basati su profilazione, posizione o precedenti penali, così come i sistemi di riconoscimento delle emozioni impiegati dalle forze dell'ordine, alle frontiere, nei luoghi di lavoro e nelle scuole ed il divieto di creare basi di dati per il riconoscimento facciale basati sull'estrazione non mirata di dati biometrici dal *web* o da sistemi di videosorveglianza a circuito chiuso.

Il Parlamento europeo ha definito, più precisamente, che i sistemi di IA con potenziale impatto sugli elettori e sui risultati delle elezioni, nonché quelli impiegati negli algoritmi di suggerimento delle piattaforme di *social media* ed altre piattaforme digitali, devono essere categorizzati come ad alto rischio. In aggiunta, il testo del regolamento, approvato dal Parlamento, pone un'enfasi sull'importanza di assicurare che i cittadini abbiano il diritto di presentare reclami e ricevere spiegazioni per le decisioni prese attraverso l'uso di sistemi di Intelligenza Artificiale ad elevato rischio, mettendo così al centro delle considerazioni le necessità e gli interessi umani nell'interazione con le macchine.

All'interno dell'AI Act, il Parlamento europeo ha riconosciuto la cruciale necessità di trovare un punto di equilibrio tra la regolamentazione del rapido avanzamento della tecnologia dell'IA, considerando le sue implicazioni per la vita dei cittadini e la preoccupazione di non ostacolare l'innovazione oppure di danneggiare le imprese di minori dimensioni, che potrebbero non essere in grado di sopportare regolamentazioni eccessivamente stringenti. Di conseguenza, sono state inserite alcune clausole per esentare le attività di ricerca e le componenti di IA offerte con licenze *open source*, al fine di promuovere la crescita e la collaborazione all'interno della comunità scientifica e tecnologica. Inoltre, sono stati istituiti ambienti sperimentali regolatori, che consentono di testare i sistemi di IA e valutare i loro impatti in termini di etica, sicurezza ed effetti sociali prima della loro introduzione sul mercato. Tuttavia, sorge un interrogativo in merito all'AI Act, in quanto il regolamento, una volta adottato, prevede un periodo di due anni durante il quale le normative saranno applicate in modo più flessibile, concedendo alle aziende il tempo necessario per adeguarsi. Poiché è probabile che questa regolamentazione non entri in vigore prima del 2026, l'Unione europea potrebbe dover affidarsi principalmente alla conformità volontaria delle aziende che operano nel settore dell'Intelligenza Artificiale, allo scopo di garantire immediatamente trasparenza, sicurezza e rispetto dei diritti e della *privacy* dei cittadini.

1.3.2 AI Act: contesto globale

L'AI Act, rappresenta quindi la prima legge al mondo creata per regolamentare l'Intelligenza Artificiale. Questo importante passo, rende l'Unione Europea un pioniere nella promulgazione di una legislazione completa per la supervisione dei sistemi di IA. Questo ha suscitato aspettative di un seguito simile da parte di altri organi legislativi, tra cui gli Stati Uniti d'America, che, sotto l'amministrazione Trump, avevano adottato una politica di “*light touch*” verso la regolamentazione dell'IA, con un'enfasi sugli investimenti in ricerca e sviluppo e sulla sicurezza etica [56]. Con l'avvento dell'amministrazione Biden, invece, gli Stati Uniti hanno intensificato l'attenzione verso l'etica e la responsabilità nell'utilizzo dei sistemi di IA. Sebbene sia stata proposta una normativa simile all'AI Act, denominata “Algorithm Accountability Act”, questa è ancora in fase di discussione. La recente riunione tra gli Stati Uniti e l'Unione Europea all'interno del Trade and Tech-

nology Council (organismo politico transatlantico che funge da forum diplomatico per coordinare la tecnologia e la politica commerciale tra gli Stati Uniti e l'Unione europea) ha evidenziato l'obiettivo comune di evitare un potenziale vuoto legislativo nell'ambito dell'IA tra le due regioni, che potrebbe insorgere con l'approvazione definitiva dell'AI Act europeo. Per affrontare questa sfida, è in fase di sviluppo un codice di condotta volontario relativo all'utilizzo dell'Intelligenza Artificiale [57]. Una volta completato, questo codice sarà presentato come proposta transatlantica congiunta ai *leader* del G7, con l'obiettivo di incentivare le aziende ad adottarlo. Il G7 è costituito dalle sette principali economie mondiali, ovvero Canada, Francia, Germania, Giappone, Italia, Regno Unito e Stati Uniti d'America. Queste nazioni sviluppate detengono un ruolo di centrale importanza a livello globale, sia dal punto di vista politico che economico, industriale e militare.

In questo modo, mentre l'approvazione dell'Artificial Intelligence Act da parte del Parlamento europeo rappresenti un passo fondamentale nella regolamentazione dell'IA in Europa, il codice di condotta sull'IA potrebbe svolgere un ruolo cruciale nell'armonizzazione delle pratiche aziendali a livello globale relative ai sistemi di Intelligenza Artificiale, garantendo un utilizzo sicuro, etico e trasparente.

2. Descrizione del progetto

Nel seguente capitolo, viene eseguita un'approfondita analisi della tecnologia utilizzata all'interno di questo progetto, concentrando-si principalmente sugli strumenti denominati “Large Language Models” (LLM). Viene fornito un contesto alla ricerca, spiegando le caratteristiche distintive di tali modelli ed evidenziando la differenza con la denominazione “Generative AI”. In particolare, questi ultimi costituiscono la categoria più estesa di modelli in grado di produrre una varietà di *output*, mentre gli LLM costituiscono un gruppo specifico, all'interno della precedente categoria, specializzato nella generazione di testo. In seguito, si esplorano le origini storiche e l'evoluzione di tali modelli di linguaggio, partendo dal test di Turing, passando per i Statistical Language Models, attraversando gli inverni dell'IA, sino a giungere a `word2vec` e `Transformers`. Questo approfondimento permette di tracciare il contesto storico e l'evoluzione tecnologica nel corso del tempo. Successivamente, si fornisce una descrizione dell'architettura degli algoritmi di “apprendimento auto-supervisionato”, approccio alla base dei LLM, che consente ai modelli di imparare dai dati non annotati in una fase precedente ma di sfruttare la struttura intrinseca dei propri *input*. Nel contesto dei Large Language Models (LLM), questa metodologia suggerisce che il modello acquisisca la comprensione del linguaggio anticipando e predicendo segmenti di testo all'interno di un contesto più ampio. In altre parole, il modello impara il linguaggio analizzando ed anticipa frammenti di testo, non limitandosi ad esaminare porzioni in isolamento ma cercando di cogliere il significato e le connessioni più ampie presenti nel contesto circostante. Tale metodo è fondamentale per migliorare la capacità di generare testo coerente e significativo, consentendo ai LLM di acquisire una profonda comprensione delle relazioni semantiche e sintattiche appartenenti ad una lingua.

A seguire, vengono esaminate in dettaglio diverse tecniche di implementazione, tra cui: *Fine-Tuning*, *Prompting* ed *Instruction Prompting*. L'obiettivo principale di queste metodologie è perfezionare le abilità del modello, adattandolo a compiti specifici attraverso una personalizzazione più accurata. In particolare, il *Fine-Tuning* coinvolge l'ottimizzazione dei parametri del modello rispetto ad uno specifico set di dati, consentendo al modello di adattarsi alle peculiarità ed ai requisiti da soddisfare. Il *Prompting* consiste l'introduzione di *input* testuali aventi una struttura specifica, al fine di guidare il modello nella produzione di *output* desiderati. L'*Instruction Prompting* segue un approccio analogo al *Prompting*, ma si concentra sulla fornitura di istruzioni specifiche per guidare il processo di generazione degli *output*. Queste strategie sono fondamentali per ottimizzare le prestazioni dei Large Language Models, pertanto analizzarle consente di avere una comprensione approfondita di come le capacità dei modelli vengono perfezionate ed adattate per rispondere alle esigenze specifiche delle attività a cui sono destinati.

Dopodiché, si procede a fornire una panoramica delle potenzialità dei Large Language Models, approfondendo anche un'analisi delle sfide e delle problematiche ad essi associate. Tale esplorazione permette di comprendere come i LLM possano essere impiegati in diversi contesti e settori, ad esempio, spaziando dalla generazione di testo creativo e la risposta a domande, fino all'elaborazione del linguaggio naturale, la traduzione automatica

e molte altre aree. Allo stesso tempo, però, vengono considerate le difficoltà nell'utilizzo di tali strumenti, come, ad esempio, la possibilità di generare contenuti ingannevoli o discriminatori oppure la necessità di gestire la complessità computazionale o l'adattamento dei modelli a contesti specifici.

Il capitolo si conclude con una sezione dedicata a delineare l'obiettivo centrale che orienta lo sviluppo del presente lavoro di tesi. In particolare, si evidenzia il ruolo cruciale dei Large Language Models e la loro integrazione nel panorama della ricerca. L'obiettivo principale consiste nell'esaminare l'efficacia di diversi modelli di linguaggio per l'estrazione di informazioni relative al contesto di ESG all'interno di *report* aziendali pubblici. L'obiettivo più ampio di questo studio è comprendere come l'utilizzo efficace degli LLM possa consentire, alle aziende interessate, di interpretare ed utilizzare correttamente i dati, migliorando così la propria competitività sul mercato ed ottenendo vantaggi strategici.

2.1 Fondamenti Tecnologici: Large Language Models

Come anticipato nell'introduzione del capitolo 2, la presente ricerca si avvale dell'impiego di modelli di Large Language Models (LLM). Tali modelli appartengono alla classe di algoritmi di Deep Learning che si distinguono per la loro capacità di riconoscere, generare, riassumere, tradurre e persino anticipare contenuti linguistici. La caratteristica distintiva di tali modelli risiede nella necessità di utilizzare considerevoli quantità di dati, determinando dimensioni e complessità notevoli. I LLM, nel corso del tempo, hanno dimostrato un'eccellenza particolare in diverse attività linguistiche, sfruttando una vasta capacità di apprendimento automatico basato su ampi *corpus* di testi, ossia raccolte estese ed organizzate di documenti linguisticamente annotati.

Nonostante la notorietà acquisita dai LLM attraverso il *chatbot*, l'interesse iniziale è emerso con GPT-3. Sviluppato nel 2020 dalla società OpenAI, GPT-3 è un modello di linguaggio capace di produrre risultati complessi. Questo modello ha sorpreso i suoi creatori poiché ha dimostrato competenze superiori alle aspettative iniziali, come, ad esempio, la capacità di generare frammenti di codice ed apprendere dalle esperienze precedenti. Tuttavia, è fondamentale sottolineare che tali abilità non devono essere considerate straordinarie o consapevoli, ovvero non sono indicatori di una reale consapevolezza o comprensione.

Esiste, pertanto, una stretta correlazione tra i modelli di Large Language Models ed una vasta gamma di contenuti, quali testi, immagini, file audio e video, nella quale si rivela una significatività rilevante. A titolo esemplificativo, il modello LLM noto come Dall-e 2 ha dimostrato un grande successo nell'identificazione di nuove proteine [58], mentre ChatGPT trova applicazione nel processo di sviluppo del software. Questi esempi evidenziano che il termine "Large Language Model" rappresenta un'etichetta intrinsecamente limitata rispetto alle reali capacità operative di questi modelli. Di seguito, viene mostrato uno schema relativo al funzionamento di un generico modello LLM (**Figura 2.1**).

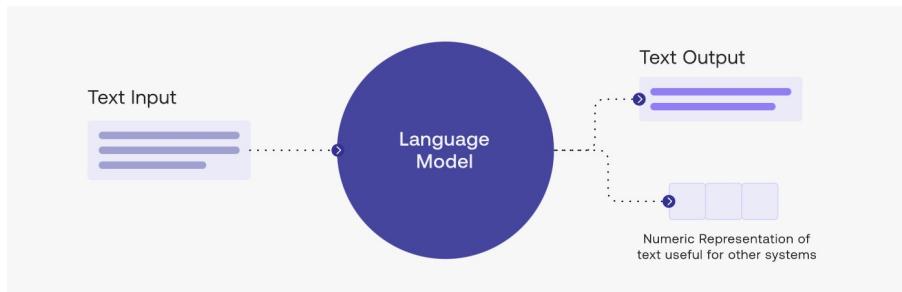


Figura 2.1. Rappresentazione schematica del funzionamento di un LLM. Il testo di *input* viene processato dal modello, il quale produce un testo di *output* ed una rappresentazione numerica del testo. Quest'ultima può essere utilizzata per applicazioni in diversi sistemi che necessitano di dati strutturati per ulteriori elaborazioni.

Fonte: [Network Digital360](#)

Prima di procedere con la descrizione dell’architettura relativa ai Large Language Models, è importante chiarire la distinzione tra questi ultimi ed i modelli di Generative AI. In questo contesto emergono tre aspetti rilevanti [59]:

- **Non tutti gli strumenti di Intelligenza Artificiale generativa sono basati su Large Language Models, ma tutti gli LLM costituiscono una forma di Intelligenza Artificiale generativa.**

L’Intelligenza Artificiale generativa rappresenta una vasta categoria di sistemi in grado di generare contenuti originali. Questa categoria fa uso di strumenti basati su modelli sottostanti, tra cui i Large Language Models. Nei contesti dell’Intelligenza Artificiale generativa, gli LLM svolgono un ruolo chiave nella generazione di testo.

- **Gli LLM producono *output* esclusivamente testuali.**

Inizialmente limitati ad accettare solo *input* testuali, i LLM, come il primo GPT-3, erano focalizzati su un approccio esclusivamente testuale. Tuttavia, con l’avanzamento dei LLM “multimodali”, questi modelli sono ora in grado di gestire *input* audio, immagini e altri formati. Un esempio di LLM multimodale è rappresentato dalla successiva iterazione di OpenAI, ovvero GPT-4. L’Intelligenza Artificiale generativa e gli LLM rivoluzioneranno settori industriali in modi distinti, con possibili impatti nella modellazione 3D, nella generazione di contenuti video, nella creazione di assistenti vocali e altri ambiti legati all’audio.

- **Gli LLM stanno continuamente evolvendo.**

Negli ultimi anni, i LLM hanno acquisito una posizione predominante, in particolare a partire dagli anni 2010, ma la loro notorietà si è ampliata significativamente con l’introduzione di potenti strumenti di Intelligenza Artificiale generativa. Tra questi spiccano ChatGPT di OpenAI e Bard di Google. Secondo l’analisi condotta da Everest Group [60], un’autorevole fonte di approfondimenti strategici nel campo delle tecnologie dell’informazione, dei processi aziendali e dei servizi di ingegneria, nonché un rinomato ente di ricerca nel settore del Business Process Outsourcing (BPO), la pronta espansione osservata nel 2023 può essere attribuita principalmente

all’ampliamento dei parametri all’interno dei LLM. In particolare, il GPT-4 annovera oltre 175 miliardi di parametri, indicando una significativa evoluzione tecnologica.

In sintesi, la distinzione tra Intelligenza Artificiale generativa e LLM è la seguente: l’IA generativa rappresenta una vasta gamma di strumenti progettati per sfruttare le informazioni provenienti dai Large Language Models, o da altri tipi di modelli di Intelligenza Artificiale che utilizzano il Machine Learning, per generare nuovi contenuti. D’altra parte, un LLM è un particolare tipologia di modello di Intelligenza Artificiale che fa uso del Machine Learning basandosi su miliardi di parametri per comprendere e generare testo.

2.1.1 Storia dei Large Language Model

Inizialmente, i *computer* erano progettati per interpretare solo un ristretto insieme di istruzioni, formulate mediante linguaggi di programmazione. Con la diffusione dell’uso dei calcolatori, è emersa la necessità di ampliare la loro capacità di comprensione, consentendo loro di elaborare anche istruzioni redatte in linguaggio naturale. Il campo dell’elaborazione del linguaggio naturale (NLP, Natural Language Processing) si occupa proprio di questo: investigare come i *computer* possano comprendere e generare testi, o altri tipi di media, in linguaggio umano. Le radici del NLP risalgono agli anni ‘50, quando Alan Turing propose il celebre “Test di Turing” [61] come indicatore di Intelligenza Artificiale. Questo esperimento mirava a stabilire se una macchina potesse manifestare comportamenti intelligenti attraverso una conversazione simulata tra un essere umano e due entità, di cui una in grado di emulare un’interazione umana. Il superamento del test si verificava nel caso in cui l’osservatore umano non riuscisse a distinguere la macchina dall’altro partecipante umano.

Nel corso dei primi decenni, la ricerca nel campo NLP si è focalizzata sull’elaborazione manuale di regole al fine di comprendere il linguaggio naturale. Successivamente, negli anni ‘80, con l’aumento della capacità di calcolo e l’introduzione del Machine Learning, sono stati adottati modelli statistici del linguaggio (Statistical Language Model) per migliorare la comprensione del linguaggio naturale. Un esempio emblematico di questa transizione è stato l’utilizzo di tali modelli per la traduzione automatica utilizzando algoritmi di apprendimento supervisionato, rappresentando una delle prime applicazioni di Machine Learning nell’ambito dell’NLP. Questi modelli hanno basato la loro metodologia sull’applicazione di approcci statistici per prevedere e continuare sequenze di parole, rappresentando un progresso significativo nel panorama delle tecnologie linguistiche. L’obiettivo principale di tali strumenti era utilizzare dati statistici per valutare la probabilità di occorrenza delle singole parole e sviluppare una struttura coerente all’interno delle frasi. L’innovazione chiave consisteva nell’utilizzo di informazioni statistiche per potenziare la comprensione e la generazione del linguaggio, aprendo la strada ad ulteriori sviluppi nelle rappresentazioni linguistiche ed introducendo i moderni Large Language Models.

In parole più semplici, i Statistical Language Models rappresentavano un approccio pionieristico che sfruttava la statistica per prevedere il flusso delle parole e costruire frasi in modo più coerente. Nella letteratura di settore, si trova spesso menzione di **Eliza**, sviluppato nel 1966 da Joseph Weizenbaum. Eliza era un semplice programma che utilizzava il riconoscimento di schemi per simulare una conversazione umana, trasformando

l’*input* dell’utente in una domanda e generando una risposta basata su una serie di regole predefinite. Sebbene era lontano dall’essere considerato in grado di leggere e comprendere conversazioni, segnò l’inizio della ricerca sull’elaborazione del linguaggio naturale (NLP) e lo sviluppo di LLM più sofisticati, come ChatGPT o Google Bard.

L’attenzione si è presto spostata dall’utilizzo di algoritmi di apprendimento supervisionato ad algoritmi semi-supervisionati e non supervisionati per analizzare ingenti quantità di dati linguistici generati su Internet. Queste metodologie consentono di elaborare informazioni linguistiche in modo più efficiente ed esaustivo, superando le limitazioni legate alla necessità di etichettare completamente ogni singolo dato. Questo è particolarmente rilevante quando si tratta di affrontare grandi quantità di testo senza dover ricorrere a un’annotazione manuale estensiva, che potrebbe risultare dispendiosa e impraticabile. Inoltre, questi approcci favoriscono una comprensione più approfondita del contesto linguistico, poiché permettono ai modelli di apprendimento automatico di scoprire e sfruttare *pattern* e relazioni presenti nei dati senza una guida dettagliata da parte dell’utente. Questo è cruciale in ambienti *online* in cui il linguaggio può essere altamente dinamico, ricco di sfumature e soggetto a evoluzioni rapide.

Tali approcci costituiscono gli strumenti chiave delle applicazioni del NLP, ovvero la modellazione del linguaggio. I modelli linguistici (LM) sono modelli statistici che rappresentano le probabilità di occorrenza di sequenze di parole in una lingua. In altre parole, un LM è una funzione che associa una probabilità a ciascuna sequenza di parole:

$$P(x^{(t+1)} | x^{(t)} + \dots + x^{(1)})$$

L’espressione calcola la distribuzione di probabilità condizionale dove $x^{(t+1)}$ può essere qualsiasi parola nel vocabolario. Pertanto, i modelli linguistici generano probabilità imparando da un insieme di testi, noto come *corpus*, ovvero una raccolta di testi in una o più lingue, che può essere annotata per fornire informazioni aggiuntive, come il significato delle parole o la loro relazione sintattica.

Uno dei primi approcci alla costruzione di un modello linguistico è basato sugli *n-gram*, ovvero una sequenza di n parole che si susseguono in un dato testo. In questo contesto, il modello assume che la probabilità della parola successiva in una sequenza dipenda solo da una finestra di n parole che la precedono. Tuttavia, i modelli linguistici *n-gram* sono stati ampiamente sostituiti da modelli linguistici neurali, ovvero che si basano su reti neurali, un sistema informatico ispirato alle reti neurali biologiche. Questi modelli utilizzano rappresentazioni continue o incorporamenti di parole per effettuare le previsioni, come mostrato in **Figura 2.2**.

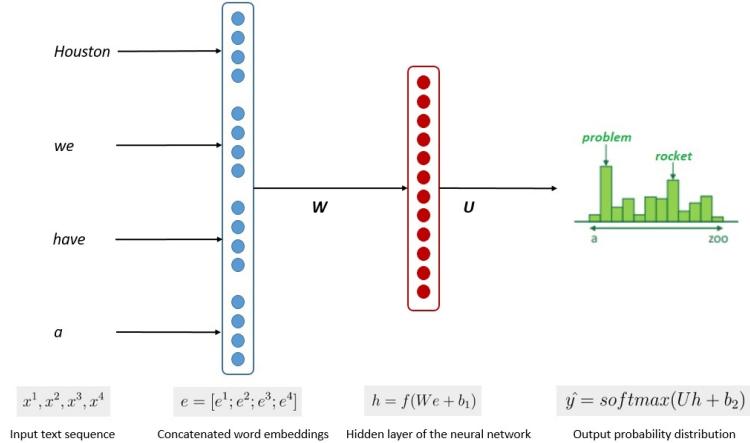


Figura 2.2. Esempio di processo di elaborazione del linguaggio naturale mediante l'utilizzo di una rete neurale. Il flusso inizia con una sequenza di testo in *input* (“Houston we have a problem”), seguita dalla trasformazione delle parole in vettori di *embedding* (simboleggiate dalle sfere blu). I vettori vengono poi concatenati ad una rete neurale, dove passano attraverso uno *layer* nascondito (simboleggia dalle sfere rosse) che esegue trasformazioni lineari e non lineari dei dati (W per i pesi del *layer*, U per i pesi dell’*output*, e b_1 e b_2 per i *bias*). L’*output* della rete è la distribuzione di probabilità delle parole successive, indicata come \hat{y} e visualizzata come un istogramma, con parole candidate come “problem” e “rocket” che ricevono le probabilità più alte.
Fonte: [Baeldung](#)

Fondamentalmente, le reti neurali emergono come un cruciale paradigma di rappresentazione, modellando le parole mediante una combinazione non lineare di pesi. Questo approccio riveste un ruolo chiave per superare la “*curse of dimensionality*”, una problematica intrinseca all’utilizzo di rappresentazioni discrete delle parole. La riduzione della dimensionalità si manifesta quando la complessità di un modello cresce in modo esponenziale rispetto al numero di variabili coinvolte. Ne consegue una dispersione eccessiva dei dati nello spazio delle *features*, rendendo difficile per i modelli linguistici tradizionali effettuare previsioni accurate e generalizzate. L’adozione di reti neurali, mediante la capacità di apprendere in maniera automatica le relazioni complesse tra le parole, costituisce un efficace antidoto a questa problematica, consentendo di superare i limiti imposti dalla *curse of dimensionality* e migliorare notevolmente le prestazioni dei modelli linguistici.

I Large Language Models (LLM) sono, quindi, rappresentazioni neurali del linguaggio implementate su una scala estremamente ampia. Questi modelli possono coinvolgere un numero potenzialmente elevato di parametri, spesso nell’ordine dei miliardi, e sono tipicamente sottoposti ad un processo di addestramento su vasti *corpus* di testo non etichettato, comprendenti anche centinaia di miliardi di parole. I LLM, noti anche come modelli di Deep learning, assumono tipicamente la forma di modelli generici eccellenti in una vasta gamma di compiti. Solitamente, vengono formati su attività relativamente elementari, come la previsione della parola successiva all’interno di una frase. Nonostante questo, tali modelli sono in grado di acquisire una notevole comprensione della sintassi e

della semantica del linguaggio umano. Di conseguenza, dimostrano competenze più precise su una vasta gamma di compiti nell’ambito della linguistica computazionale. Questo rappresenta un notevole cambio di paradigma rispetto all’approccio precedente nelle applicazioni di NLP, in cui modelli linguistici specializzati venivano addestrati per eseguire compiti specifici. Infatti, i ricercatori hanno rilevato l’emergere di molteplici competenze nei LLM, competenze per le quali essi non erano stati precedentemente addestrati. A titolo esemplificativo, è stato dimostrato che gli LLM sono in grado di eseguire operazioni aritmetiche complesse in più passaggi, decodificare le lettere di una parola ed individuare contenuti offensivi nelle lingue parlate.

È importante sottolineare come la storia dell’evoluzione dei LLM è stata, in realtà, intrisa di frammentazione, caratterizzata da fasi di disinteresse notoriamente designate come gli “inverni dell’Intelligenza Artificiale”. Tali periodi, contrassegnati da una marcata diminuzione dell’attenzione e degli investimenti nella ricerca nel campo dell’IA, hanno esercitato una significativa influenza sugli LLM durante il secondo e il terzo inverno dell’IA, verificatisi rispettivamente agli inizi degli anni ‘80 e poi proseguiti negli anni ‘90. Durante questi periodi, gli sforzi di sviluppo e miglioramento dei modelli linguistici subirono un rallentamento significativo a causa della limitata focalizzazione e dell’incertezza riguardo alla direzione futura della ricerca per l’Intelligenza Artificiale. In particolare, nel primo inverno, le risorse e gli investimenti, in precedenza vigorosamente destinati ad esplorare le potenzialità dei Large Language Models, subirono una contrazione notevole, con conseguente impatto sulla crescita e sull’avanzamento della disciplina. Il secondo inverno dell’IA, manifestatosi nei primi anni ‘80, fu caratterizzato da un diffuso scetticismo riguardo alle aspettative e alle possibilità realistiche della ricerca nell’Intelligenza Artificiale. Questa fase di arretramento comportò una riduzione delle risorse finanziarie e umane assegnate alla progettazione e allo sviluppo dei LM, con il risultato di una temporanea stagnazione nell’avanzamento di queste tecnologie. Il terzo inverno dell’IA, emergente negli anni ‘90, presentò sfide simili ma con una dimensione diversa. In questo contesto, la percezione pubblica ed industriale sull’IA oscillò tra l’entusiasmo iniziale e la delusione, influenzando notevolmente gli investimenti nel settore. La ricerca per i LM subì ulteriori limitazioni, con molte organizzazioni che ridussero o cessarono completamente il loro impegno in questo campo, riflettendo un clima generale di incertezza e cauzione.

In sintesi, gli “inverni dell’IA” hanno rappresentato fasi critiche nella storia dei modelli linguistici, segnando periodi di rallentamento ed incertezza che hanno influenzato profondamente il corso dello sviluppo di queste tecnologie. La comprensione di tali contesti storici è essenziale per valutare appieno la resilienza e la crescita dei LM nel contesto dell’evoluzione più ampia dell’Intelligenza Artificiale.

Nel 2013, è stato sviluppato da Tomas Mikolov ed il suo team presso Google, **word2vec** [62], un algoritmo di *embedding* capace di convertire le parole in vettori fornendo così un’innovativa prospettiva nella comprensione delle relazioni semantiche tra termini. L’aspetto rivoluzionario di **word2vec** risiede nella sua capacità di convertire le parole in vettori numerici densi, in cui parole simili dal punto di vista semantico sono rappresentate come vettori vicini nello spazio. Questo approccio, noto come *embedding word*, ha superato le limitazioni dei metodi tradizionali basati su rappresentazioni sparsi ed ha

permesso di catturare relazioni semantiche complesse tra le parole. L'algoritmo opera su principi di apprendimento non supervisionato e può essere addestrato su grandi quantità di testo per catturare le relazioni contestuali tra le parole.

Successivamente si arriva al 2020, anno cruciale per l'IA con il rilascio di GPT-3 (Generative Pre-trained Transformer 3). Questo modello linguistico ha guadagnato notorietà per la sua straordinaria capacità di generare testi coerenti e significativi, dimostrando un considerevole passo in avanti rispetto alle sue precedenti iterazioni. GPT-3 ha rivitalizzato l'interesse nell'ambito dell'IA, catalizzando l'attenzione del pubblico e degli studiosi. Tuttavia, sarebbe un'analisi troppo superficiale circoscrivere lo sviluppo dei modelli linguistici solamente all'avvento di GPT-3. Quest'ultimo, pur rappresentando un apice nella capacità di comprensione del linguaggio, costituisce solo un capitolo in una storia più ampia, ovvero risulta cruciale considerare GPT-3 come un punto di riferimento significativo, ma non esaustivo, all'interno di un panorama più ampio di modelli e metodologie. Il campo dell'IA continua ad evolversi e nuove scoperte e sviluppi offrono costantemente nuove prospettive e sfide.

2.1.2 Architettura generale dei Large Language Models

All'inizio del loro sviluppo, i Large Language Model sono stati principalmente concepiti mediante l'impiego di algoritmi basati sull'apprendimento auto-supervisionato (*self supervised*). Tale tipologia di apprendimento condivide con il metodo supervisionato la caratteristica di utilizzare dati di addestramento dotati di etichette. Tuttavia, è importante sottolineare che in questo caso l'etichettatura avviene in maniera automatica, eliminando la necessità di un intervento umano. Per questo motivo è possibile considerare tale metodologia una forma di apprendimento non supervisionato. Nell'ambito dell'apprendimento auto-supervisionato, il *software* è in grado di dedurre autonomamente le etichette adeguate, sfruttando le correlazioni esistenti tra i dati e le strutture implicitamente presenti in essi.

In molte circostanze, gli algoritmi di apprendimento auto-supervisionato fanno ampio ricorso a modelli basati su reti neurali artificiali (Artificial Neural Networks, ANN). Le ANN costituiscono una classe di modelli computazionali ispirati al funzionamento del cervello umano e sono utilizzati per apprendere *pattern* complessi dai dati. Nella formulazione di tali reti, è possibile configurare diverse architetture ma, per molte applicazioni di apprendimento auto-supervisionato, la scelta preponderante è stata l'adozione della rete neurale ricorrente (Recurrent Neural Network, RNN). Le RNN sono una particolare tipologia di ANN progettate per lavorare con dati sequenziali, come sequenze temporali o testi. La caratteristica distintiva delle RNN è la presenza di cicli all'interno della struttura, consentendo loro di mantenere un *layer* interno che possa essere aggiornato con l'arrivo di nuovi *input*. Questa capacità rende le RNN particolarmente adatte per catturare dipendenze a lungo termine nei dati sequenziali e, pertanto, efficaci in contesti in cui l'ordine delle informazioni è significativo. La **Figura 2.3** mostra un diagramma schematico di una tipica architettura di RNN.

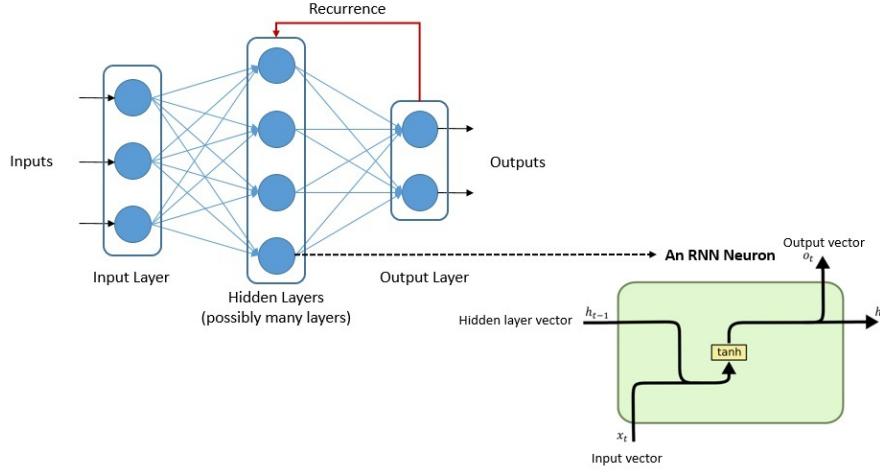


Figura 2.3. Architettura di una RNN. Viene illustrato il flusso di *input* attraverso i diversi strati (*input layer*, *hidden layer* e *output layer*) e la ricorrenza che collega l'*output* di un *hidden layer* al suo *input* al tempo successivo. Inoltre, viene mostrato in dettaglio la struttura interna di un neurone RNN, mostrando la trasformazione dell'*input* vettoriale attraverso la funzione di attivazione **tanh** (tangente iperbolica, essa trasforma un *input* numerico in un *output* compreso nell'intervallo $[-1, 1]$) e la produzione di un vettore di strato nascosto ed un vettore di *output*.

Fonte: [Baeldung](#)

Un modello RNN elabora ciascuna parola o carattere individualmente, producendo un *output* una volta completato il processo di elaborazione dell'intero testo di *input*. Sebbene il funzionamento sia generalmente efficace, talvolta si verifica un fenomeno in cui il modello sembra “dimenticare” gli eventi iniziali della sequenza quando si avvicina alla sua conclusione del processo. Per tale motivo sono state sviluppate delle varianti di RNN, tra cui le Long Short-Term Memory (LSTM) [63] e le Gated Recurrent Units (GRU) [64]. L'architettura LSTM offre un supporto significativo alla RNN per gestire a lungo termine le dipendenze temporali, consentendo di mantenere e dimenticare informazioni nel corso dell'apprendimento. La struttura di GRU, invece, pur essendo meno complessa rispetto a LSTM, presenta notevoli vantaggi. In particolare, GRU prevede due porte principali: una porta di reset (*reset gate*) ed una porta di aggiornamento (*update gate*). Queste porte consentono alla GRU di regolare l'accesso e l'aggiornamento della memoria interna, facilitando il flusso delle informazioni attraverso la rete. In tal modo, vengono richieste minori risorse di memoria durante la fase di addestramento e si dimostra una maggiore efficienza computazionale rispetto a LSTM. Tuttavia, è importante sottolineare che l'efficacia di GRU è generalmente più evidente in contesti caratterizzati da set di dati di dimensioni ridotte. In altre parole, GRU si configura come una scelta preferibile quando la complessità computazionale e l'occupazione di memoria devono essere ottimizzate, soprattutto in situazioni in cui le dimensioni dei dati sono limitate.

Per un lungo periodo, le LSTM e le GRU hanno rappresentato la scelta predominante per lo sviluppo di sistemi di elaborazione del linguaggio naturale avanzati. Tuttavia,

questi modelli presentano una problematica nota come il problema del gradiente evanescente (*vanishing gradient*), caratterizzato dalla rapida diminuzione dei gradienti mentre vengono propagati attraverso i diversi *layer* della rete. Questa riduzione drastica complica notevolmente l'aggiornamento dei pesi, i quali rappresentano i parametri della rete neurale che vengono adattati durante il processo di addestramento per minimizzare l'errore complessivo del modello. In particolare, nei primi strati della rete, il gradiente può diventare estremamente piccolo, impedendo efficacemente ai pesi di subire modifiche significative nel loro valore. Questo fenomeno, noto come gradiente evanescente, si manifesta durante l'addestramento delle reti neurali utilizzando metodi di apprendimento basati sul gradiente con *backpropagation*. In alcuni casi, il problema del gradiente evanescente può addirittura bloccare ulteriormente l'addestramento della rete neurale, compromettendo le capacità di apprendimento del sistema neurale nel suo complesso. Queste complicazioni rendono sostanzialmente inefficiente l'utilizzo delle reti neurali ricorrenti per compiti di elaborazione del linguaggio naturale (NLP).

Alcuni dei problemi associati alle RNN sono stati mitigati mediante l'introduzione del meccanismo di attenzione nella loro struttura. Nelle architetture ricorrenti, come le LSTM, si è constatato che la capacità di propagare informazioni è limitata e la finestra temporale per la conservazione delle informazioni è piuttosto breve. Tuttavia, l'implementazione del meccanismo di attenzione ha consentito un significativo ampliamento di questa finestra informativa. L'attenzione rappresenta una strategia volta a migliorare specifiche porzioni dei dati in *input*, enfatizzandone alcune parti e riducendone altre. Tale approccio è motivato dal principio che la rete dovrebbe focalizzare la sua attenzione sulle parti più rilevanti dei dati a disposizione.

Nel contesto della ricerca sui modelli di elaborazione del linguaggio naturale, è importante comprendere la sottile distinzione tra il concetto di “attenzione” e quello di “auto-attenzione”, sebbene la loro motivazione fondamentale rimanga invariata. Mentre il meccanismo di attenzione implica la capacità di concentrarsi su diverse parti di una sequenza esterna, l’auto-attenzione si concentra sulla capacità di rivolgere l’attenzione a diverse porzioni della sequenza corrente. L’auto-attenzione, come illustrato nella **Figura 2.4**, offre al modello la possibilità di accedere alle informazioni provenienti da qualsiasi elemento all’interno della sequenza di *input*. Questo è particolarmente rilevante nell’ambito dell’elaborazione del linguaggio naturale, poiché consente al modello di acquisire informazioni significative anche da *token* (unità con cui viene suddiviso un testo) distanti all’interno di una sequenza di testo. Grazie a questo meccanismo, il modello può catturare dipendenze nell’intera sequenza senza dover fare affidamento su finestre fisse o meccanismi di scorrimento.

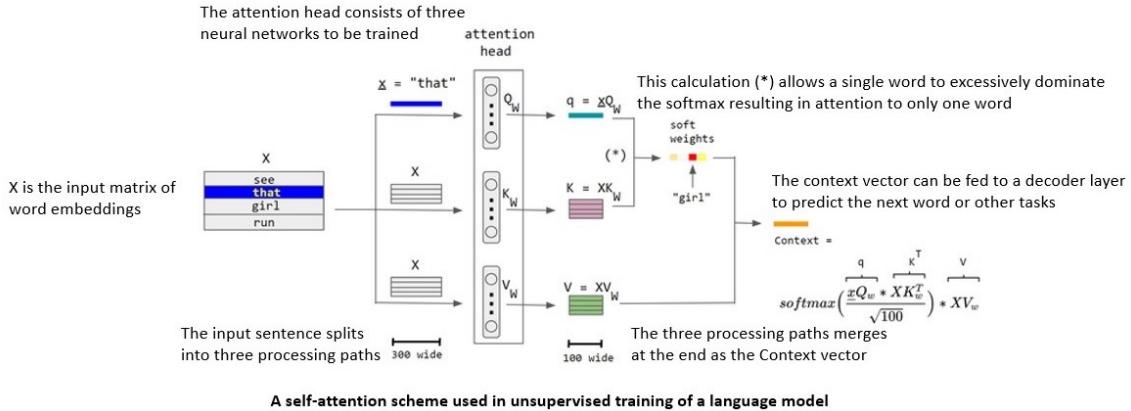


Figura 2.4. Schema di auto-attenzione utilizzato nell’addestramento non supervisionato di un modello linguistico. L’auto-attenzione consente alla rete di apprendere quali parole sono più importanti per la comprensione del contesto. In questo schema, viene utilizzata per calcolare un vettore contesto, che rappresenta la relazione tra tutte le parole nell’*input*. Il vettore contesto può essere utilizzato per prevedere la prossima parola nell’*input* o per altre attività.

Fonte: [Baeldung](#)

Nel contesto delle applicazioni di NLP, la rappresentazione delle parole o dei simboli presenti nel linguaggio naturale gioca un ruolo cruciale. Nei modelli di LLM, il testo in *input* viene suddiviso in *token* e ciascun *token* viene trasformato mediante un processo di incorporamento delle parole (*word embedding*) all’interno di un vettore con valori reali. Questo incorporamento delle parole è in grado di catturare il significato intrinseco di ciascuna parola, in modo tale che parole vicine nello spazio vettoriale rappresentino concetti simili. Gli *embedding* di parole si presentano con diverse forme, una delle quali consiste nell’esprimere le parole come vettori di contesti linguistici in cui la parola compare. Per generare questi incorporamenti di parole vengono adottati diverse metodologie ma l’approccio basato sull’architettura di reti neurali che risulta essere il più diffuso. Nel 2013, il *team* di ricerca di Google ha introdotto *word2vec*, come precedentemente anticipato nella sezione 2.1.1, un *toolkit* per l’*embedding* di parole che utilizza una rete neurale per apprendere le associazioni tra parole da un vasto *corpus* di testi. Studi successivi hanno dimostrato che l’utilizzo di incorporamenti di parole e di frasi migliora le *performance* nelle attività di NLP, come l’analisi sintattica e l’analisi del *sentiment* (in cui vengono estratte e valutate le emozioni, le opinioni ed il tono espressi all’interno di un testo).

I modelli basati su reti neurali ricorrenti (RNN), implementati con meccanismi di attenzione, hanno mostrato un notevole miglioramento nelle loro prestazioni. Tuttavia, è importante notare che i modelli ricorrenti presentano intrinsecamente difficoltà di scalabilità. D’altra parte, il meccanismo di auto-attenzione si è dimostrato straordinariamente efficace, superando addirittura la necessità di elaborazione sequenziale ricorrente. Di conseguenza, l’introduzione dei Transformers da parte del *team* di Google Brain nel 2017 (come illustrato nella **Figura 2.5**) segna un momento di svolta cruciale nella storia dei modelli di linguaggio basati su reti neurali. Questo è dovuto al fatto che i Transformers sono modelli di apprendimento profondo che sfruttano il meccanismo di auto-

attenzione, rivoluzionando l'approccio al trattamento dell'intero *input* simultaneamente e rappresentando un significativo avanzamento nel campo.

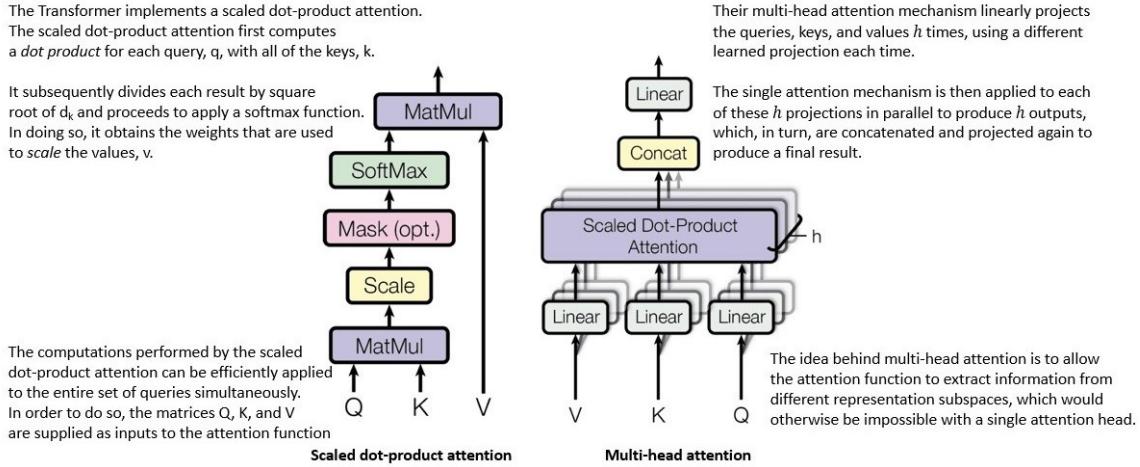


Figura 2.5. Struttura e funzionamento dell'attenzione a prodotto scalare e dell'attenzione multi-testa, utilizzata nei Transformers. Inizialmente, le *query*, le chiavi ed i valori (Q , K e V) vengono proiettati h volte, con una proiezione lineare diversa per ciascuna iterazione. Successivamente, per ogni *query* (q), viene calcolato il prodotto scalare con tutte le chiavi (k) ed il risultato viene diviso per la radice quadrata della dimensione d , seguita dall'applicazione di una funzione **softmax**. Questo passaggio determina i pesi utilizzati per scalare i valori (v), che vengono poi sommati per ottenere l'*output* dell'attenzione. Infine, l'attenzione “scaled dot-product” viene applicata h volte in parallelo, utilizzando diverse proiezioni lineari per ogni testa. Gli *output* delle h teste vengono concatenati e proiettati nuovamente per ottenere l'*output* finale. Questo processo è essenziale per l'efficacia del modello Transformer nell'elaborazione di sequenze di *input* in modo simultaneo ed efficiente.

Fonte: [Baeldung](#)

In netta deviazione dai modelli precedenti basati su reti neurali ricorrenti (RNN), i Transformer si distinguono per l'assenza di una struttura ricorrente. Attraverso un opportuno insieme di dati di addestramento, il solo meccanismo di attenzione integrato nell'architettura dei trasformatori può raggiungere prestazioni paragonabili a quelle di un modello RNN equipaggiato con un analogo meccanismo di attenzione. Un ulteriore vantaggio, di notevole rilievo, nell'utilizzo dei trasformatori è rappresentato dalla loro maggiore parallelizzazione e dalla riduzione dei tempi di addestramento. Questa peculiarità si configura come aspetto strategico di primaria importanza, risultando fondamentale per la costruzione di modelli di linguaggio su un vasto *corpus* di dati testuali, tenendo conto delle limitazioni temporali e delle risorse disponibili.

I modelli Transformer sono stati integrati nel contesto dei modelli Sequence-to-Sequence [65] (noti anche come Seq2Seq), i quali sono progettati per ricevere una sequenza di *input* e produrre una sequenza di *output*. Questi modelli sono particolarmente efficaci nelle applicazioni di traduzione. Un'opzione comune per tali modelli sono le reti neurali LSTM

che, grazie alla loro natura sequenziale, sono in grado di dare significato alla sequenza “memorizzando” le parti rilevanti, siccome l’ordine delle parole all’interno delle frasi è cruciale per la comprensione del significato. I modelli Seq2Seq sono generalmente composti da un *encoder* ed un *decoder*. L’*encoder* elabora la sequenza di *input* e la proietta in uno spazio dimensionale più ampio (vettore n-dimensionale). Questo vettore astratto viene quindi utilizzato dal *decoder* per generare la sequenza di *output* desiderata. La sequenza di *output* potrebbe essere in un’altra lingua, rappresentata da simboli o una copia dell’*input* stesso. Sia l’*encoder* che il *decoder* condividono lo stesso spazio n-dimensionale, il che significa che entrambi sono in grado di rappresentare l’informazione in modo astratto. Questo consente all’*encoder* di trasformare l’*input* nello spazio dimensionale, mentre al *decoder* di leggere e mappare i valori in una sequenza di *output*. Inizialmente, né l’*encoder* né il *decoder* hanno conoscenza dello spazio n-dimensionale. Per apprenderlo, vengono addestrati.

Nella **Figura 2.6**, viene fornita una panoramica dettagliata dell’architettura del modello Transformer, che comprende diversi componenti chiave, tra cui il Blocco di Input Embedding, l’Encoding Posizionale, l’Encoder, l’Attenzione Multi-Head, il Decoder, l’Attenzione Multi-Head Mascherata e lo Strato Softmax di *output*. Per addestrare un Transformer, è necessario fornire testo in *input* e testo in *output*. Nel Blocco di Encoding, situato a sinistra nell’architettura, vengono passate le *feature*, ovvero il testo in *input*, mentre nel Decoder, a destra, vengono inserite le sequenze di *output*, ovvero il testo che si prevede che il Transformer genererà.

Il processo inizia col Blocco di Input Embedding. Gli *embedding*, o incorporamenti, sono rappresentazioni vettoriali complesse che sono fondamentali nel campo del Deep Learning e vengono utilizzate anche all’interno del modello Transformer. Gli *embedding* sono necessari per convertire le stringhe testuali in numeri, vettori o matrici, rendendo possibile il loro trattamento da parte dei modelli. Nel Transformer, il meccanismo di *embedding* arricchisce le rappresentazioni testuali con un *encoding* posizionale, che fornisce al modello informazioni sulla posizione dei *token* all’interno dell’*input*. Un aspetto rivoluzionario del Transformer è il meccanismo di attenzione, che permette al modello di selezionare le parti rilevanti di una frase in base al significato ed al contesto. Questo meccanismo è evidenziato come il primo componente del Blocco di Encoding.

Il Blocco di Attenzione ha lo scopo di individuare le parti importanti di una frase, basandosi sulle rappresentazioni vettoriali create dai *layer* precedenti. Ogni parola della frase riceve un vettore di attenzione in *output*, che viene quindi elaborato parallelamente nel Blocco Feed-forward, costituito da una rete neurale classica. Questo *layer* trasforma l’*output* dell’attenzione in una matrice comprensibile per l’unità di decodifica.

Nell’unità di decodifica, il Blocco di Embedding e Codifica Posizionale funziona in modo simile a quello del Blocco di Encoding. Tuttavia, la differenza fondamentale si manifesta con l’introduzione dell’unità di attenzione mascherata. Questa unità maschera le parole successive alla parola di *output* durante il processo di addestramento, garantendo che il modello apprenda a generare correttamente le parole successive nella sequenza. Successivamente, l’unità di attenzione *multi-head*, nota anche come attenzione Encoder-Decoder, riceve l’*input* sia dal Decoder che dall’Encoder e valuta le relazioni tra la sequenza di *input* e quella di *output*. Questo blocco è cruciale poiché consente al Transformer di

mappare l'*input* a *output*, comprendendo le relazioni tra i *token* e quindi le sequenze. Infine, l'*output* dell'attenzione *multi-head* viene elaborato da un'altra rete neurale *feed-forward* prima di essere trasformato linearmente dal *layer* Softmax. Quest'ultimo, utilizza la funzione *softmax* per restituire la distribuzione di probabilità dei *token* rispetto a tutti quelli presenti nel set di addestramento. Questa distribuzione di probabilità consente al modello di selezionare il *token* più probabile da restituire in base al contesto. In definitiva, questa architettura complessa e ben definita del Transformer sfrutta efficacemente meccanismi chiave come l'attenzione e la funzione *softmax* per generare *output* di alta qualità nel contesto del trattamento del linguaggio naturale.

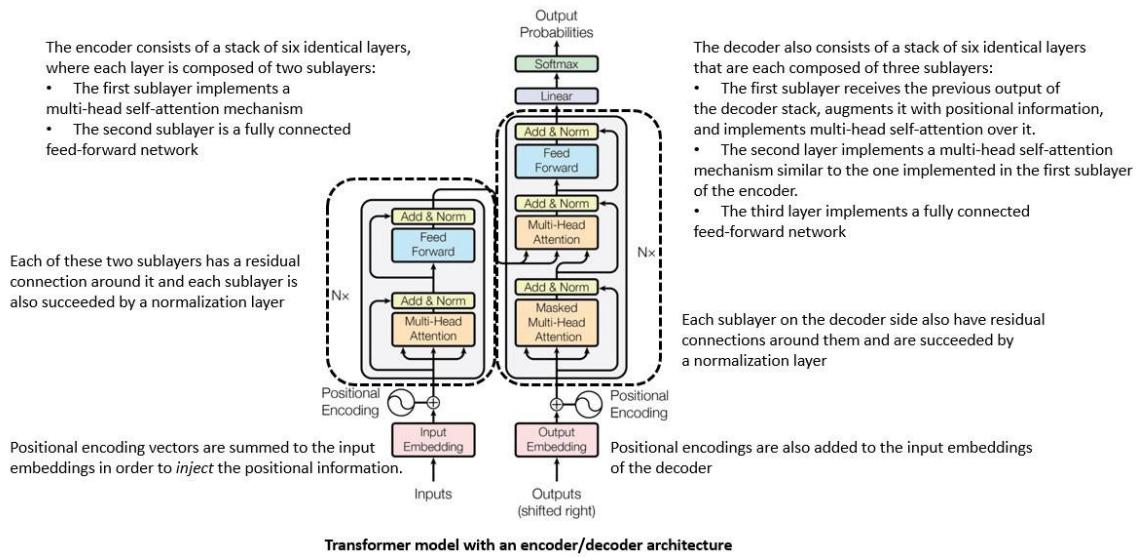


Figura 2.6. Architettura di un Transformer, focalizzata sulle componenti del codificatore e decodificatore. Il codificatore è composto da sei *layer* identici, ogni strato è caratterizzato da due sottostrati principali: un meccanismo di *self-attention* multi-testa ed una rete *feed-forward* completamente connessa. Al contrario, il decodificatore segue una struttura simile ma con alcune variazioni: oltre al meccanismo di *self-attention* multi-testa ed alla rete *feed-forward*, è presente il meccanismo di *self-attention* multi-testa con mascheramento, che permette al modello di focalizzarsi solo sulle informazioni precedenti nella sequenza durante la fase di generazione. Importante sottolineare che ogni sottostrato è dotato di una connessione residua per facilitare il flusso delle informazioni e garantire la stabilità del modello, ed ogni *layer* è seguito da un livello di normalizzazione per mantenere la distribuzione delle attivazioni stabile nel corso dell'addestramento. Questa struttura è stata progettata per massimizzare le prestazioni del modello Transformer nelle applicazioni di elaborazione del linguaggio naturale e di altri tipi di dati sequenziali.

Fonte: [Baeldung](#)

2.1.3 Tecniche di applicazione degli LLM

Come discusso nella sezione 2.1.2, i LLM vengono comunemente sottoposti ad un processo di addestramento su un vasto *corpus* di dati caratterizzato da attività semplici e generiche. Tuttavia, per applicare un LLM ad una specifica attività di elaborazione del linguaggio naturale (NLP), è essenziale impiegare tecniche specializzate come il *fine-tuning*, il *prompting* o l'*instructor tuning*.

Fine-Tuning

Il *fine-tuning* rappresenta un approccio della tecnica *transfer learning*. Il *transfer learning* è una tecnica di Machine Learning in cui la conoscenza del modello, appresa da un'attività, viene riutilizzata per migliorare le prestazioni in un'attività correlata. Il processo di *fine-tuning* viene tipicamente associato alle reti neurali e può essere applicato sia all'intera rete che ad un sottoinsieme specifico di suoi *layer*. Questo approccio comporta l'adattamento di un modello pre-addestrato ad un nuovo compito o ad un insieme di dati specifico. Nel caso delle reti neurali linguistiche, per esempio, il *fine-tuning* implica l'introduzione di un nuovo set di pesi collegato all'ultimo *layer* del modello, orientato verso l'*output* del compito successivo. Inoltre, è comune mantenere “congelati” i pesi originali del modello, garantendo che i parametri non vengano aggiornati durante il *training*, come illustrato nella **Figura 2.7**. Questa pratica, tuttavia, non è limitata esclusivamente alle reti neurali e può essere estesa ad altri tipi di modelli di apprendimento automatico, in base alle specifiche esigenze del problema ed alla natura dei dati disponibili.

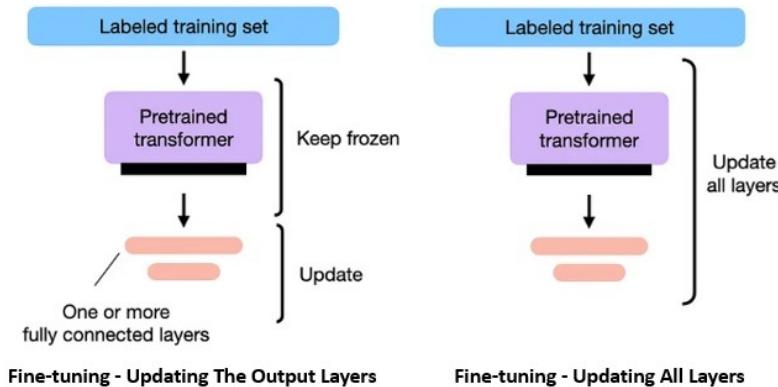


Figura 2.7. L'immagine mostra due differenti schemi per il *fine-tuning* nei modelli di apprendimento automatico. Sul lato sinistro descrive il processo di aggiornamento degli *output layers* in cui il trasformatore rimane “congelato” e solo uno o più *layers* completamente connessi vengono aggiornati con un nuovo set di dati etichettato. Sul lato destro illustra l'aggiornamento di tutti i *layers* del trasformatore con il set di dati etichettato. Quest'ultimo approccio è più complesso e computazionalmente oneroso ma può portare a prestazioni migliori se il set di dati di *fine-tuning* è sufficientemente grande e rappresentativo per lo specifico compito.

Fonte: [Baeldung](#)

Il *fine-tuning* viene generalmente realizzato mediante l'apprendimento supervisionato, il quale coinvolge un insieme notevolmente più limitato di dati etichettati specifici e pertinenti al compito, come mostrato nella **Figura 2.8**.

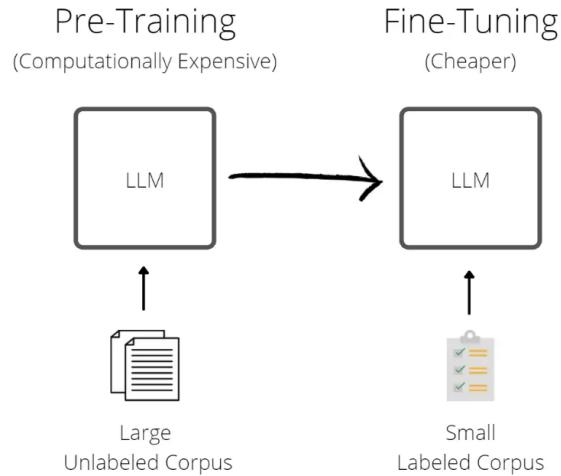


Figura 2.8. Processo di apprendimento degli LLM. Il pre-addestramento del modello avviene prima su un ampio *corpus* non etichettato, richiedendo un'alta capacità computazionale. Successivamente, il *fine-tuning* è eseguito su un *corpus* più piccolo ed etichettato, risultando meno dispendioso.

Fonte: [Network Digital360](#)

Altri approcci spesso impiegati per il *fine-tuning* includono l'apprendimento con supervisione debole e il *reinforcement learning*. L'apprendimento con supervisione debole si distingue per la sua capacità di sfruttare informazioni parziali, incerte o affette da rumore durante la fase di addestramento. In questo contesto, il modello non si basa esclusivamente su etichette chiaramente definite, ma può incorporare indicazioni vaghe o informazioni probabilistiche, consentendo una maggiore flessibilità nell'adattamento ai dati. Per quanto concerne il *reinforcement learning*, si tratta di una metodologia di Machine Learning che coinvolge un sistema informatico denominato “agente”. Quest’ultimo apprende a svolgere un determinato compito mediante iterazioni ripetute, basate sul principio di *trial-and-error* (ovvero “tentativi ed errori”), interagendo con un ambiente in costante evoluzione. Questa tipologia di apprendimento consente all’agente di prendere una serie di decisioni volte a massimizzare una specifica metrica di ricompensa per l’attività, senza richiedere una programmazione esplicita per tale operazione e senza coinvolgimento umano diretto. L’incorporazione di tali metodologie nel processo di *fine-tuning* consente di affrontare diverse sfide, tra cui la scarsa disponibilità di dati etichettati di elevata qualità e la necessità di adattare il modello a contesti dinamici, contribuendo significativamente al miglioramento delle *performance* complessive.

Il *fine-tuning* rappresenta la strategia più comunemente utilizzata per adattare i Large Language Models all’esecuzione di compiti successivi nel campo dell’elaborazione del linguaggio naturale. Questi compiti includono *sentiment analysis* (che permette di identificare e classificare le emozioni, o opinioni, espresse in un testo, determinando se sono

positive, negative o neutre), *named-entity recognition* (che consente di individuare e classificare entità nominate, come persone, luoghi, organizzazioni, date o altro) e *part-of-speech tagging* (che assegna etichette grammaticali a ciascuna parola in una frase al fine di identificarne la funzione grammaticale). Tuttavia, con l'aumentare delle dimensioni dei LLM, sono emerse tecniche più semplici, come ad esempio il *prompting*, che hanno cominciato a guadagnare maggiore popolarità.

Prompting

Con l'introduzione dei Large Language Models, come GPT-3, il *prompting* ha assunto un ruolo più diretto e popolare nell'utilizzo di tali modelli al fine di effettuare compiti specifici. In questo metodo, il problema da risolvere viene presentato al modello attraverso un *prompt* testuale, che il modello deve affrontare fornendo un completamento. Un *prompt* costituisce quindi un *input* testuale fornito ad un LLM, concepito per indurre il modello ad una risposta specifica. Il processo di formulazione di tale *input* è noto come *prompt engineering*, rappresentando una combinazione di creatività e metodologia nel delineare le istruzioni in maniera precisa con lo scopo di ottenere risposte desiderate dal sistema.

Un *prompt* può assumere diverse forme e strategie al fine di guidare un modello linguistico nella generazione di risposte coerenti. Una modalità comune è il Zero-shot Prompting, in cui una singola richiesta viene presentata al modello senza ulteriori suggerimenti. In alternativa, si può adottare il Few-shot Prompting, nel quale si fornisce al modello un *prompt* contenente esempi di domande correlate alle relative risposte, seguiti dalla effettiva domanda. Questo approccio, noto anche come “*in-context learning*”, consiste nell'inserire specifici esempi di coppie problema-soluzione, denominati “*shots*”, allo scopo di istruire il modello attraverso l'apprendimento da esempi concreti. Una variante interessante del *in-context learning* è il Chain of Thought Prompting. In questo caso, il *prompt* include esempi *few-shot* ma introduce una serie di passaggi intermedi di ragionamento. Ciò incoraggia il modello a ragionare seguendo la struttura del *prompt*, eseguendo una catena di pensiero. Questo approccio si rivela particolarmente vantaggioso quando il modello è chiamato ad affrontare compiti che richiedono un pensiero logico e ragionamento, come evidenziato nella **Figura 2.9**. Tale metodologia impone al modello di elaborare attentamente tutti i passaggi del problema, scoraggiando risposte superficiali e promuovendo una riflessione più approfondita.

Poiché i Large Language Models racchiudono notevoli quantità di conoscenza, sono capaci di generare risposte coerenti anche mediante la tecnica *Zero-shot Prompting*. Tuttavia, la loro *performance* è fortemente influenzata dalla formulazione precisa del *prompt*. Questa sensibilità ha dato vita al campo del *prompt engineering*, emerso come disciplina critica per ottimizzare le prestazioni degli LLM, richiedendo una progettazione attenta dei *prompt* al fine di guidare il modello verso risposte accurate ed informative.

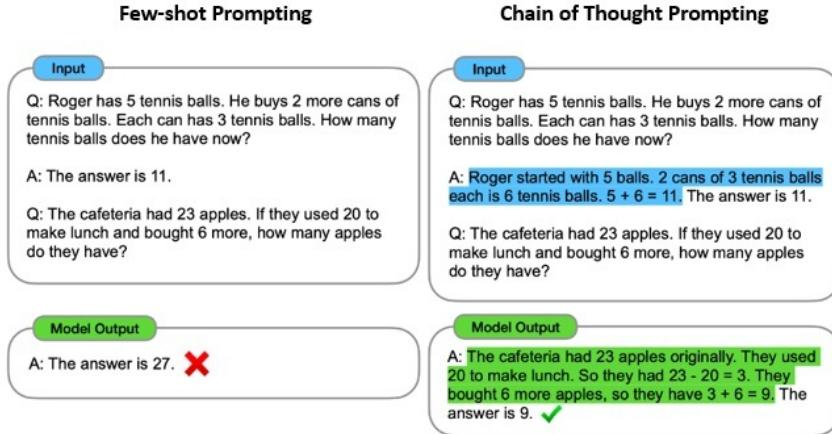


Figura 2.9. Confronto delle metodologie Few-shot Prompting e Chain of Thought Prompting per la risoluzione di problemi matematici tramite due esempi di domande e le relative risposte fornite da un modello di AI. Nella parte a sinistra dell'immagine, il Few-shot Prompting presenta una risposta errata al secondo quesito, mentre nella parte a destra, l'approccio Chain of Thought Prompting conduce ad una risposta corretta, dimostrando come la spiegazione dettagliata del processo di pensiero possa migliorare la precisione del modello.

Fonte: [Baeldung](#)

Instructor Tuning

Un'ulteriore strategia di implementazione per i Large Language Models è l'*instruction tuning*. Questa tecnica rappresenta una forma di *fine-tuning* progettata per migliorare l'efficacia della tecnica di Zero-shot Prompting introducendo istruzioni o suggerimenti specifici con l'obiettivo di rendere le interazioni con il modello più naturali ed accurate. L'*instruction tuning* prevede la personalizzazione dei parametri del modello per ottimizzare le proprie *performance* su compiti specifici, indicati attraverso istruzioni dettagliate. Durante il processo di addestramento, il modello è esposto a numerosi esempi di compiti formulati come istruzioni in linguaggio naturale, accompagnate da risposte appropriate. Ciò consente al modello di apprendere come rispondere in modo coerente e preciso alle richieste formulate secondo le istruzioni fornite, amplificando così la sua capacità di eseguire con successo compiti complessi attraverso lo Zero-shot Prompting.

Esistono diverse tecniche di metodologie per implementare la tecnica di *instructor tuning*, una tra le più utilizzate è, ad esempio, il Reinforcement Learning From Human Feedback (RLHF), applicato con successo in modelli di rilievo come ChatGPT [66] e Sparrow, sviluppato da Google. Il processo si articola in diverse fasi, come mostrato in **Figura 2.10**. Il RLHF è una tecnica che utilizza il *feedback* umano per ottimizzare direttamente un modello di linguaggio, allineando l'addestramento del modello con valori umani complessi. A differenza dei modelli tradizionali di apprendimento automatico, che si basano su semplici funzioni di perdita, RLHF mira a superare queste limitazioni, offrendo un approccio più versatile e accurato nella rappresentazione della conoscenza umana.

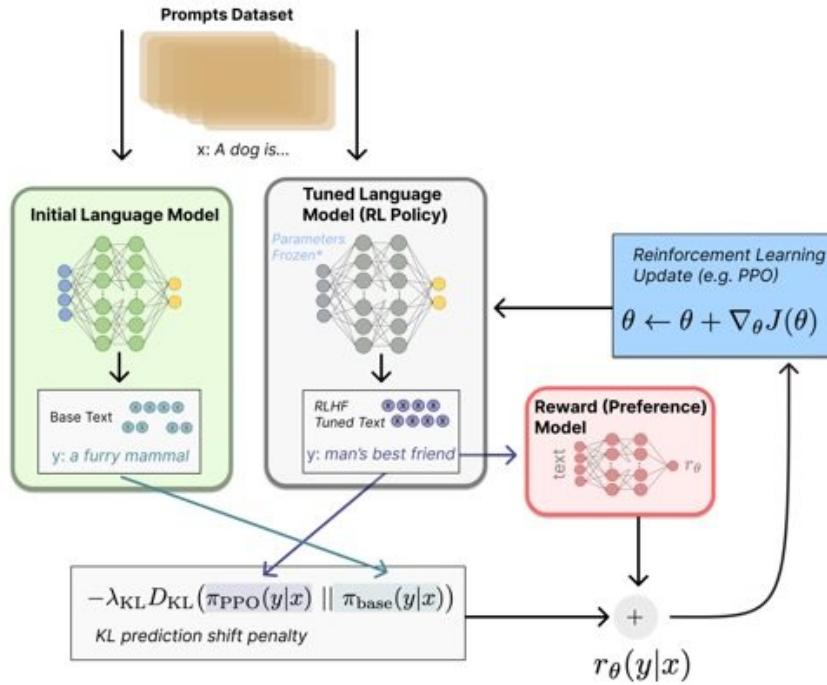


Figura 2.10. Processo di affinamento di un modello di linguaggio attraverso RLHF. Il processo inizia con un dataset di *prompt*, ovvero *input* testuali, che vengono forniti al modello iniziale. Il modello linguistico di base riceve i *prompt* e genera delle risposte base. Tali risposte vengono poi modificate attraverso un modello di ricompensa (*Reward Model*), addestrato per prevedere i *feedback* umani basandosi su coppie di esempi di prompt e risposta (rappresentati da un insieme di nodi rosa). All'interno della fase *RL Policy*, poi, il modello linguistico iniziale viene aggiornato utilizzando il modello di *reward* per generare risposte che meglio si allineano con i feedback umani. Questo aggiornamento viene fatto congelando i parametri originali ed aggiustando solo una parte del modello per riflettere i *feedback* umani (*RLHF Tuned Text*). Durante il processo di aggiornamento, viene applicata una *Shift Penalty* per assicurarsi che le predizioni del modello non si discostino troppo da quelle del modello originale, questo è importante per mantenere la coerenza e l'affidabilità del modello stesso. Il modello viene, poi, aggiornato iterativamente utilizzando un algoritmo di apprendimento rinforzato, come il *Proximal Policy Optimization* (PPO), guidato dal gradiente della funzione obiettivo, che è una combinazione delle previsioni del modello di *reward* e della penalità.

Fonente: [Baeldung](#)

La metodologia RLHF è basata su tre processi chiave:

1. Pre-addestramento di un modello di linguaggio

Il processo inizia con il pre-addestramento di un modello di linguaggio utilizzando obiettivi tradizionali di apprendimento automatico, come GPT-3 o altri modelli di tipo Transformer. Il modello iniziale può essere ulteriormente perfezionato con l'aggiunta di testo supplementare o condizioni specifiche, se necessario, prima di procedere alla fase successiva.

2. Raccolta di dati e addestramento del modello di ricompensa

La fase successiva coinvolge la raccolta di dati con *feedback* umano. Gli annotatori forniscono valutazioni sulle risposte del modello di linguaggio in diverse situazioni e svolgono una comparazione tra le risposte alternative. I dati raccolti vengono impiegati per addestrare il “modello di ricompensa” (*Reward Model*), che rappresenta una stima della qualità del modello attuale e delle sue risposte, in base alle valutazioni degli annotatori.

3. Fine-tuning del modello con *Reinforcement Learning Policy*

Infine, il modello di linguaggio originale viene perfezionato attraverso il processo di apprendimento per rinforzo (*Reinforcement Learning Policy*), ottimizzando il modello in base alle ricompense ottenute dal modello di ricompensa. Fondamentalmente, il modello impara a migliorare le sue risposte e le sue prestazioni generali basandosi sul feedback umano e sulle valutazioni ricevute.

Il Reinforcement Learning From Human Feedback si configura come una strategia efficace nell’ambito dell’*instructor tuning*, contribuendo ad ottimizzare le performance dei modelli mediante iterazioni mirate e l’integrazione di feedback umano. Tuttavia, una delle sfide principali del RLHF riguarda la sua scalabilità ed il costo associato al coinvolgimento umano. Di conseguenza, questo approccio può risultare più lento ed oneroso rispetto all’apprendimento non supervisionato. In aggiunta, il feedback umano può mostrare variazioni significative a seconda del compito e delle preferenze individuali delle persone coinvolte.

2.1.4 Applicazioni dei LLM

Il settore dei Large Language Models costituisce, tutt’ora, un’area attiva di ricerca e sviluppo. La complessità legata alle risorse necessarie per la creazione e la formazione di un LLM impone che tali sforzi siano prevalentemente intrapresi da organizzazioni di ampia portata. Tuttavia, l’introduzione di ChatGPT da parte di OpenAI ha innescato una significativa svolta in questo contesto.

I LLM possono essere suddivisi all’interno di tre categorie, come mostrato in **Figura 2.11**:

- **Pre-training models**, quali GPT-3/GPT-3.5, T5 [67] e XLNet [68], vengono addestrati su estese quantità di dati, consentendo loro di acquisire una vasta gamma di modelli e strutture linguistiche. Questi modelli eccellono nella generazione di testo

coeso e grammaticalmente corretto su una molteplicità di argomenti e vengono impiegati come punto di partenza per ulteriori fasi di formazione e perfezionamento destinati a compiti specifici.

- **Fine-tuning models**, come BERT [69], RoBERTa [70] e ALBERT [71], subiscono una fase iniziale di pre-addestramento su un vasto insieme di dati, seguita da un'ottimizzazione su un set di dati più ristretto, mirato ad una specifica attività. Questi modelli dimostrano una notevole efficacia nelle attività quali la *sentiment analysis*, la risposta alle domande e la classificazione del testo, trovando spesso impiego in contesti industriali in cui si richiedono modelli linguistici adattati a compiti specifici.
- **Multimodal models**, quali CLIP [72] e DALL-E , entrambi sviluppati da OpenAI, integrano il testo con altre modalità, come immagini o video, per creare modelli linguistici più robusti. Queste architetture sono in grado di comprendere le relazioni tra immagini e testo, consentendo la generazione di descrizioni testuali di immagini o addirittura la generazione di immagini da descrizioni testuali.

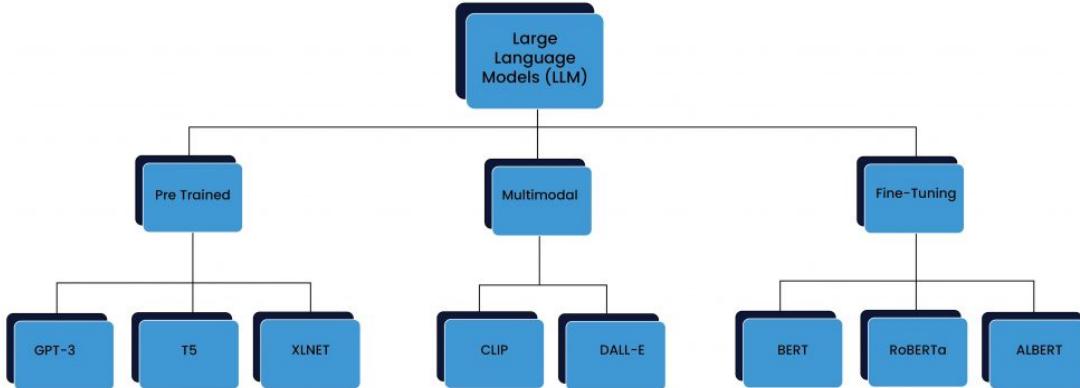


Figura 2.11. Classificazione dei modelli linguistici di grandi dimensioni (LLM) in tre categorie (*Pre-Trained*, *Multimodal*, *Fine-tuning*) indicando, inoltre, i modelli più utilizzati per ciascuna di essa.

Fonte: [ScribbleData](#)

Ciascuna categoria di LLM presenta punti di forza e di debolezza unici, la scelta del modello da utilizzare dipende strettamente dal caso d'uso specifico.

I modelli linguistici rappresentano una tecnologia versatile con un'ampia gamma di applicazioni. Tra le applicazioni più diffuse si trovano gli assistenti virtuali, che facilitano l'interazione tra Intelligenza Artificiale ed utenti umani, soprattutto nel contesto del supporto per i clienti aziendale, la traduzione automatica e la generazione di contenuti. Tuttavia, l'utilizzo dei modelli linguistici si estende ben oltre queste applicazioni, trovando contesti specifici in cui la loro utilità raggiunge una profondità significativa. Ad esempio, i motori di ricerca sfruttano i LLM per migliorare la precisione dei risultati restituiti, mentre nel settore dell'istruzione, questi modelli possono personalizzare l'apprendimento e suggerire testi e approfondimenti rilevanti.

Un'applicazione altrettanto cruciale si riscontra nel miglioramento del *Natural Language Processing* (NLP), utilizzato per rendere più accurata l'analisi delle impressioni dei clienti, inclusa l'interpretazione del *sentiment*. Settori di specializzazione avanzati, come il settore sanitario, impiegano i LLM per analizzare in dettaglio i dati dei pazienti, al fine di offrire trattamenti personalizzati oppure vengono impiegati in maniera più approfondita per esaminare le interazioni comunicative tra medici e pazienti, con l'obiettivo di creare fascicoli sanitari più approfonditi ed accurati dal punto di vista diagnostico.

Le possibilità di impiego dei modelli linguistici sono ampie e diverse, estendendosi anche a settori come il *marketing* e i media per la pubblicazione di contenuti. Tuttavia, il livello più avanzato di astrazione emerge dalla sinergia tra reti neurali e l'utilizzo dei LLM, soprattutto quando si tratta di letture dei testi e di estrazione di contenuti utili per analisi approfondite. All'interno di questo lavoro, i LLM si configurano come strumenti indispensabili per comprendere, elaborare e sfruttare in modo significativo il vasto mondo di informazioni contenute nei *report* aziendali, in particolare all'interno di *report ESG*, contribuendo così alla creazione di approcci innovativi ed avanzati nell'automatizzazione per il reperimento ed analisi di informazioni espresse in linguaggio naturale.

2.1.5 Rischi correlati all'utilizzo di LLM

Nonostante il notevole impatto rivoluzionario della tecnologia dei LLM, essa si trova ancora in una fase embrionale dello sviluppo, affrontando problemi e limitazioni ampiamente riconosciuti che richiedono una ponderata attenzione. Tali questioni presentano possibili impedimenti per diverse applicazioni, necessitando di un superamento accurato. Inoltre, considerando che gli LLM sono un'entità accessibile al pubblico, sono suscettibili ad un utilizzo sia a scopi benefici che malevoli. Nonostante la capacità dell'Intelligenza Artificiale generativa di produrre contenuti utili e precisi a beneficio della società, è essenziale affrontare il concreto rischio di diffusione di disinformazione, la quale potrebbe ingannare gli utenti attraverso la creazione di contenuti fuorvianti.

Alcune problematiche che si possono riscontrare, sono le seguenti [73]:

- **Allucinazioni**

Il fenomeno delle allucinazioni emerge come un elemento significativo nella capacità generativa della tecnologia basata sui LLM. Questa tecnologia può produrre risultati che non sono sempre strettamente vincolati al contesto o alla conoscenza incorporata nel modello. Questa dinamica si riflette nella produzione di testi che potrebbero mancare di coerenza logica rispetto all'input fornito o essere semanticamente scorretti, pur mantenendo un'apparente plausibilità per il lettore umano.

- **Inclinazione**

La tendenza predominante nelle applicazioni dei LLM consiste nell'utilizzare modelli "pre-addestrati", poiché la creazione di un modello completamente nuovo risulta eccessivamente onerosa per la maggior parte delle organizzazioni. Tuttavia, è essenziale notare che la disponibilità di dati perfettamente bilanciati è un caso raro,

perciò ogni modello presenterà inevitabilmente distorsioni in determinati aspetti. Ad esempio, potrebbe verificarsi un disequilibrio tra testi in lingua inglese e cinese o tra conoscenze relative al liberalismo ed al conservatorismo. Quando gli utenti basano le proprie decisioni sulle raccomandazioni di tali modelli, c'è il rischio che i loro pregiudizi o inclinazioni influenzino le decisioni in modo ingiusto o discriminatorio.

- **Consistenza**

La coerenza si configura come un elemento complesso nella produzione di *output* da parte della tecnologia basata su LLM, poiché non può garantire la ripetizione sistematica di risultati identici quando sono forniti *input* uguali. Questa variabilità nel processo è intrinseca alla natura dei LLM, i quali funzionano come modelli probabilistici, prevedendo costantemente la parola successiva in base a specifiche distribuzioni di probabilità durante l'analisi dei dati di *input*.

- **Bypassare il Filtro**

Gli strumenti basati su LLM sono comunemente dotati di dispositivi di sicurezza per prevenire la generazione di contenuti indesiderati, come materiali per adulti, violenti o proprietari. Tuttavia, in alcune situazioni, tali filtri possono essere aggirati mediante la manipolazione dei dati in ingresso, come nel caso del *prompt injection*, ovvero la pratica di fornire *input* specifici al fine di malipolare le risposte di un modello. In tal modo gli LLM possono essere utilizzati come strumenti in pericolosi attacchi di *social engineering* oppure utilizzati per creare posta elettronica di *spear phishing*, contenuti *deep fake* o anche codici dannosi, aumentando così il rischio di manipolazione ed inganno.

- **Privacy dei Dati**

Dal punto di vista della progettazione, i modelli linguistici, sono strutturati solo per ricevere *input* e produrre *output* in un formato non crittografato. Quando un LLM di proprietà viene erogato come servizio, come avviene nel caso di OpenAI, i fornitori di servizi raccolgono un vasto volume di dati sensibili o classificati. Le conseguenze di un incidente di violazione della sicurezza dei dati possono rivelarsi, perciò, catastrofiche.

Un altro aspetto di notevole importanza riguarda la potenziale violazione della proprietà intellettuale. Gli LLM, grazie all'abilità di generare contenuti che somigliano a materiale coperto da *copyright*, costituiscono un rischio significativo per le organizzazioni che basano il proprio vantaggio competitivo sulla proprietà intellettuale.

In conclusione, oltre alle sfide di natura tecnica, è di fondamentale importanza esaminare ed affrontare le implicazioni etiche e di sicurezza associate all'utilizzo degli LLM. La consapevolezza di tali rischi riveste un ruolo cruciale per permettere di adottare misure preventive ed assicurare un utilizzo responsabile di queste avanzate tecnologie linguistiche.

2.2 Obiettivo del progetto

Come evidenziato nella sezione 1.2, all'interno del contesto degli investimenti sostenibili i *report ESG* (Environmental, Social and Governance) assumono un ruolo di fondamentale

importanza. Tali rapporti consentono agli *stakeholder* di valutare le prestazioni delle aziende in termini di sostenibilità e responsabilità sociale. L'aumento dell'interesse nei confronti della sostenibilità ha generato un significativo aumento sia nel numero che nella complessità di tali rapporti, rendendo l'analisi di questi sempre più impegnativa e *time-consuming*. Nonostante ciò, l'importanza di condurre un'analisi accurata e sistematica dei rapporti ESG è imprescindibile per garantire decisioni di investimento informate e responsabili.

All'interno di tale contesto, l'obiettivo della presente ricerca è sviluppare ed implementare una soluzione basata su modelli avanzati di Large Language Models per l'estrazione automatica dei *Key Performance Indicators* (KPI), metriche quantificabili utilizzate per valutare il successo e le prestazioni di un'organizzazione, presenti all'interno di *report* ESG in formato PDF. Lo scopo principale è semplificare il processo di raccolta dati, riducendo la necessità di intervento manuale ed aumentando l'efficienza dell'analisi ESG.

Pertanto l'obiettivo primario di questa ricerca consiste nella progettazione ed implementazione di un *framework* basato sull'intelligenza artificiale, che sfrutta le più recenti innovazioni dei modelli generativi per condurre un'analisi semantica avanzata dei testi. Il fine è estrarre in modo efficiente e preciso i valori associati ad una lista di KPI dai *report* ESG a disposizione.

Questo lavoro si propone di sviluppare un sistema in grado di:

- Identificare i KPI rilevanti nell'ambito ESG oggetto dell'analisi;
- Estrarre le informazioni descrittive correlate a ciascun KPI, inclusi i rispettivi valori;
- Creare un database strutturato che contenga i KPI estratti e le aziende corrispondenti, facilitando ulteriori analisi e confronti.

L'obiettivo principale consiste nell'esaminare l'efficacia di tale struttura attraverso l'implementazione di diversi modelli di linguaggio. Pertanto, lo scopo più ampio di questo studio è comprendere come l'utilizzo efficace dei LLM possa consentire, alle aziende interessate, di interpretare ed utilizzare correttamente i dati, migliorando così la propria competitività sul mercato ed ottenendo vantaggi strategici.

L'approccio adottato si baserà su tecnologie avanzate nel campo dei modelli generativi, focalizzandosi sull'ottimizzazione della precisione e dell'efficienza nell'elaborazione dei dati testuali. La corretta identificazione ed estrazione di queste informazioni sono fondamentali per fornire una base affidabile per le analisi successive e per facilitare la comparazione tra diverse entità aziendali.

Il presente studio si posiziona all'incrocio tra due ambiti cruciali: la *data science* e la finanza sostenibile. Tale connubio rappresenta un settore in costante crescita, rivestendo un ruolo fondamentale nell'orientamento delle future politiche di investimento. L'approccio proposto si caratterizza per l'impiego di metodologie all'avanguardia nel trattamento del linguaggio naturale, mirando a superare le limitazioni intrinseche ai convenzionali metodi di analisi documentale, spesso basati su regole rigide o estrazione di testo mediante

parole chiave.

L'implementazione di questo sistema di estrazione di KPI avrà un impatto significativo sulla capacità delle parti interessate al fine di monitorare e valutare le *performance* ESG delle aziende. Per gli analisti finanziari, ciò si tradurrà in una riduzione dei tempi di analisi ed in un incremento della qualità delle informazioni a loro disposizione. Per le aziende, l'automazione del processo di *reporting* potrebbe comportare una maggiore coerenza e comparabilità dei dati ESG, promuovendo una maggiore trasparenza e responsabilità, agevolando un allineamento più stretto tra gli obiettivi finanziari con quelli di sostenibilità.

3. Modelli e strumenti utilizzati

Nel presente capitolo, viene esaminata l’architettura implementata all’interno del presente lavoro, delineando il flusso dei dati e descrivendo gli strumenti fondamentali utilizzati per l’analisi. Inizialmente, si esplora la sorgente dati, rappresentata dal sito di *Carbon Disclosure Project* (CDP) [74]. Utilizzando l’ambiente Azure Databricks, una robusta piattaforma per l’analisi e la manipolazione dei dati, si estraggono i *report ESG* pertinenti presenti nella piattaforma CDP, per poi memorizzarli tramite il servizio di archiviazione di oggetti Azure Blob Storage, offerto anch’esso da Microsoft Azure, che consente una gestione efficiente dei dati su larga scala. Successivamente, i documenti vengono nuovamente elaborati in Azure Databricks attraverso un processo di modellazione e trasformazione impiegando modelli avanzati di generative AI, tra cui **GPT-3.5 Turbo**, **Dolly**, **LLAMA2** e **Gemini**. I risultati di tali operazioni vengono quindi archiviati nuovamente nel Azure Blob Storage e successivamente visualizzati tramite Power BI, la piattaforma di analisi aziendale di Microsoft, che permette la visualizzazione, l’analisi e la condivisione intuitiva ed interattiva dei dati attraverso *dashboard* e *report* dinamici.

Dopo aver fornito una panoramica sull’architettura adoperata, si procede con un’analisi relativa alle caratteristiche e le funzionalità delle famiglie dei modelli impiegati, ovvero **GPT**, **Dolly**, **LLAMA2** e **Gemini**. Ogni modello viene esaminato in profondità, illustrando le varie tipologie ed iterazioni di cui è composto, oltre a descrivere le sue peculiari caratteristiche tecniche. Questo approfondimento mira a fornire una visione esaustiva degli strumenti utilizzati, indicando le specifiche tipologie scelte per affrontare le analisi e permettendo di dare un quadro più completo della ricerca effettuata.

3.1 Architettura

L’architettura impiegata in questo progetto di tesi è strutturata tramite diversi componenti. Di seguito, viene fornita una panoramica dei principali elementi di cui è composta:

- **Data Source:** All’inizio del flusso dei dati, la sorgente primaria è rappresentata dal sito della compagnia “Carbon Disclosure Project” (CDP). CDP è un’organizzazione *no-profit* di portata internazionale che offre un sistema globale di misurazione e rendicontazione ambientale rivolto ad imprese, autorità locali, governi ed investitori. L’obiettivo principale di tale organizzazione è, quindi, fornire un meccanismo per misurare, monitorare, gestire e condividere informazioni sul cambiamento climatico su scala globale. L’organizzazione sostiene quattro programmi distinti: *Climate Change Program*, *Water Program*, *Forests Program* e *Supply Chain Program*. Inoltre, gestisce un programma specifico dedicato a città e regioni, ovvero *Cities, States and Regions Program*.

Per ciascuno di questi programmi, CDP ha sviluppato questionari specifici, contenenti le informazioni relative ai programmi di riferimento, che le parti interessate (come imprese, città ed enti governativi) devono fornire. Queste informazioni vengono raccolte all’interno di una piattaforma accessibile [74] e successivamente

valutate dagli *stakeholder*. Questo processo contribuisce alla creazione di una base comune di informazioni rilevanti sia per gli investitori che per i governi, facilitando la comprensione e la gestione dei rischi ambientali e sostenendo azioni concrete per affrontare il cambiamento climatico.

In particolare, l'azione di CDP si articola attorno a quattro obiettivi strategici volti a promuovere un sistema economico globale, improntato alla sostenibilità ambientale e capace di mitigare gli impatti dannosi del cambiamento climatico. In particolare hanno come obiettivi strategici:

1. *Incremento della trasparenza aziendale.*

CDP collabora strettamente con analisti finanziari per aumentare la trasparenza delle aziende riguardo al loro impatto ambientale. Questo coinvolgimento mira ad identificare le realtà con maggiore rischio ambientale ed a fornire informazioni utili ai cittadini per orientare le loro decisioni di acquisto e di investimento.

2. *Integrazione della performance ambientale nelle decisioni aziendali e di investimento.*

L'organizzazione promuove l'integrazione della *performance* ambientale nelle decisioni aziendali e di investimento. Ciò comporta l'adattamento delle pratiche di reportistica ambientale alle esigenze del settore finanziario e l'incoraggiamento degli investitori a considerare i programmi di riduzione delle emissioni come parte integrante della valutazione delle imprese.

3. *Supporto alle città nel ridurre l'inquinamento e nell'adattarsi al cambiamento climatico.*

Il CDP offre sostegno alle città nel loro impegno a ridurre l'inquinamento ed a sviluppare strategie di adattamento al cambiamento climatico. Questo supporto include la diffusione globale delle migliori pratiche ed il sostegno a programmi volti a gestire l'impatto ambientale delle aree urbane.

4. *Promozione di politiche e leggi a difesa dell'ambiente.*

L'organizzazione lavora a fianco di governi e organizzazioni sovranazionali per promuovere politiche e leggi a difesa dell'ambiente. Questo impegno si traduce nel fornire dati utili per supportare le decisioni normative e per sensibilizzare l'opinione pubblica sull'influenza delle aziende sulle scelte politiche relative all'ambiente.

All'interno di questo lavoro sono stati utilizzati i *report* relativi al programma *Climate Change*, in linea con l'obiettivo generale delineato nella sezione 2.2.

- **Ingest:** I *report* presenti nella fonte dati CDP, vengono scaricati ed elaborati attraverso l'ambiente Azure Databricks, piattaforma di analisi dati basata su Apache Spark che offre un ambiente collaborativo per lo sviluppo, la gestione e l'esecuzione di *workflow* di analisi dati avanzati, sfruttando il calcolo distribuito e l'intelligenza artificiale. All'interno di questa fase, i *report* vengono scaricati mediante la metodologia di *web scraping*, ovvero mediante l'estrazione dei dati per mezzo di un codice *python* automatizzato. Tale procedura è stata condotta in conformità con le politiche della piattaforma per garantire la conformità normativa e legale.

- **Store:** I dati elaborati, nella fase precedente, sono memorizzati all'interno del Azure Blob Storage, servizio di archiviazione *cloud* di Microsoft progettato per memorizzare grandi quantità di dati non strutturati, come immagini, video, documenti e file di backup, offrendo scalabilità, affidabilità ed accesso tramite API REST. In questo modo si garantisce un'archiviazione strutturata e sicura dei dati, pronta per essere sottoposta ad ulteriori analisi.
- **Orchestration:** In questa fase, i dati vengono processati all'interno di Azure Databricks al fine di affrontare la fase “**Transform & Model**”. Questa fase comprende la ricerca dei KPI tramite l'utilizzo di Large Language Models avanzati, in particolare vengono testati i modelli: GPT 3.5 Turbo, LLAMA2, Dolly V2 e Bard. Prima di intraprendere la fase successiva, ovvero l'ultima, i dati prodotti dai modelli ritornano alla fase di **Store**, questo perchè ogni volta che un modello elabora degli *output* viene salvato all'interno del Azure Blob Storage.
- **Visualize:** Infine, i dati trasformati e modellati vengono prelevati dal Azure Blob Storage e visualizzati attraverso Power BI, piattaforma di analisi dei dati e di *business intelligence* sviluppata da Microsoft, che consente di visualizzare, analizzare e condividere dati per ottenere *insight* aziendali. In questo modo si sviluppa un'interfaccia intuitiva ed interattiva per esplorare i risultati in modo efficace.

Al fine di fornire una rappresentazione visiva, viene presentato un diagramma che illustra l'architettura del sistema attraverso le diverse fasi del processo, precedentemente descritte, all'interno della **Figura 3.1**, evidenziando le interazioni tra le componenti principali.

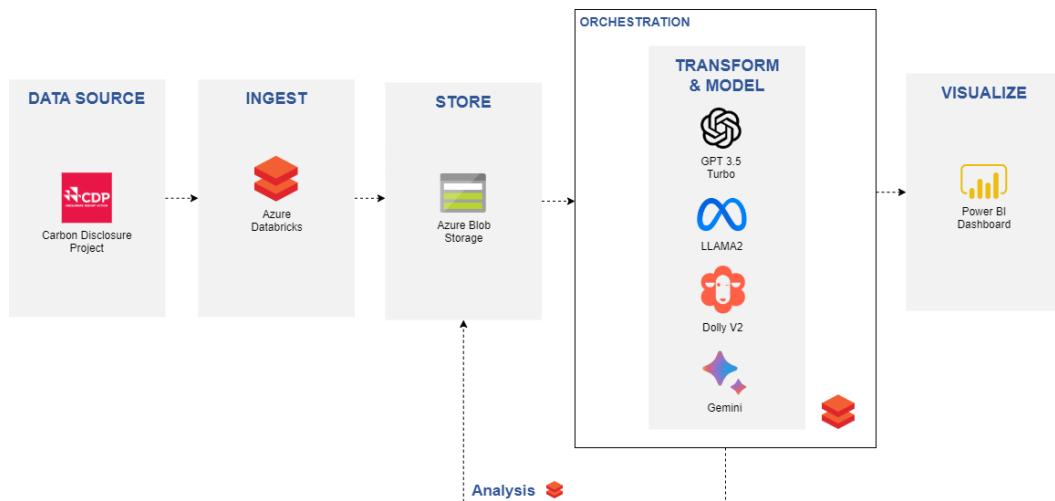


Figura 3.1. Architettura utilizzata all'interno del presente progetto di tesi.

Nel contesto di questa ricerca, è stato impiegato l'ambiente di Azure Databricks come piattaforma di sviluppo per la realizzazione dei codici implementativi. In particolare, è

stato configurato un *cluster* 12.2 LTS ML, il quale includeva Apache Spark 3.3.2, supporto GPU e Scala 2.12, come rappresentato nella **Figura 3.2**. Per *cluster* si intende l’insieme di risorse computazionali a disposizione all’interno dell’ambiente Azure Databricks. Questa configurazione specifica è stata ottimizzata per l’analisi e l’apprendimento automatico, integrando strumenti come Apache Spark, un *framework open-source* progettato per l’elaborazione distribuita dei dati, ed il supporto GPU, che consente prestazioni di calcolo elevate grazie alle unità di elaborazione grafica. Scala 2.12 è il linguaggio di programmazione principale utilizzato per scrivere e gestire il codice all’interno di questo ambiente. La versione 2.12 è una delle *release* stabili di Scala, che offre miglioramenti e nuove funzionalità rispetto alle versioni precedenti, contribuendo così ad una maggiore efficienza nello sviluppo e nella gestione del codice.



Figura 3.2. Configurazione del *cluster* all’interno dell’ambiente Azure Databricks. È stato implementato un *cluster* 12.2 LST ML, che rappresenta il nucleo operativo per le attività di analisi e predizione. Tale *cluster* è alimentato da una combinazione di tecnologie avanzate, tra cui Apache Spark 3.3.2, un motore di calcolo distribuito noto per la sua capacità di elaborare grandi quantità di dati in modo efficiente e scalabile. L’integrazione delle unità di elaborazione grafica (GPU) contribuisce ad amplificare ulteriormente la potenza computazionale del sistema, consentendo la gestione di compiti di Machine Learning che richiedono elevate risorse computazionali con una rapidità ed una precisione. Il *framework* Scala 2.12 costituisce l’infrastruttura su cui si basa l’intero sistema, offrendo una sintassi potente per lo sviluppo di algoritmi di Machine Learning complessi.

All’interno del capitolo 4, denominato “Implementazione”, verranno forniti maggiori dettagli tecnici per ognuna delle fasi descritte.

3.2 Modelli utilizzati

Questa sezione è dedicata all’analisi dei modelli impiegati nel processo di definizione dei KPI. In particolare, sarà fornita una descrizione generale delle caratteristiche di ciascuna categoria di modelli e, successivamente, un approfondimento sul modello specificamente adottato nell’ambito della presente ricerca.

3.2.1 GPT

I modelli GPT, acronimo di “Generative Pre-trained Transformer” [2], denotano una serie di sofisticati modelli linguistici nell’ambito dell’Intelligenza Artificiale, realizzati dal laboratorio di Artificial Intelligence americano OpenAI. Questi modelli sono eredi concettuali dell’architettura Transformer, ovvero modelli di apprendimento automatico basati sull’attenzione (*attention-based*) che sono stati introdotti per la prima volta nel 2017 da Vaswani et al. nella loro pubblicazione “Attention Is All You Need” [75], precedentemente anticipato all’interno della sezione 2.1.2. Tale architettura costituisce il fondamento concettuale su cui si erigono le diverse istanze appartenenti alla famiglia GPT. I Transformers, avvalendosi del meccanismo di auto-attenzione, dimostrano una notevole capacità di apprendimento relazionale tra gli elementi che costituiscono una sequenza. Questa peculiarità permette l’elaborazione parallela dei dati in *input*, portando ad un incremento sostanziale dell’efficienza del modello nel trattamento di sequenze dati. In particolare, la facoltà di rilevare ed interpretare le interdipendenze tra parole o elementi sequenziali, indipendentemente dalla loro posizione all’interno del testo o della sequenza, costituisce un avanzamento significativo per l’analisi di dati sequenziali. Tale capacità risulta essere di primaria importanza nelle applicazioni di elaborazione del linguaggio naturale (NLP), dove la comprensione e l’interpretazione delle relazioni semantiche e sintattiche tra le parole, considerate non più come singole unità ma come componenti di un sistema di significati interconnessi, rappresentano un requisito fondamentale per l’efficacia dell’elaborazione. Pertanto, proprio grazie a questa architettura che oggi siamo in grado di beneficiare di strumenti come ChatGPT che si basa quasi completamente sull’architettura Transformer, con solo alcune lievi variazioni. L’origine di questa tecnologia risale a GPT-1, quando OpenAI introdusse parte dell’architettura Transformer, concentrandosi principalmente sul suo componente *decoder*, per addestrare il modello. OpenAI ha apportato alcune modifiche al componente *decoder* del Transformer per consentire a GPT di essere *task-independent*, ovvero in grado di affrontare una vasta gamma di problemi senza dipendere da un compito specifico. Queste modifiche includono l’aggiunta di componenti come *dropout* e strati di attivazione **GeLu** (Gaussian Error Linear Unit), combinazione di una funzione lineare ed una funzione non lineare che introduce una componente probabilistica tramite la funzione di errore gaussiana migliorando la propagazione del gradiente durante l’addestramento.

La *self-attention* rappresenta un aspetto fondamentale nei modelli Transformer, permettendo al modello di comprendere il contesto di ciascuna parola all’interno di una frase analizzando tutte le altre parole presenti. Questo processo assegna un punteggio tra un vettore di “query” e uno di “chiave” per ogni parola, determinando così l’importanza di ciascuna parola presente nella frase per interpretare una specifica parola. I punteggi vengono quindi normalizzati attraverso una funzione **softmax**, garantendo che la loro somma sia pari a 1 e possano essere utilizzati come pesi. Questo meccanismo permette a ciascuna parola di acquisire informazioni da tutte le altre parole, indipendentemente dalla loro posizione nella frase, superando le limitazioni dei modelli ricorrenti nella gestione delle dipendenze a lunga distanza. L’*output* di questa auto-attenzione ponderata, basato sui punteggi di attenzione, viene utilizzato per creare una nuova rappresentazione per ogni parola, catturando le informazioni specifiche del contesto. Nei modelli Transformer,

viene impiegato il concetto di “attenzione multi-testa”, che prevede l’esecuzione di diverse operazioni di auto-attenzione in parallelo, ognuna con trasformazioni lineari distinte dei vettori di *query*, chiave e valore originali. Gli *output* di ciascuna “testa” vengono quindi combinati e trasformati linearmente per produrre l’*output* finale, permettendo al modello di catturare varie relazioni nei dati. Di conseguenza, la *self-attention* rappresenta un meccanismo cruciale che consente ai modelli GPT ed ai Transformer di considerare l’intera sequenza di *input* e valutare dinamicamente l’importanza di ciascuna parte durante l’elaborazione della sequenza, rendendoli estremamente efficaci in diversi compiti di elaborazione del linguaggio naturale.

Nel corso del suo sviluppo, la tecnologia GPT ha attraversato diverse fasi evolutive, culminando nelle quattro iterazioni principali: GPT-1, GPT-2, GPT-3 e GPT-4. Di seguito viene data una breve descrizione per ognuna delle iterazioni:

- **GPT-1**

Durante l’avvento dei modelli linguistici di grandi dimensioni (LLM), quali BERT [69], nel 2018 OpenAI pubblicò il suo articolo intitolato “Improving Language Understanding by Generative Pre-Training” [76], che introduceva il pionieristico sistema di trasformatore generativo pre-addestrato, noto come GPT-1. Il modello GPT-1, rappresenta il primo modello della serie GPT segnando un significativo avanzamento nel campo dell’elaborazione del linguaggio naturale (NLP). Addestrato su 40GB di dati testuali e caratterizzato da 117 milioni di parametri, fu sottoposto ad un processo di addestramento utilizzando un vasto *corpus* di dati testuali provenienti dal *web*. Tale modello ha ottenuto risultati all'avanguardia per compiti di modellazione come LAMBADA [77] e ha dimostrato una prestazione competitiva per compiti come GLUE [78] e SQuAD [79]. LAMBADA (LAnguage Model Benchmark for Broad Applications) è un *benchmark* progettato per valutare la capacità dei modelli di linguaggio di comprendere e generare testo in contesti più ampi e complessi. GLUE (General Language Understanding Evaluation) è un ulteriore *benchmark* progettato per valutare le capacità di comprensione del linguaggio naturale dei modelli di intelligenza artificiale. SQuAD (*Stanford Question Answering Dataset*), invece, è un dataset progettato per valutare la capacità dei modelli di linguaggio di rispondere a domande basate su un contesto di testo.

Con una lunghezza massima relativa al contesto pari a 512 *token* (circa 380 parole), il modello GPT-1 era in grado di conservare informazioni per frasi o documenti relativamente brevi per richiesta. Tuttavia, le impressionanti capacità di generazione di testo del modello e la sua forte *performance* sui compiti standard hanno fornito lo stimolo per lo sviluppo del modello successivo.

- **GPT-2**

Rilasciato nel 2019, GPT-2 costituiva una versione ingrandita e più potente del modello GPT originale. In particolare, il modello GPT-2 conserva le stesse caratteristiche architetturali del modello precedente, tuttavia, viene addestrato su un *corpus* di dati testuali ancora più ampio. Infatti, GPT-2 può gestire il doppio della dimensione dell’*input*, consentendogli di elaborare campioni di testo più estesi. Con un impressionante numero di 1,5 miliardi di parametri, questo sistema dimostrava

una maggiore capacità nella generazione di testi coerenti e contemporaneamente pertinenti dal punto di vista contestuale. A causa delle preoccupazioni legate ad un possibile uso improprio, specialmente nella generazione di notizie false, OpenAI inizialmente limitò l'accesso alla versione completa del modello, ma successivamente la rese disponibile al pubblico.

Il modello GPT-2 rappresenta un notevole avanzamento rispetto al suo predecessore. Una delle sue innovazioni principali è l'utilizzo della tecnica del Modified Objective Training (MO) durante la fase di pre-addestramento. Questa tecnica mira ad ottimizzare i modelli linguistici, affrontando il problema delle previsioni potenzialmente incoerenti o poco pertinenti. Essa incorpora informazioni aggiuntive come *Parts of Speech* (ovvero le classi grammaticali alle quali le parole possono appartenere in una lingua, come sostantivi, verbi, etc) e *Subject-Object Identification* (processo di identificazione e distinzione delle entità grammaticali che fungono da soggetto e oggetto all'interno di una frase o di un testo). Tale operazione è necessaria per garantire che il contesto venga adeguatamente considerato durante la generazione di testo, consentendo al modello di produrre *output* più coerenti ed informativi.

Un'altra innovazione consiste nella normalizzazione dei livelli, che migliora sia il processo di addestramento che le prestazioni della rete neurale. Questa tecnica mira a mitigare il problema della *Internal Covariate Shift*, causato dalle variazioni nei parametri della rete durante il processo di apprendimento. Inoltre, GPT-2 implementa algoritmi di campionamento avanzati rispetto al modello precedente GPT. Tra questi, il “campionamento top-p” è particolarmente significativo: seleziona solo i *token* con una probabilità cumulativa superiore a una determinata soglia durante il processo di generazione, evitando così *token* a bassa probabilità e producendo un testo più coerente e diversificato. La regolazione della temperatura dei *logit* controlla il grado di casualità nel testo generato: temperature più basse producono un testo più conservativo e prevedibile, mentre temperature più alte generano un testo più creativo e inaspettato. Infine, l'opzione di campionamento incondizionato consente agli utenti di esplorare le capacità generative del modello, permettendo la generazione di risultati originali e creativi senza vincoli specifici di *input*.

- **GPT-3**

Introdotto nel 2020, il modello GPT-3 ha elevato le capacità dei suoi predecessori ad un livello completamente nuovo. Dotato di un impressionante numero di 175 miliardi di parametri, si è affermato come uno dei più vasti modelli linguistici mai concepiti. Il GPT-3 ha dimostrato una straordinaria comprensione linguistica, generando risposte estremamente realistiche ed al contempo pertinenti in diversi contesti ed alle molteplici richieste effettuate.

In aggiunta alle sue dimensioni maggiorate, il modello GPT-3 ha introdotto una serie di miglioramenti significativi. Tra questi, spicca l'implementazione di GShard, un sistema di parallelismo del modello gigante, che permette di distribuire il modello su più acceleratori. Ciò semplifica l'addestramento e l'esecuzione parallela, specialmente per modelli linguistici di notevoli dimensioni con miliardi di parame-

tri. Inoltre, è stata introdotta la capacità di apprendimento *zero-shot*, che consente a GPT-3 di compiere compiti per i quali non è stato specificamente addestrato. Questo significa che può generare testo in risposta a stimoli nuovi, basandosi sulla sua comprensione generale del linguaggio e del compito richiesto. Allo stesso modo, è stata implementata la capacità di apprendimento *few-shot*, che permette a GPT-3 di adattarsi rapidamente a nuovi compiti e domini con un numero minimo di esempi di addestramento. Questa abilità dimostra una notevole flessibilità nel processo di apprendimento del modello. In aggiunta, GPT-3 supporta la generazione di testo in circa trenta lingue diverse, inclusi l'inglese, il cinese, il francese, il tedesco e l'arabo. Questo ampio sostegno multilingue lo rende un modello linguistico estremamente versatile per una vasta gamma di applicazioni.

Infine, sono state apportate migliorie al processo di campionamento: GPT-3 utilizza un algoritmo di campionamento migliorato che offre maggiore controllo sulla casualità nel testo generato, simile a quanto visto in GPT-2. Inoltre, è stata introdotta l'opzione di campionamento “*prompted*”, che consente la generazione di testo basata su stimoli o contesti specificati dall'utente, ampliando ulteriormente le sue capacità di adattamento ed utilizzo contestuale.

– GPT-3.5

Introdotto nel 2022, i modelli della serie GPT-3.5, analogamente ai suoi predecessori, sono derivati dai modelli GPT-3. Tuttavia, la caratteristica distintiva dei modelli GPT-3.5 risiede nella loro aderenza a politiche specifiche basate sui valori umani, incorporate utilizzando una tecnica chiamata Reinforcement Learning from Human Feedback (RLHF). L'inserimento della componente umana nell'addestramento non supervisionato ha dato grande vantaggio a OpenAI per la creazione di modelli che rispondessero in maniera simile ad un umano ad una determinata *query*. Nel paper “Deep reinforcement learning from human preferences” [80] OpenAI esplora l'apprendimento questa tecnica per rinforzo dal *feedback* umano che, come anticipato nella sottosezione 2.1.3, prevede l'addestramento del modello utilizzando principi di apprendimento per rinforzo, dove il modello riceve ricompense o penalità in base alla qualità ed all'allineamento dei suoi *output* generati con gli esaminatori umani. Integrando questo *feedback* nel processo di addestramento, il modello acquisisce la capacità di imparare dagli errori e migliorare le sue prestazioni, producendo *output* molto simili a quelli umani e permettendo quindi la conversazione. L'obiettivo principale mirava ad allineare i modelli in maniera più precisa alle intenzioni dell'utente, mitigare la tossicità e dare priorità alla veridicità nei loro *output* generati. Questa evoluzione rappresenta uno sforzo cosciente per migliorare l'uso etico e responsabile dei modelli linguistici al fine di fornire un'esperienza utente più sicura ed affidabile.

– GPT-3.5 Turbo

GPT-3.5 Turbo, rilasciato il 1 marzo 2023, rappresenta un'evoluzione migliorata del modello GPT-3.5. OpenAI ha utilizzato, anche in questo caso, la tecnica di RLHF, coinvolgendo il *feedback* umano per valutare le prestazioni del modello in fase di sviluppo. GPT-3.5 Turbo offre molte delle stesse funzio-

nalità di GPT-3.5, tuttavia, si è dimostrato in grado di rispondere a domande molto più versatili e di agire su una gamma più ampia di comandi, dimostrando anche di avere meno probabilità di “allucinazioni”. Questo è dovuto anche all’inclusione di una maggiore complessità del modello, ottenuta tramite l’uso di set di dati più ampi e diversificati durante l’addestramento, insieme ad un’ottimizzazione dell’algoritmo di generazione del testo. In particolare, GPT-3.5 Turbo dimostra una migliore coerenza e coesione nel testo generato, con una riduzione degli errori di comprensione e risposte ambigue. Inoltre, il modello offre maggiori capacità di controllo e personalizzazione per gli utenti, consentendo una maggiore adattabilità alle esigenze specifiche degli utenti.

- **GPT-4**

Il modello GPT-4 costituisce l’ultima iterazione dei modelli di linguaggio sviluppati da OpenAI, pubblicata il 14 marzo 2023. Essendo la quarta versione della serie GPT, questo modello linguistico multimodale di ampie dimensioni dimostra la capacità di comprendere sia il testo che le immagini, generando sempre *output* testuali. Questo modello è in grado di gestire diversi formati di immagini, tra cui documenti con testo, fotografie, diagrammi, grafici, schemi e *screenshot*.

Il processo di addestramento del modello GPT-4 è articolato in diverse fasi, tra cui assume particolare rilievo la fase di “*pre-training*”. In questa fase, il modello è esposto ad un vasto e diversificato *corpus* di dati, consentendogli di sviluppare una comprensione profonda del contesto linguistico e delle relazioni semantiche tra le parole. In particolare, GPT-4 è progettato per predire la parola successiva nelle frasi, sfruttando le informazioni contenute nel testo di *input*. Anche in questa iterazione viene adottato un approccio basato sul *reinforcement learning*, precedentemente discussa nella sezione 2.1.3. Questo metodo permette al modello di migliorare ulteriormente le sue capacità di generazione del linguaggio, integrando sia il *feedback* fornito dagli esseri umani che l’analisi delle risposte generate dal modello stesso. In questo modo, il modello viene costantemente adattato per garantire che le sue risposte siano coerenti con le aspettative umane e rispecchino adeguatamente il contesto comunicativo in cui si inseriscono.

Sebbene OpenAI non abbia divulgato dettagli tecnici come dimensioni del modello, architettura, metodologia di addestramento o pesi del modello in questione, alcune stime suggeriscono che esso comprenda circa 1 trilione di parametri. Il modello base di GPT-4 segue un obiettivo di addestramento simile ai modelli GPT precedenti, mirando a prevedere la parola successiva data una sequenza di parole. Il processo di addestramento ha coinvolto l’uso di un enorme set di dati pubblicamente disponibili su *internet* e dati con licenza. GPT-4 ha dimostrato una prestazione superiore rispetto a GPT-3.5 nei *benchmark* pubblici come TruthfulQA [81]. TruthfulQA è un *benchmark* pubblico utilizzato per valutare la capacità dei modelli di linguaggio naturale di generare risposte accurate ed affidabili. In questo *benchmark*, le domande sono progettate per essere “factuali”, ovvero richiedono risposte basate su fatti oggettivi e verificabili. I modelli sono quindi valutati sulla precisione e l’accuratezza

delle risposte fornite rispetto a un insieme di domande e risposte verificate da fonti affidabili.

– GPT-4 Turbo

Il modello **GPT-4 Turbo**, l'ultima iterazione del modello sviluppato da OpenAI, rappresenta un notevole passo avanti nel campo dell'elaborazione del linguaggio naturale (NLP). Una delle principali migliorie di **GPT-4 Turbo** rispetto al suo predecessore è la sua ampia base di conoscenze. Rispetto al **GPT-4** originale, che integrava dati fino a settembre 2021, il **GPT-4 Turbo** include dati aggiornati fino ad aprile 2023, estendendo significativamente la comprensione del modello sugli sviluppi e gli argomenti recenti. Inoltre, **GPT-4 Turbo** presenta una straordinaria finestra di contesto di 128k *token*, che gli consente di elaborare l'equivalente di circa 96.000 parole in un singolo *prompt*, espandendo notevolmente la sua capacità di gestione dei dati rispetto alla precedente capacità di 32.000 *token*. Tuttavia, va notato che l'*output* del modello è ancora limitato a 4000 *token*.

GPT-4 Turbo migliora ulteriormente le capacità del suo predecessore introducendo funzioni multimodali, consentonogli di elaborare anche le immagini. Questo significa che ora è possibile inserire immagini in **GPT-4 Turbo** per la creazione automatica di didascalie, l'analisi del contenuto visivo ed il riconoscimento del testo all'interno delle immagini. Sebbene questa funzionalità sia attualmente in anteprima, l'implementazione della comprensione visiva amplia considerevolmente le possibilità delle applicazioni di visione artificiale, avvicinandosi sempre più al modo in cui gli esseri umani percepiscono e interagiscono con l'ambiente circostante. Attualmente, l'API consente l'*input* di immagini in formato **base64** oppure un URL diretto dell'immagine. Il costo di questa funzionalità varierà in base alla dimensione dell'immagine in *input*. **GPT-4 Turbo** supera i modelli precedenti nell'esecuzione di attività che richiedono il rispetto preciso delle istruzioni, in particolare nella generazione di formati designati (come la risposta in XML). Introduce inoltre l'innovativa modalità **JSON**, garantendo risposte **JSON** valide. Questo è facilitato dal nuovo parametro API, “**response_format**”, che indirizza il modello a produrre oggetti **JSON** sintatticamente accurati.

Infine, **GPT-4 Turbo** introduce un parametro “**seed**” che garantisce che il modello fornisca completamenti coerenti per la maggior parte del tempo, consentendo risultati riproducibili. Questa funzionalità beta è particolarmente utile per riprodurre le richieste durante il *debug*, creare *unit test* dettagliati (ovvero *test software* per la verifica della correttezza di unità di un programma simulando il suo comportamento in modo controllato) ed ottenere un maggiore controllo sul comportamento del modello.

I modelli GPT costituiscono un esempio significativo di apprendimento non supervisionato, un approccio dove i modelli assimilano informazioni da *corpus* testuali estesi privi di annotazioni o etichette esplicite. Nel corso della fase preliminare di pre-addestramento,

tali modelli elaborano e interiorizzano il linguaggio ed il suo contesto, affinando la loro capacità di comprensione. Successivamente, nella fase di adattamento, essi vengono ottimizzati per eseguire specifici compiti linguistici come traduzione, generazione di riassunti, generazione di risposte a domande specifiche ed altre attività correlate. Questo processo progressivo consente ai modelli di acquisire una maggiore adattabilità e precisione nell’ambito di determinati contesti applicativi.

La società OpenAI offre l’accesso ai suoi modelli GPT-3.5 Turbo, GPT-4 e GPT-4 Turbo tramite un servizio di API REST, consentendo ai programmati di integrare e sfruttare le funzionalità di tali modelli all’interno dei loro progetti. Il costo associato all’utilizzo di questi strumenti varia a seconda del modello selezionato e si basa su una metrica chiave che considera il numero di *token* forniti in *input* al modello ed il numero di *token* generati in *output*. Tali *token* rappresentano le unità di base del testo, che includono singole parole, segni di punteggiatura ed altri elementi linguistici. I modelli Transformer in stile GPT hanno in genere un limite di *token* ben definito: ad esempio, `gpt-3.5-turbo` (ChatGPT) ha un limite di 4096 *token* e `gpt-4-32k` ha un limite di 32768. Poiché viene fornito al modello Transformer la concatenazione dei *token* di ingresso e di tutti i *token* generati in uscita fino a quel momento, il limite del modello si riferisce al numero totale di *token* di *input* più *output*. Di conseguenza, il costo di utilizzo di un modello GPT è direttamente proporzionale al numero di *token* elaborati durante l’interazione con il sistema. Nella **Tabella 3.1**, di seguito, viene illustrata la struttura dei costi.

Modello	Input	Output
<code>gpt-3.5-turbo-0125</code>	\$ 0.0005 / 1K tokens	\$ 0.0015 / 1K tokens
<code>gpt-3.5-turbo-instruct</code>	\$ 0.0015 / 1K tokens	\$ 0.0020 / 1K tokens
<code>gpt-4</code>	\$ 0.03 / 1K tokens	\$ 0.06 / 1K tokens
<code>gpt-4-32k</code>	\$ 0.06 / 1K tokens	\$ 0.12 / 1K tokens
<code>gpt-4-0125-preview</code>	\$ 0.01 / 1K tokens	\$ 0.03 / 1K tokens
<code>gpt-4-1106-preview</code>	\$ 0.01 / 1K tokens	\$ 0.03 / 1K tokens
<code>gpt-4-1106-vision-preview</code>	\$ 0.01 / 1K tokens	\$ 0.03 / 1K tokens

Tabella 3.1. Costi dei modelli GPT.

Dato il costo effettivo dell’utilizzo e la quantità di dati da elaborare, all’interno di questo lavoro di tesi è stato utilizzato il modello `gpt-3.5-turbo-0125`.

3.2.2 Dolly

All’interno del panorama delle soluzioni per l’analisi dati e per l’Intelligenza Artificiale, è possibile notare **Dolly 2.0**: un modello linguistico di grandi dimensioni (LLM) sviluppato in modalità *open-source* dalla società Azure Databricks. Fondata su una piattaforma *cloud*, Azure Databricks offre agli utenti la possibilità di gestire ed analizzare grandi volumi di dati in modo efficiente e flessibile. Il lancio di **Dolly** nel 2023 ha segnato un momento significativo nel campo dei modelli linguistici, poiché è risultato essere il primo LLM ad adottare un approccio “*instruction-tuned*”, conferendogli una notevole versatilità

ed adattabilità per una vasta gamma di compiti linguistici. Questo metodo si concentra sull’addestramento del modello per rispondere in modo efficace ed accurato ad istruzioni linguistiche specifiche, migliorando notevolmente le prestazioni del modello su compiti di natura istruttiva, come la generazione di testi, la risposta a domande specifiche ed il riassunto di documenti.

Tale versione, denominata ufficialmente **Databricks Dolly 2.0** [3], è un modello *open source* che si distingue per la sua ottimizzazione su un set di dati di addestramento raccolto tramite *crowdsourcing* (ovvero attraverso la partecipazione di un vasto numero di persone) da parte dei dipendenti della società Databricks. La peculiarità che rende Dolly 2.0 una scelta valida, risiede principalmente nella sua natura *open source* e nell’accessibilità al *dataset* impiegato per il suo addestramento, il quale è trasparente e liberamente disponibile. Ciò implica che il modello può essere utilizzato per scopi commerciali senza la necessità di pagare per l’accesso alle API REST o di condividere dati con terze parti. Tale scelta riflette un impegno verso la trasparenza e l’accessibilità nel contesto della ricerca e dello sviluppo di soluzioni avanzate di intelligenza artificiale.

Lo sviluppo di Dolly ha avuto inizio con la creazione del modello Dolly 1.0, rilasciato nel Marzo del 2023. Tale versione adottava una metodologia che integrava le risposte generate da ChatGPT nel set di dati utilizzato per addestrare il modello. Questa prima versione si fondava sulle avanzate capacità linguistiche di ChatGPT, modello di notevole potenza e riconoscimento nel momento del rilascio di Dolly 1.0. Sfruttando l’*output* di ChatGPT come parte del processo di addestramento, Dolly 1.0 riusciva a beneficiare delle solide prestazioni linguistiche già sviluppate da ChatGPT, garantendo risposte fluide ed accurate. Tuttavia, Dolly 1.0 non è stato autorizzato per l’uso commerciale dato che il *dataset* conteneva *output* rilasciati da ChatGPT, soggetto ai termini di servizio di OpenAI. Inoltre, l’efficacia di Dolly 1.0 risultava limitata dalla genericità delle risposte di ChatGPT, non sempre ottimali per ogni contesto o specifico compito linguistico. Nell’aprile 2023 è stato introdotto Dolly 2.0, portando miglioramenti sostanziali. Questa nuova versione, è stata ispirata dal rivoluzionario documento di ricerca di OpenAI relativo a InstructGPT [82]. InstructGPT è un modello linguistico appositamente addestrato per seguire istruzioni ed eseguire compiti complessi. Tale modello è fondato sulla famiglia di modelli Pythia [83] di EleutherAI [84], una serie di modelli di intelligenza artificiale specializzati nel trattamento del linguaggio naturale. Si basa su 12 miliardi di parametri, forniti da EleutherAI ed addestrato sul set di dati, generato da umani, chiamato **Databricks Dolly 15k** al fine di sviluppare capacità simili nell’interazione con le istruzioni. Questo set di dati rappresenta un’innovazione fondamentale nel campo, essendo il primo di tale genere, *open-source* e creato manualmente, mirato a potenziare le capacità interattive dei modelli linguistici di ampia portata ed autorizzato sia per uso di ricerca che commerciale, grazie alla natura delle sue coppie di domande e risposte, generate tramite il *crowdsourcing* dai dipendenti Databricks. Questo significa che può utilizzarlo essere utilizzato da tutti per creare applicazioni interattive senza pagare l’accesso all’API REST o condividere dati con terze parti.

EleutherAI è una organizzazione no-profit dedicata alla ricerca nell’Intelligenza Artificiale. Pythia rappresenta una serie di 16 LLM che si distinguono per la loro capacità

di elaborare il linguaggio naturale in modi specializzati. I modelli inclusi nella serie di **Pythia** sono stati addestrati con dati pubblici seguendo un approccio sistematico, concepito per agevolare la ricerca scientifica. È importante sottolineare che **Pythia** si distingue come unica suite di modelli accessibili al pubblico che condivide la specificità di essere stati addestrati sugli stessi dati, secondo lo stesso ordine. Questa caratteristica unica offre un vantaggio significativo, in quanto permette agli utenti di confrontare e sperimentare con modelli che coprono una vasta gamma di dimensioni, da 70 milioni a 12 miliardi di parametri. Il *team* di EleutherAI ha reso disponibili 154 *checkpoint* per ciascuno dei 16 modelli di **Pythia**. I *checkpoint* rappresentano punti di controllo nel processo di addestramento dei modelli, catturando lo stato del modello in momenti specifici durante il processo di apprendimento. Questa pratica consente agli sviluppatori ed ai ricercatori di riprendere l’addestramento da un punto specifico, evitando di dover ricominciare dall’inizio in caso di interruzioni o di dover sperimentare con diverse configurazioni di modello senza perdere il progresso fatto fino a quel momento. Inoltre, i *checkpoint* possono essere utilizzati per valutare le prestazioni del modello su dati di test o per distribuire il modello addestrato per utilizzi pratici. In sostanza, rappresentano dei “salvataggi” del modello in diversi stati durante il processo di apprendimento. Tutti gli strumenti necessari per scaricare e riprodurre il processo di addestramento sono liberamente accessibili al pubblico. Questa trasparenza e disponibilità dei dati e degli strumenti sono fondamentali per agevolare ulteriori ricerche nel campo dell’intelligenza artificiale.

L’architettura di **Pythia** integra le migliori pratiche e gli avanzamenti recenti nel campo del linguaggio su larga scala. Basandosi ampiamente sul lavoro di Brown et al. [85], **Pythia** presenta alcune deviazioni significative che ne migliorano l’efficacia e l’efficienza. **Pythia** adotta esclusivamente strati completamente densi. Questo approccio, supportato dai risultati di ricerca successivi, si è dimostrato efficace per migliorare le prestazioni dei modelli. Durante il processo di addestramento, **Pythia** implementa Flash Attention [86], una tecnica progettata per ottimizzare la quantità di dati elaborati dai dispositivi. Questo contribuisce ad una maggiore efficienza nell’addestramento dei modelli, consentendo di ridurre i tempi di elaborazione e di sfruttare al meglio le risorse computazionali disponibili. Un altro elemento chiave dell’architettura di **Pythia** sono gli *embedding* rotativi, ovvero *embedding* che incorporano una componente rotativa che permette al modello di catturare informazioni sull’ordine delle parole introducendo una componente dinamica che evolve nel tempo durante il processo di apprendimento del modello. Tali componenti sono diventati una scelta preferita per **Pythia** grazie alla loro efficacia ed alla loro ampia adozione nella comunità della ricerca. **Pythia** utilizza anche tecniche di parallelizzazione dell’attenzione ed un approccio *feedforward* al fine di migliorare l’efficienza dell’addestramento senza compromettere le prestazioni complessive dei modelli. Infine, **Pythia** utilizza matrici di *embedding*/smontaggio non vincolate, ovvero l’operazione di *embedding* e la sua operazione inversa, chiamata “smontaggio”, non sono necessariamente la stessa. Questo significa che l’*embedding* di un *token* ed il suo corrispondente smontaggio possono rappresentare informazioni diverse, semplificando la ricerca sull’interpretazione dei modelli.

L’evoluzione da **Dolly** 1.0 a **Dolly** 2.0 rappresenta un chiaro impegno da parte di Databricks nel fornire soluzioni più avanzate ed accessibili nel campo dei modelli linguistici

di grandi dimensioni. Questo progresso non solo amplia le possibilità di interazione con tali modelli, ma apre anche nuove opportunità per il loro utilizzo commerciale, consentendo ad un'ampia gamma di organizzazioni di sfruttare completamente il potenziale delle tecnologie basate sull'Intelligenza Artificiale. A differenza dei modelli GPT, attualmente Dolly 2.0 è in grado, tuttavia, di fornire risposte solamente in lingua inglese.

Databricks ha reso disponibili varie iterazioni *open source* della libreria Dolly 2.0 tramite la piattaforma Hugging Face, come mostrato nella **Figura 3.3**. Hugging Face è una piattaforma *open source* dedicata al Deep Learning ed all'intelligenza artificiale, nota per offrire una vasta gamma di risorse e strumenti per lo sviluppo, l'addestramento e la distribuzione di modelli di linguaggio naturale ed altre applicazioni di Intelligenza Artificiale.



Figura 3.3. Elenco dei modelli Dolly disponibili su Hugging Face.
Fonte: [Hugging Face - Databricks](#)

All'interno di questo progetto si è scelto di utilizzare il modello **databricks/dolly-v2-7b** per motivi computazionali dell'ambiente di sviluppo a disposizione. Tale modello è un LLM con 6,9 miliardi di parametri derivato da **Pythia-6.9b** di **EleutherAI**. Esso, inoltre, è stato sottoposto a *fine-tuning* su un corpus di istruzioni di circa 15.000 record generato da dipendenti di Databricks.

3.2.3 LLAMA

La famiglia di modelli linguistici di grandi dimensioni (LLM) denominata **LLAMA** (Language Learning Multi-Agent) è stata sviluppata dalla società madre del noto *social network* Facebook, ovvero Meta, in collaborazione con Microsoft. Nel febbraio 2023, Meta ha introdotto **LLAMA 1** in risposta ai modelli linguistici proposti da OpenAI e Google. **LLAMA 1** è stato addestrato utilizzando un vasto set di dati composto da testo e codice contenente 1 trilione di *token*. Questo livello di esposizione ai dati consente al modello di produrre testi complessi, beneficiando, anche, di una lunghezza di contesto pari a 2.048 *token*, ovvero il numero di *token* considerati prima di generare un *output*. L'adozione di una lunghezza di contesto più estesa permette al modello di integrare una maggiore quantità di informazioni durante la generazione del testo, il che potenzialmente migliora la coerenza e la coesione delle risposte. **LLAMA 1** è stato sviluppato con un'attenzione particolare all'efficienza, cercando di mantenere elevate prestazioni con minori risorse computazionali

rispetto a modelli più grandi come GPT-3.

Inizialmente disponibile solo tramite una licenza non commerciale a scopo di ricerca, la crescente domanda di accesso a LLAMA 1 ha portato Meta a sviluppare il suo successore, LLAMA 2 [4], rilasciato nel luglio 2023 con una licenza commerciale. LLAMA 2 è stato progettato per raggiungere prestazioni paragonabili a quelle di GPT-3.5 su una varietà di *benchmark* accademici, pur distinguendosi nettamente da GPT-4, soprattutto nei compiti di codifica. Un vantaggio significativo di LLAMA 2 risiede nella sua capacità di ridurre in modo significativo i punteggi di violazione della sicurezza rispetto ai suoi concorrenti, rappresentando quindi un notevole progresso nell'aderire alle linee guida etiche relative ai modelli linguistici di ampia diffusione.

Per addestrare il modello LLaMA 2, viene utilizzata un'architettura autoregressiva basata sul trasformatore *decoder-only*, simile ai suoi predecessori, questo significa che i dati di *input* vengono alimentati direttamente nel decodificatore senza essere trasformati in una rappresentazione più alta ed astratta da un codificatore. Tuttavia, viene implementato un *layer* aggiuntivo di complessità attraverso l'apprendimento per rinforzo con *feedback* umano (RLHF), al fine di migliorare l'allineamento del modello con il comportamento e le preferenze umane. Nonostante sia un approccio computazionalmente costoso, è essenziale per migliorare la sicurezza e l'efficacia complessiva del modello. L'innovazione principale di LLaMA 2 risiede nel suo regime di pre-addestramento. Pur basandosi sul lavoro del suo predecessore, LLaMA 1, introduce diversi miglioramenti significativi per potenziarne le prestazioni. In particolare, vi è un aumento del 40% del numero totale di *token* addestrati ed una duplice espansione della lunghezza del contesto. Questo ampliamento del *dataset* ha consentito ai modelli LLaMA 2 di essere esposti ad una gamma più ampia di dati, permettendo loro di generare testi più complessi ed informativi. Il *corpus* di addestramento di LLaMA 2 comprende circa due trilioni di *token* di dati provenienti da differenti fonti disponibili pubblicamente, escludendo dati relativi a prodotti o servizi di Meta, come dichiarato dalla società stessa. Inoltre, questi modelli, rispetto alla versione precedente, includono una maggiore lunghezza del contesto, passando da 2048 a 4096 *token*, il doppio rispetto a quella dei modelli LLaMA 1. Questo significa che i modelli LLaMA 2 sono in grado di comprendere ed elaborare porzioni di testo più estese, rendendoli particolarmente utili per compiti come rispondere a domande specifiche e tradurre testi. In sintesi, LLaMA 2 rappresenta un notevole avanzamento rispetto al suo predecessore LLaMA 1, mostrando *performance* superiori in quasi tutti gli ambiti.

Con l'ultima iterazione dei modelli LLaMA, si è assistito a significativi miglioramenti che hanno ampliato le capacità e le prestazioni dei modelli stessi, mantenendo la solida base dell'architettura del trasformatore di Google, su cui si fonda il modello LLaMA originale. Tra le innovazioni più rilevanti, emerge l'adozione della pre-normalizzazione RMSNorm [87], un'evoluzione ispirata alla struttura di GPT-3. La RMSNorm, o Root Mean Square Normalization, rappresenta una tecnica di normalizzazione che opera sui singoli vettori di *input* di un *layer*. Contrariamente alla normalizzazione *layer-wise* tradizionale, in cui i dati in ingresso di ciascun *layer* vengono normalizzati per ridurre la varianza e regolarizzare il processo di apprendimento, RMSNorm normalizza ciascun vettore utilizzando la radice quadrata della media dei quadrati dei valori, mantenendo inalterato il segno. Que-

sto approccio garantisce una regolarizzazione più efficace del processo di apprendimento, promuovendo la stabilità e la convergenza dei modelli neurali durante la fase di addestramento. In aggiunta, è stata implementata la funzione di attivazione SwiGLU, tratta dall’architettura PaLM (*Pattern-Label Memory*), LLM sviluppato da Google. La SwiGLU, derivata dalla combinazione di Swish e GLU (Gated Linear Unit), offre un’elevata capacità di apprendimento grazie alla sua natura non lineare ed alla capacità di catturare relazioni complesse tra i dati. Questa funzione di attivazione si è dimostrata particolarmente efficace nell’incrementare le *performance* dei modelli LLaMA, contribuendo alla creazione di rappresentazioni più ricche ed informative dei dati. Un ulteriore avanzamento è stato rappresentato dall’introduzione degli Rotary Position Embedding (RoPE) [88]. Questi incorporamenti, differentemente dai tradizionali incorporamenti posizionali statici, sono in grado di variare dinamicamente la loro rappresentazione in base al contesto circostante, consentendo ai modelli di catturare informazioni più ricche riguardo alla posizione relativa delle parole all’interno delle sequenze di *input*. Questa flessibilità nell’*encoding* delle informazioni spaziali contribuisce significativamente alla capacità di comprensione e generazione dei modelli LLaMA. Infine, l’addestramento dei modelli LLaMA 2 ha visto l’impiego dell’ottimizzatore AdamW [89]. AdamW, un’ulteriore evoluzione dell’algoritmo di ottimizzazione Adam, incorpora una regolarizzazione *weight decay* direttamente nel processo di aggiornamento dei pesi del modello aggiungendo un termine supplementare alla funzione di perdita del modello, che dipende dalla norma dei pesi. Questo permette di controllare meglio la crescita dei pesi durante l’addestramento, riducendo il rischio di *overfitting* e migliorando la capacità di generalizzazione dei modelli neurali.

Infine, un’ulteriore e significativa innovazione introdotta è stata, anche, l’adozione del nuovo metodo di attenzione denominato “Grouped-Query Attention” (GQA), invece dell’attenzione alle “Multi-Query Attention” (MQA) [90]. Il metodo MQA si riferisce all’attenzione *multi-query*, in cui il modello considera più *query* contemporaneamente durante il processo di elaborazione. Questo metodo consente ai modelli di condurre inferenze con una notevole celerità rispetto ai precedenti modelli LLaMA 1, nonostante la maggiore complessità e dimensione dei modelli LLaMA 2. La tecnica GQA ottimizza il processo di attenzione all’interno di un modello neurale, gestendo in modo efficiente e simultaneo gruppi di *query*, favorendo una rapida elaborazione dei dati e delle relazioni presenti nel contesto dell’applicazione considerata. Questa innovazione riveste particolare importanza nell’ambito dell’Intelligenza Artificiale, poiché migliora significativamente le prestazioni dei modelli in termini di velocità e scalabilità, rendendoli adatti ad affrontare sfide computazionali di vasta portata. LLaMA 2-Chat, inoltre, è stato ottimizzato rigorosamente utilizzando sia il Supervised Fine-Tuning (SFT) che il Reinforcement Learning with Human Feedback (RLHF). Nella SFT, il modello viene addestrato su un set di dati etichettato, mentre nel RLHF, l’ottimizzazione è guidata dal *feedback* umano. Nel processo di SFT, il modello pre-addestrato viene esposto ad un set di dati etichettato, dove gli algoritmi di apprendimento supervisionato aggiornano i pesi del modello in base ai gradienti calcolati da una specifica funzione di perdita. Questa ottimizzazione consente al modello di catturare i modelli complessi e le sfumature presenti nel set di dati etichettato, specializzandosi nel compito *target* con elevata precisione. Successivamente, l’apprendimento per rinforzo viene utilizzato per allineare ulteriormente il comportamento del modello con le preferenze umane. In questa fase, vengono impiegate tecniche come il Importance Sampling [91],

ovvero un metodo Monte Carlo per valutare le proprietà di una particolare distribuzione, e la Proximal Policy Optimization [92], un algoritmo nel campo dell'apprendimento per rinforzo che addestra la funzione decisionale di un agente informatico per portare a termine compiti difficili, per introdurre rumore algoritmico, permettendo al modello di esplorare soluzioni alternative e di evitare minimi locali sub-ottimali. Questa iterazione continua non solo migliora il modello, ma allinea anche i suoi risultati alle aspettative umane. LLaMA 2-Chat raccoglie dati sulle preferenze umane attraverso un protocollo di confronto binario, che informa i modelli di ricompensa utilizzati per ottimizzare il modello di Intelligenza Artificiale conversazionale.

Anche in questo caso, come per l'utilizzo di Dolly 2, all'interno di questo lavoro di tesi è stato possibile utilizzare un modello LLAMA 2 tramite la piattaforma Hugging Face. All'interno della **Figura 3.4** è possibile verificare quali siano i modelli messi a disposizione.

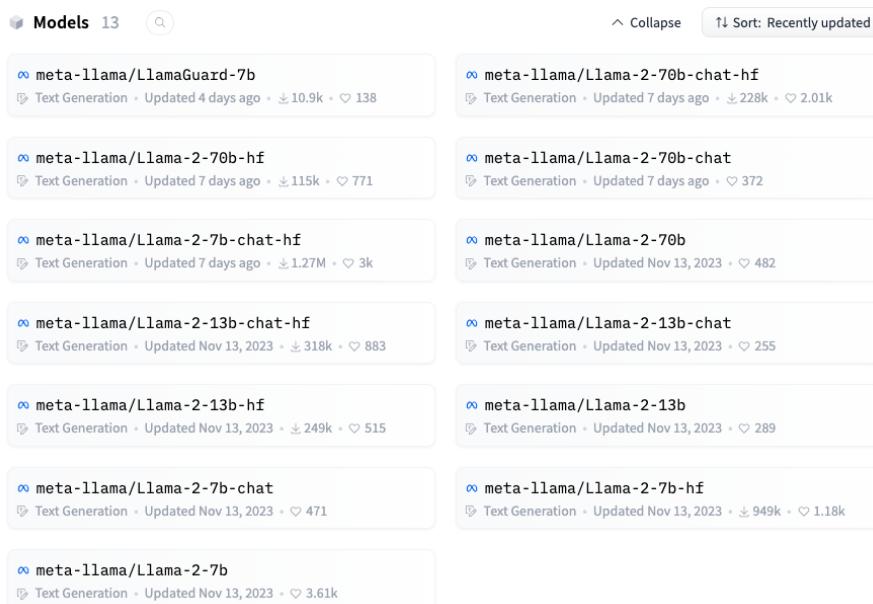


Figura 3.4. Elenco dei modelli LLAMA2 disponibili su Hugging Face.

Fonte: [Hugging Face - Databricks](#)

Pertanto, è possibile vedere come Llama 2 rappresenti una collezione di modelli di testo generativi preaddestrati ed affinati, caratterizzati da una gamma di parametri che varia da 7 miliardi a 13 miliardi fino a raggiungere i 70 miliardi. I modelli LLM ottimizzati per il dialogo, denominati Llama-2-Chat, sono stati sviluppati con l'obiettivo di eccellere nei contesti di utilizzo in conversazioni. Nel corso delle valutazioni effettuate, i modelli Llama-2-Chat hanno dimostrato prestazioni superiori rispetto ai modelli di *chat open-source* in numerosi *benchmark*. Inoltre, le valutazioni umane effettuate per valutare l'utilità e la sicurezza dei modelli hanno posizionato i Llama-2-Chat ad un livello paragonabile a quello di alcuni rinomati modelli *closed-source*, come ChatGPT e PaLM. I modelli indicati, invece, con "hf" sono in riferimento alla conversione nel formato Hugging Face Transformers.

All'interno di questo progetto, data la capacità computazionale dell'ambiente di sviluppo a disposizione, si è deciso di utilizzare il modello `meta-llama/Llama-2-7b-chat-hf`.

3.2.4 Gemini

Un anno dopo l'esordio di **ChatGPT**, il quale ha segnato l'avvio di una corsa all'avanguardia nell'ambito dell'Intelligenza Artificiale, Google ha presentato un progetto mirato a consolidare la propria posizione di *leadership* mondiale nel settore. Questo progetto ha preso forma nel Marzo 2023 con il lancio del modello **Bard**, il primo esperimento della compagnia nel campo della Generative AI. **Bard** è stato concepito non come un sostituto, ma come complemento alla ricerca Google. Gli utenti possono accedere direttamente al motore di ricerca di Google attraverso l'interfaccia di **Bard**, permettendo loro di approfondire ulteriormente le risposte fornite dal modello. Inoltre, **Bard** è in grado di elaborare *query* in linguaggio naturale, consentendo agli utenti di formulare domande simulando l'interazione con un essere umano. Questo aspetto rende l'esperienza di ricerca più intuitiva e *user-friendly*, facilitando l'accesso e l'assimilazione delle informazioni.

Dietro **Bard** si cela **LaMDA** [93], un *software* sviluppato da Google per la creazione di *chatbot*. Questo strumento, noto come “Language Models for Dialog Applications”, è stato progettato appositamente per gestire conversazioni in modo efficace e naturale. Per integrare **LaMDA** in **Bard**, Google ha scelto di utilizzare una versione ottimizzata del *software*, caratterizzata da una minore richiesta di potenza computazionale. Tale scelta consente una maggiore scalabilità del servizio, permettendo di assistere un numero più elevato di utenti e migliorare costantemente la qualità dell'esperienza, grazie al *feedback* continuo degli utenti. **Bard** si integrava anche con diversi servizi ed applicazioni Google, tra cui YouTube, Maps, Hotels, Flights, Gmail, Docs e Drive, consentendo agli utenti di applicare **Bard** ai propri contenuti personali.

Inizialmente Google ha concesso l'utilizzo del modello **Bard** solamente ad un ristretto gruppo di utenti al fine di testare lo strumento in maniera controllata. Successivamente, nel Maggio 2023, Google ha rimosso la lista di attesa e reso **Bard** disponibile tramite *browser* in oltre 180 paesi e territori, ma non l'utilizzo per gli sviluppatori tramite un'API ufficiale. Pertanto, **Bard** ha permesso di esplorare le molteplici potenzialità offerte dall'Intelligenza Artificiale nella generazione di testi, immagini e codice di programmazione, nonché nell'ottimizzazione della produttività e dell'efficienza dei compiti quotidiani.

Il giorno 8 Febbraio 2024, Google ha ufficialmente annunciato il nuovo nome del suo *chatbot* di intelligenza artificiale: **Gemini** [5]. Il *rebranding* in “Gemini” mira ad unificare l'intero ecosistema di Intelligenza Artificiale di Google sotto un'unica denominazione, abbandonando il precedente nome e qualsiasi eventuale connotazione negativa associata ad esso, soprattutto in riferimento alle restrizioni di sicurezza. In particolare, i ricercatori di Check Point Research (azienda israeliana produttrice di dispositivi di rete e *software*, specializzata in prodotti relativi alla sicurezza quali *firewall* e VPN), attraverso una serie di analisi attentamente condotte [94], sono riusciti a dimostrare che **Bard** può essere utilizzato per creare *email di phishing* tramite una richiesta non esplicitamente dichiarata, a differenza di GPT che risulta essere più moderato quando si tratta di contenuti malevoli.

Oltre alla versione base gratuita, basata sul modello chiamato **Gemini Pro**, Google ha introdotto un livello a pagamento al prezzo di \$20 al mese. Gli utenti che optano per questa opzione ottengono accesso a **Gemini Advanced**, il quale opera su **Gemini Ultra**, una versione più avanzata dei modelli di linguaggio **Gemini**. Tramite questo approccio, ovvero l'offerta di livelli di servizio diversificati, Google potrebbe rispondere alle critiche riguardo alla limitatezza delle funzionalità della versione gratuita, offrendo agli utenti un'opzione più completa ed avanzata a un costo accessibile. Attualmente, però, **Gemini Pro** è accessibile in più di 230 paesi e territori, supportando oltre 40 lingue, mentre **Gemini Advanced** è fruibile in 150 paesi.

Gemini, è uno strumento *chatbot* alimentato dal modello di linguaggio omonimo e progettato per simulare conversazioni umane utilizzando l'elaborazione del linguaggio naturale e l'apprendimento automatico. Oltre ad integrarsi con la ricerca di Google, **Gemini** può essere integrato in siti *web*, piattaforme di messaggistica o applicazioni per fornire risposte realistiche ed in linguaggio naturale alle domande degli utenti.

La prima versione di **Bard** utilizzava una versione più leggera del modello LaMDA che richiedeva meno potenza di calcolo per gestire un maggior numero di utenti simultanei. L'incorporamento del modello di linguaggio PaLM 2 ha permesso a **Bard** di essere più visuale nelle risposte alle domande degli utenti. **Bard** includeva anche Google Lens, che consentiva agli utenti di caricare immagini oltre alle richieste scritte. L'incorporamento del modello di linguaggio **Gemini** ha consentito ragionamenti, pianificazioni e comprensioni più avanzati. **Gemini Pro** aggiunto a **Bard** nel Febbraio 2024, aggiunge funzionalità migliorate in oltre 40 lingue. **Bard** utilizzava anche il modello **Imagen 2**, che conferiva allo strumento capacità di generazione di immagini.

I modelli **Gemini** si basano sui *decoder Transformer*, ma presentano miglioramenti specifici per garantire un addestramento stabile su larga scala e per ottimizzare l'inferenza sulle unità di elaborazione tensore di Google. Questi modelli sono stati progettati per gestire un contesto più ampio, fino a 32k, sfruttando meccanismi di attenzione efficienti come l'attenzione *multi-query*. La loro architettura è concepita per gestire *input* testuali arricchiti con una vasta gamma di dati audio e visivi, compresi immagini naturali, grafici, *screenshot*, PDF e video, consentendo di produrre *output* sia testuali che visivi. Il set di dati utilizzato per il pre-addestramento è diversificato ed include fonti come documenti Web, libri e codice, oltre a dati relativi a immagini, audio e video. La tokenizzazione avviene mediante l'utilizzo del tokenizzatore SentencePiece, libreria *open-source* per la tokenizzazione di testo sviluppata da Google. Inoltre, vengono applicati filtri di qualità ai dati, sia mediante l'impiego di regole euristiche che attraverso l'utilizzo di classificatori basati su modelli. Infine, per garantire l'integrità dei dati, viene eseguito un filtro di sicurezza per rimuovere eventuali contenuti dannosi ed i set di valutazione vengono accuratamente separati dal corpus di addestramento.

Gli utenti di **Gemini**, inoltre, devono soddisfare specifici requisiti per poter interagire col modello: devono avere almeno 18 anni e possedere un profilo Google personale. Le restrizioni sull'età per l'utilizzo dell'applicazione *web* di **Gemini** variano a seconda della

regione di utilizzo. In Europa, ad esempio, gli utenti devono essere maggiorenni mentre in altri paesi l'età minima è di 13 anni, salvo diversa disposizione di legge. È importante notare, anche, che gli utenti di età inferiore ai 18 anni possono interagire con **Gemini** solamente in lingua inglese.

Attualmente, **Gemini** offre un'API REST gratuita ma non disponibile per le nazioni europee. Come alternativa, Google ha reso disponibile il servizio Vertex AI, una piattaforma di Intelligenza Artificiale fornita da Google Cloud. Questa piattaforma offre un insieme di strumenti e servizi per lo sviluppo, l'addestramento ed il *deployment* di modelli di Machine Learning in modo semplice e scalabile. È possibile usufruire di una prova gratuita con un credito di 300 \$, dopodiché è possibile proseguire con l'utilizzo a pagamento, secondo i prezzi indicati sulla piattaforma, reperibili al seguente link: cloud.google.com/vertex-ai/pricing.

Nel contesto di questo progetto, poiché lo sviluppo del codice è iniziato prima del rilascio dell'API REST per l'utilizzo da codice del modello **Gemini**, si è adottato un metodo non ufficiale proposto da sviluppatori terzi [95]. Questo metodo consiste nell'utilizzo di *cookie* relativi alla pagina *web* del *chatbot* per interrogarlo all'interno del codice. Nella sottosezione 4.3.4 vengono forniti maggiori dettagli relativi all'implementazione di tale approccio.

4. Implementazione

All'interno del seguente capitolo, viene fornita un'analisi dettagliata relativa all'implementazione del progetto, descrivendo le diverse fasi che ne compongono il processo, come illustrato nella **Figura 4.1**.

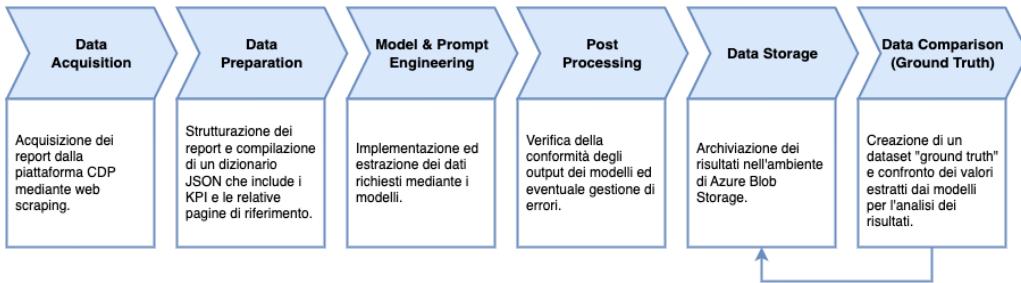


Figura 4.1. Rappresentazione schematica delle diverse fasi coinvolte nel processo di implementazione del progetto.

Si procede con la prima fase, denominata “Data Acquisition”, in cui viene dettagliato il metodo di estrazione dei *report* dalla piattaforma *Carbon Disclosure Project* ([74]) e la definizione dei KPI necessari per guidare l'estrazione delle informazioni da parte dei modelli. Successivamente, si procede con la fase di “Data Preparation”, in cui si descrive il processo di suddivisione dei *report* in singole pagine e la creazione di un dizionario in formato JSON, contenente le informazioni relative ai KPI (contesto e nome della metrica di riferimento) e le pagine corrispondenti. Tale algoritmo di indicizzazione mira a focalizzare l'attenzione dei modelli esclusivamente su porzioni specifiche di testo anziché fornire l'intero *report*, al fine di ottimizzare l'elaborazione e ridurre i relativi costi. Si prosegue con la fase di “Model and Prompt Engineering”, in cui si espone il processo di interrogazione di ciascun modello per l'estrazione dei KPI dal testo, relativo alle pagine specificate nella fase precedente. Inoltre, si forniscono dettagli relativi alla struttura dei *prompt*, utilizzati per guidare i modelli durante l'elaborazione, ed ulteriori aspetti tecnici relativi all'implementazione dei singoli modelli. Successivamente, si affronta la fase di “Post Processing”, in cui gli *output* forniti nella precedente fase vengono esaminati e corretti al fine di ottenere solamente dizionari in formato JSON ed eliminando eventuali commenti verbali aggiunti dai modelli. Si effettua anche una revisione dei valori estratti per garantirne la coerenza (ad esempio, rimuovendo i duplicati, controllando la consistenza dei dati, ed altro), con lo scopo di agevolare una successiva analisi dei risultati in maniera efficiente. Gli *output* di ciascun modello vengono, quindi, inseriti all'interno di un DataFrame Pandas. Successivamente, tramite la fase di “Data Storage”, ciascun DataFrame, risultante dall'estrazione dei modelli, viene memorizzato in formato CSV in una locazione appositamente designata nel Azure Blob Storage. Tale fase è seguita, infine, dalla procedura di “Data Comparison”, durante la quale viene generato un insieme di dati valido come “ground truth”, che rappresenta i valori reali e corretti presenti all'interno dei *report*. Si procede, quindi, con la creazione, ed il successivo salvataggio all'interno del Azure Blob Storage, di un *dataset* complessivo in cui i valori estratti vengono confrontati con i valori reali al fine di condurre le analisi conclusive.

4.1 Data Acquisition

Come già illustrato nel capitolo di Introduzione, il presente lavoro è stato concepito e realizzato nel contesto dello svolgimento del tirocinio curricolare presso l’azienda Power Reply Srl [6], società del gruppo Reply focalizzata sul mercato “Energy & Utilities”. Il progetto si è focalizzato sull’analisi e lo sviluppo di soluzioni mirate a soddisfare le necessità specifiche di un cliente aziendale. A tal fine, i KPI e le fonti dei dati analizzati sono stati definiti in stretta collaborazione con la committenza, al fine di garantire un allineamento con le esigenze.

In particolare, è stato sviluppato un *dataset* che include i KPI necessari da identificare all’interno dei *report*, con le seguenti caratteristiche:

- **context**: descrive il contesto specifico a cui il KPI fa riferimento. Questo attributo è generalmente indicato nei titoli delle sezioni all’interno dei *report* ed è essenziale per poter indicare al modello la metrica corretta, poiché all’interno del testo possono coesistere diverse metriche con lo stesso nome, rendendo necessaria una chiara distinzione;
- **kpi_name**: rappresenta la denominazione del KPI;
- **measure_unit**: indica la scala di misura utilizzata per quantificare il KPI.

I valori da ricercare all’interno dei documenti sono i seguenti:

Context name	KPI name	Measure unit
Reporting year	Gross global Scope 1 emissions (metric tons CO2e)	Metric tons CO2e
Reporting year	Scope 2, location-based	Metric tons CO2e
Reporting year	Scope 2, market-based (if applicable)	Metric tons CO2e
Purchased goods and services	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Capital goods	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Fuel-and-energy-related activities (not included in Scope 1 or 2)	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Upstream transportation and distribution	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Waste generated in operations	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Business travel	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Employee commuting	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e

Upstream leased assets	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Downstream transportation and distribution	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Processing of sold products	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Use of sold products	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
End of life treatment of sold products	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Downstream leased assets	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Franchises	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Investments	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Other (upstream)	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Other (downstream)	Emissions in reporting year (metric tons CO2e)	Metric tons CO2e
Non-fuel energy consumption in the reporting year	Consumption of purchased electricity (MWh)	MWh
Non-fuel energy consumption in the reporting year	Consumption of self-generated electricity (MWh)	MWh
Non-fuel energy consumption in the reporting year	Consumption of purchased heat, steam, and cooling (MWh)	MWh
Non-fuel energy consumption in the reporting year	Consumption of self-generated heat, steam, and cooling (MWh)	MWh
Non-fuel energy consumption in the reporting year	Total non-fuel energy consumption (MWh) [Auto-calculated]	MWh
Engagement & incentivization (changing supplier behavior)	% of suppliers by number	%
Engagement & incentivization (changing supplier behavior)	% total procurement spend (direct and indirect)	%
Engagement & incentivization (changing supplier behavior)	% of supplier-related Scope 3 emissions as reported in C6.5	%
Details of the climate-related requirements	% suppliers by procurement spend that have to comply with this climate-related requirement	%
Details of the climate-related requirements	% suppliers by procurement spend in compliance with this climate-related requirement	%

Poiché il progetto mira ad analizzare i *report* ESG pubblicati negli ultimi tre anni, è stato necessario apportare delle modifiche al *dataset* iniziale a causa di variazioni lessicali all'in-

terno della denominazione del contesto e del nome del KPI necessari per la ricerca. Tali modifiche includono ad esempio il cambiamento del KPI denominato “Emissions in reporting year (metric tons CO2e)” in “Metric tonnes CO2e” per i *report* pubblicati nell’anno 2021 oppure l’esclusione dei KPI nominati “Consumption of self-generated heat, steam, and cooling (MWh)” e “Total non-fuel energy consumption (MWh) [Auto-calculated]” per i *report* nel 2022 poiché non presenti. Queste variazioni sono state necessarie al fine di rendere più efficiente il processo di ricerca effettuato dai modelli.

I *report* richiesti per lo svolgimento di questo progetto sono relativi alle seguenti compagnie del mondo *energy*: Italgas, Snam, Enagas, Naturgy Energy Group SA, A2A, ENEL SpA, Iren SpA, ACEA SpA, Korea Gas Corp, Nippon Gas Co Ltd, EDF, Hera.

Considerato che la sorgente dati contenente tali *report*, ovvero la piattaforma CDP, non offre un servizio API REST per poter scaricare in maniera diretta i dati, è stato necessario recuperare tali documenti tramite la tecnica di *web scraping* (processo automatizzato di estrazione e raccolta di dati da pagine *web*). È importante evidenziare che tale procedimento è stato eseguito in conformità con le politiche della piattaforma, consultabili nella sezione “Terms of website use” del sito della piattaforma CDP [74], per garantire la conformità normativa e legale. Di conseguenza, è stato sviluppato uno *script* all’interno dell’ambiente Azure Databricks al fine di implementare un sistema di automazione avanzato per la navigazione *web* e la raccolta di dati dal sito. Utilizzando il modulo **Selenium WebDriver** in Python, il programma simula l’interazione di un utente con il *browser* Chrome per eseguire una sequenza di compiti specifici. I passaggi che vengono eseguiti all’interno di questo programma sono:

- **Configurazione delle opzioni del *browser*.** In questa fase, vengono specificati i dettagli di configurazione del *browser*. In particolare, si definisce la cartella predefinita per il salvataggio dei *file*, si attiva la modalità *headless* per eseguire il *browser* senza una finestra grafica, ottimizzando così l’esecuzione per test automatizzati. Inoltre, si disabilita l’utilizzo della memoria condivisa, comunemente impiegata da Chrome per gestire grafica e *rendering*. Questa disattivazione può risultare vantaggiosa in contesti di limitata memoria condivisa o per ridurre il consumo complessivo. Si procede anche con la disattivazione del *Sandbox* di sicurezza di Chrome, il quale limita l’accesso dei processi *browser* alle risorse del sistema. Tale disattivazione potrebbe essere necessaria per eseguire test automatizzati in ambienti controllati. Viene effettuato, inoltre, il *download* di un *file* dall’archivio Azure Blob Storage per consentire l’utilizzo di un’estensione specifica nel *browser* Chrome chiamata **Anti-Captcha Blocker Extension**. Questo passaggio è fondamentale per evitare il blocco del processo di scaricamento automatico dei documenti. Infine, si definisce il formato grafico in cui devono essere salvati i documenti, ad esempio impostando il formato della pagina su A4 ed i margini con uno spessore nullo.
- **Accesso alla piattaforma CDP.** Attraverso l’utilizzo di Selenium WebDriver è possibile immettere le chiavi di accesso per accedere alla piattaforma CDP.
- **Gestione automatica dei *cookie*.** Poiché durante l’apertura automatica del *browser* possono comparire *cookie*, che richiedono l’accettazione obbligatoria per

continuare la navigazione, si implementa un comando in grado di selezionarli automaticamente.

- **Selezione dei documenti da scaricare.** Durante questa fase, vengono identificati i documenti da scaricare utilizzando la libreria python BeautifulSoup. Inizialmente, si impostano i filtri di ricerca, come illustrato nella **Figura 4.2**, specificando l’anno di riferimento, il programma di interesse (in questo studio viene selezionato il programma “Climate Change”, poiché la ricerca si concentra su un confronto nell’ambito ESG, come descritto nella sezione 2.2), ed impostando lo stato con “Submitted” per filtrare solo i documenti effettivamente pubblicati dalle aziende. Successivamente, viene selezionato il nome dell’azienda all’interno della barra di ricerca.

The screenshot shows a search interface for CDP responses. At the top, it says "Search and view past CDP responses". Below that is a search bar with placeholder text "Search for a city or company name" and a magnifying glass icon. Underneath the search bar is a note: "You can also refine your search". There are four dropdown filters: "Country/Area" set to "2023", "Year" set to "2023", "Program" set to "Climate Change", and "Status" set to "Submitted". Below the filters is a "Search" button and a "Next" button. The main area displays "12,460 results" in bold. A table follows, with columns: "Name", "Response", "Year", "Status", and "Score". The first few rows of the table are:

Name	Response	Year	Status	Score
(ACIP) Alexandria Company for Industrial Packages	Climate Change 2023	2023	Submitted	Not Available
(Sichuan) Tianqi Lithium Corporation	Climate Change 2023	2023	Submitted	B
1&1 AG	Climate Change 2023	2023	Submitted	D
1000mercis SA (holding of Numberly)	Climate Change 2023	2023	Submitted	Not Available
16 POINTS CONSULTING LLC	Climate Change 2023	2023	Submitted	Not Available

At the bottom of the table, there is a note: "Note: Not all companies requested to respond to CDP do so. Companies who are requested to disclose their data and fail to do so, or fail to provide sufficient information to CDP to be evaluated will receive an F. An F does not indicate a failure in environmental stewardship." Below the table are pagination controls: "Show 5 10 20" and page numbers "1 2 3 4 5 ... 2491 2492 >".

Figura 4.2. Schermata di esempio della piattaforma CDP [74] all’interno della quale vengono selezionati i *report* da scaricare in relazione al campo “Name” (in riferimento al nome dell’azienda), “Response” (in riferimento al programma CDP), “Year” (ovvero l’anno di pubblicazione).

- **Scaricamento tramite la funzionalità integrata di Chrome.** In questa fase conclusiva, vengono definiti dei controlli per accertare che il programma sia correttamente posizionato sulla pagina relativa al *report*, verificando che il numero di

pagine non sia nullo o abbia una numerosità troppo bassa. Se il controllo non viene superato, presumibilmente a causa di un errore di caricamento, il codice rigenera la pagina. Successivamente, utilizzando la funzionalità Chrome DevTools Protocol integrata nel *browser* Chrome, si procede alla stampa della pagina corrente all'interno di un file PDF, poiché la piattaforma CDP non dispone di un pulsante per scaricare in maniera diretta il documento. Il *file* viene quindi salvato all'interno dell'archivio Azure Blob Storage per l'archiviazione e la successiva consultazione. Durante questa fase, inoltre, viene impostato un *timeout* per gestire eventuali eccezioni nel caso ci siano problemi durante lo scaricamento.

4.2 Data Preparation

Dopo aver completato la fase di Data Acquisition, si procede con la fase di Data Preparation, durante la quale il dato viene adeguatamente trattato per essere utilizzato dai modelli in modo accurato e controllato.

In questa fase, i *report* sono stati organizzati in singole pagine e salvati all'interno del Azure Blob Storage come tali. Successivamente, è stata implementato un algoritmo di indicizzazione al fine di individuare all'interno di ciascuna pagina le stringhe relative al contesto ed al nome di ogni KPI. Qualora sia stato individuato sia il contesto che il nome del KPI nella stessa pagina (indicando all'interno del metodo che il contesto debba essere trovato prima del nome del KPI, essendo generalmente parte integrante del titolo della sezione), l'indice della pagina viene aggiunto ad un elenco. Vengono aggiunti alla lista anche gli indici delle pagine in cui si verifica che la posizione del contesto si trovi all'interno di una pagina ed il nome del KPI nella pagina successiva. In questo caso, si vuole tener conto anche delle eventuali eccezioni, pertanto, entrambe le pagine vengono annotate nell'elenco. Quest'operazione è stata eseguita al fine di ridurre i costi associati all'utilizzo di modelli a pagamento (GPT) e per non sovraccaricare il *cluster* di Azure Databricks, a causa del precedente caricamento di alcuni modelli come LLAMA e Dolly.

Una volta estratte le pagine rilevanti, vengono memorizzate all'interno di un dizionario JSON, in cui la chiave è rappresentata dall'unione del contesto ed il nome del KPI di riferimento ed il valore è costituito dalla lista di pagine correlate. Successivamente, il dizionario viene salvato come file JSON all'interno dello spazio di archiviazione Azure Blob Storage, in relazione a ciascun *report* di ogni azienda ed anno considerati.

4.3 Model and Prompt Engineering

All'interno della fase “Model and Prompt Engineering”, si procede con la formulazione di procedure standard da seguire per l'elaborazione di tutti i modelli e si prosegue con la descrizione relativa all'implementazione ed all'adattamento, della modalità precedentemente stabilita, per ogni singolo modello.

In primo luogo è stato definito che, al fine di specificare il contesto di ricerca e fornire una specifica guida relativamente a come i modelli debbano affrontare il compito

assegnato, venga inizialmente sottoposto un *prompt*, denominato `instruction_prompt`, strutturato come segue:

You are an expert in extracting sustainability Key Performance Indicators (KPI) from reports. I will provide you with some KPIs and your objective is to detect them from a given text, with name exactly as provided and extract their values.

You will output the detected KPIs with their values in a json format, or an empty json if no KPI among the provided ones is detected.

L'approccio discusso rappresenta un tentativo di avvicinamento al metodo noto come RAG, ovvero Retrieval Augmented Generation [96]. RAG rappresenta un approccio innovativo nell'ambito dell'Intelligenza Artificiale poiché integra due paradigmi tradizionalmente distinti: i modelli basati sul recupero e quelli generativi. Fondamentalmente, l'obiettivo è dotare un modello linguistico di grandi dimensioni (LLM) di un meccanismo per accedere dinamicamente ad un database esterno al fine di arricchire le sue capacità di generazione di testo. Questa integrazione non solo amplia il repertorio di conoscenze del modello ma trasforma radicalmente il suo processo decisionale durante le interazioni con le *query*. Il funzionamento di un sistema RAG può essere suddiviso in due fasi principali: la fase di recupero e la fase di generazione. Nella prima fase, il sistema esplora il database esterno per individuare informazioni rilevanti in risposta ad una *query* specifica. Questo database può variare da una raccolta curata di vettori di conoscenza a fonti più ampie e dinamiche come il *web*. Una volta ottenuti i dati pertinenti, nella fase di generazione il modello sintetizza queste informazioni insieme alla sua conoscenza pregressa per produrre risposte coerenti e contestualmente ricche. Uno dei principali vantaggi attribuiti a RAG è la sua capacità di ridurre le “allucinazioni”, ovvero la generazione di informazioni false o irrilevanti da parte dei modelli di Intelligenza Artificiale. Grazie al continuo aggiornamento e all’incorporazione di dati contestualizzati nel processo di generazione, i modelli RAG sono in grado di fornire risposte più accurate e ancorate alla realtà rispetto ai modelli generativi tradizionali. Nel contesto di questa ricerca, non è stato possibile implementare un database specifico per il dominio di interesse a causa delle limitazioni di tempo e risorse disponibili per la costruzione accurata di un tale database. Pertanto, è stato adottato un approccio generale, pur cercando di fornire indicazioni specifiche sul contesto di ricerca e le modalità di implementazione dei modelli.

Successivamente, è stato determinato che, per ogni *report* preso in esame, venga richiamato il file JSON, precedentemente archiviato durante la fase di Data Preparation (come descritto nella sezione 4.2), che contiene il dizionario indicante le pagine specifiche del *report* in cui è stata verificata la presenza del KPI necessario per l'estrazione. Dopo aver estratto il dizionario JSON, ogni KPI viene analizzato singolarmente in maniera iterativa. Per ognuno di essi, si estraе il contesto ed il nome relativo alla metrica da individuare e si procede con la costruzione della seguente stringa, aggiungendo i parametri mancanti utilizzando le informazioni presenti nella lista dei KPI e relativi alla singola esecuzione del programma:

```
kpi_list:  
  - Context Name: {context}; KPI name: {kpi_name}; Measure Unit: {measure};  
    Year: {reporting_year}
```

Successivamente, la stringa `kpi_list` è inserita nel *prompt*, denominato `user_input`, in cui vengono specificate le istruzioni per guidare il modello nell'estrazione corretta del valore del KPI considerato. All'interno del *prompt* viene anche incluso il testo completo delle pagine in cui è stato individuato il KPI tramite una stringa chiamata `extracted_text`. Inoltre, è stato definito che tutti i modelli devono generare in *output* una lista di dizionari in formato JSON al fine di agevolare la gestione e la creazione di un DataFrame Pandas finale contenente tutti gli *output* del modello effettuati durante l'estrazione. Di seguito è riportata la struttura di tale *input* testuale:

```
user_input:  
    I will provide you with individual pages extracted from sustainability  
    reports. You will search the content for all and only the KPIs listed  
    below in the "KPIs and metrics to search for" section structured as  
    follows:  
        - Context name: "context_name_example"; KPI name: "kpi_name_example";  
        Measure Unit: measure_unit_example; Year: year_example".
```

The string of the context is always before the name of the kpi.

Find in the input the value of the KPI contained in the below list "KPIs and Metrics to Search". Store the found KPI names exactly as they appear in the input and their value in a dictionary JSON-like with this structure:

```
[ {  
    "context": "context_name_example",  
    "kpi_name": "kpi_name_example",  
    "value": value_example,  
    "measure_unit": measure_unit_example,  
    "year": year_example  
, ... ]
```

If you find a different KPI not listed in "KPIs and Metrics to Search", do not add it to the json. Before adding a KPI to the json, check its value: all the values have to be numerical, otherwise ignore it.

KPIs and Metrics:

```
{kpi_list}
```

Input:

```
{pages_text}
```

Successivamente, viene fornito un secondo *prompt* (`check_input`) affinché il modello possa esaminare ulteriormente il testo, verificare la correttezza dell'*output* fornito nella fase precedente (`previous_response`) ed assicurarsi che il formato dell'*output* sia conforme

al dizionario JSON richiesto nelle istruzioni del *prompt*. Di seguito viene mostrata la struttura di questo *prompt*:

`check_input:`

You are now performing a refining phase.

Check that the output in json format contains in the "context" field the "Context name" value provided in the "KPI and metrics" list and in the "kpi_name" field the "KPI name" value provided in the "KPI and metrics" list.

If the output in json format is already correct, rewrite it.

`KPIs and Metrics:`

`{kpi_list}`

`Output to correct:`

`{previous_response}`

L'intera procedura di esecuzione di ciascun modello, per l'intera gamma di *report* disponibili, è stata automatizzata mediante l'implementazione di un orchestratore ("orchestrator"). Tale strumento, operante a livello di codice, coordina il passaggio dei parametri pertinenti (ovvero l'anno ed il nome della compagnia) a ciascun *notebook* dedicato all'elaborazione di un modello specifico, facilitandone l'attivazione e monitorando eventuali problematiche insorte durante l'esecuzione.

Nelle seguenti sotto-sezioni, verrà fornita un'analisi dettagliata dell'implementazione di ciascun modello, presente nella **Tabella 4.2**, e delle relative caratteristiche nel fornire una risposta conforme alle indicazioni fornite nei *prompt*.

Company	Modello	Numero di parametri	Context window	Dati di training
OpenAI	gpt-3.5-turbo-0125	175 miliardi	16k	Multi-genere da fonti pubbliche (Common Crawl, libri, Wikipedia)
Databricks	dolly-v2-7b	7 miliardi	2k	Coppie domande e risposta dai dipendenti Databricks
Meta	llama-2-7b-chat-hf	7 miliardi	4k	Materiali pubblici (Common Crawl, Wikipedia, libri da Project Gutenberg)
Google	gemini	100 miliardi	1 milione	Dati multimodali e multilingue dal Web, libri e dati interni

Tabella 4.2. Tabella riassuntiva delle caratteristiche dei modelli analizzati.

4.3.1 GPT3.5 Turbo

Per interagire con il modello GPT-3.5 Turbo, ospitato su Azure OpenAI, è stato necessario, in primo luogo, configurare le credenziali e le variabili di ambiente. Questa fase ha coinvolto l'acquisizione delle chiavi API per l'autenticazione e la definizione delle informazioni relative all'*endpoint*, il nome del modello e la versione, al fine di agevolare la comunicazione con l'API di OpenAI.

Successivamente, è stata sviluppato il metodo `get_completion` al fine di rendere più agile l'interazione col modello considerato. Tale metodo accetta in *input* messaggi di testo (*prompt*) e restituisce la risposta generata dal modello. Il processo prevede un massimo di cinque tentativi in caso di sovraccarico del *server* o di indisponibilità dello stesso. Nel caso di insuccesso dopo cinque tentativi, viene generata un'eccezione bloccando l'esecuzione del programma. Nel caso in cui si verifichi un errore di sovraccarico o di indisponibilità del *server*, il codice prova, fino a cinque volte, di ottenere una risposta con una pausa di 10 secondi tra ogni tentativo per evitare sovraccarichi ulteriori. Esauriti i tentativi, o nell'eventualità che venisse generata un'eccezione non prevista, il modello restituisce un messaggio di avviso contenente l'errore. Nel caso di completamento riuscito, viene fornita la risposta generata dal modello per la richiesta inviata.

I *prompt* vengono sottoposti al modello nel seguente modo:

```
messages=[{"role":"system", "content": instruction_prompt},
          {"role":"user", "content": user_input(kpi_list, extracted_text)}]
response_gpt = get_completion(messages=messages)
response = response_gpt.choices[0].message["content"]

messages_check =[{"role":"system", "content":instruction_prompt},
                  {"role":"user", "content":check_input(kpi_list, response)}]

time.sleep(2)
response_gpt_check = get_completion(messages=messages2)
response_check = response_gpt_check.choices[0].message["content"]
```

Le variabili `messages` e `messages_check` sono liste di dizionari che rappresentano messaggi strutturati con ruoli di sistema ed utente. Questo significa che ogni messaggio contiene sia l'*input* destinato al sistema, al fine di specificare il contesto di ricerca e le linee guida per il compito richiesto, che l'*input* fornito dall'utente contenente la richiesta da portare a termine. Questi messaggi sono strutturati come dizionari con chiavi “role” e “content”, che indicano rispettivamente il ruolo del messaggio (sistema o utente) ed il suo contenuto. In questo modo è stato possibile inserire il *prompt* `instruction_prompt` e, attraverso la costruzione di funzioni specifiche, sono stati sottomessi anche i *prompt* di `user_input` e `check_input` fornendo le informazioni necessarie per il completamento di questi testi. Dopodichè, le risposte generate dal modello (`response_gpt` e `response_gpt_check`) vengono salvate attraverso la chiamata al metodo `get_completion`, specificando i messaggi in riferimento. Dopo aver ottenuto le risposte dal modello, viene estratto il contenuto testuale dalla struttura dei dati all'interno di due variabili: `response` e `response_check`.

4.3.2 DollyV2-7b

Per poter usufruire dell’interazione col modello **DollyV2-7b**, è stato necessario configurare l’ambiente di utilizzo, poiché il modello viene caricato localmente. Inizialmente, è stata impostata una variabile d’ambiente specifica in modo da regolare la memoria allocata per l’utilizzo della GPU durante l’esecuzione del codice, in questo lavoro è stata impostata a 128 megabyte (MB). Dopodiché, viene eseguito l’accesso al portale Hugging Face, tramite *script*, ed al successivo caricamento del modello in questione tramite il percorso all’interno del portale, ovvero “`databricks/dolly-v2-7b`”. Successivamente, il codice procede al caricamento del modello pre-addestrato ed il *tokenizzatore* associato al fine di abilitare l’elaborazione e la generazione di testo basata su Transformers, con particolare attenzione alla gestione della memoria GPU ed alle configurazioni del modello durante il caricamento.

Per permettere l’interazione con il modello **DollyV2-7b**, è stata sviluppato un metodo denominato `generate_response`. Tale metodo, essenziale per il processo di generazione delle risposte, accetta come argomento una stringa rappresentante l’*input*, contenente la richiesta per la quale si intende ottenere una risposta. In aggiunta, richiede come parametri il modello neurale pre-addestrato ed i rispettivi *tokenizzatori*. La presenza dei *tokenizzatori* è essenziale per convertire la stringa di *input* in un formato comprensibile per il modello neurale, al fine di produrre una risposta coerente.

Lo *script* opera nel seguente modo:

- **Preparazione dell’*input***

Al metodo viene passato il *prompt* relativo all’istruzione che il modello deve eseguire (composta prima da `user_input` e poi da `check_input`, mostrate all’inizio della sezione 4.3), dopodiché il *prompt* viene formattato all’interno di una stringa denominata `PROMPT_FORMAT` insieme all’`instruction_prompt` (mostrato anch’esso all’inizio della sezione 4.3). Tale stringa presenta la seguente struttura:

```
PROMPT_FORMAT = """ Below is an instruction that describes a task.  
Write a response that appropriately completes the request.  
  
### Context:  
{instruction_prompt}  
  
### Instruction:  
{input}  
  
### Response:  
"""
```

La stringa `PROMPT_FORMAT`, viene quindi *tokenizzata* (processo di suddivisione della stringa in *token* più piccoli). I *token* risultanti vengono convertiti in tensori PyTorch, al fine di rappresentare l’*input* in un formato compatibile con il modello neurale, e spostati sulla GPU, se disponibile, per massimizzare le prestazioni del sistema durante la generazione delle risposte.

- **Generazione della risposta**

Utilizzando il modello neurale fornito, viene generata una sequenza di *token* che rappresentano la risposta alla richiesta. Questa generazione avviene mediante la chiamata al metodo `generate` del modello, specificando parametri come la massima lunghezza della sequenza di *token*, la probabilità cumulativa massima (`top_p`) ed altre opzioni per controllare il processo di generazione.

- **Identificazione della risposta generata**

La sequenza di *token* generata viene analizzata per identificare la porzione relativa alla risposta effettiva. Questa porzione inizia con un *token* speciale indicato da “### Response:” e termina con un *token* indicato da “### End”. Il metodo individua la posizione di questi *token* nella risposta generata e restituisce la porzione compresa tra di essi.

- **Gestione delle eccezioni**

Il codice gestisce eventuali eccezioni che potrebbero verificarsi durante il processo di generazione della risposta, in particolare il caso in cui la GPU esaurisca la memoria. In tal caso, il metodo prova a generare la risposta nuovamente, riducendo la massima lunghezza dei nuovi *token*.

Durante le prove di esecuzione del codice, è emerso che il modello in questione non supporta l'estrazione diretta di dizionari in formato JSON. Di conseguenza, è stato necessario modificare l'approccio, richiedendo al modello di estrarre esclusivamente il valore numerico del KPI desiderato. Successivamente, l'*output* ottenuto viene passato come argomento ad uno *script* il quale si occupa di generare un dizionario JSON valido, inserendo il valore estratto dal modello nel campo “`value`” del dizionario e popolando gli altri campi con i valori appropriati per tale estrazione.

4.3.3 LLAMA2-7b

Come per DollyV2-7b, l'implementazione di LLAMA richiede il caricamento in locale di oltre 7000 parametri, pertanto, è stato necessario stabilire un ambiente di esecuzione ottimizzato al fine di evitare sovraccarichi di memoria e potenziali errori durante la fase di compilazione del codice. A tal fine, sono state definite specifiche variabili di ambiente per la gestione delle risorse CUDA di PyTorch, che rappresentano la memoria utilizzata dalla GPU durante le operazioni di calcolo. In particolare, è stata configurata la dimensione massima di allocazione per ciascuno *split* della memoria CUDA a 128 megabyte (MB). Questo parametro riveste un'importanza fondamentale nel controllo dell'allocazione di memoria sulla GPU, consentendo di prevenire possibili errori di esaurimento della memoria e di ottimizzare le prestazioni del sistema in determinate circostanze. In aggiunta, è stata abilitata l'espansione dinamica dei segmenti di memoria CUDA. Questa funzionalità permette a PyTorch di regolare dinamicamente le dimensioni dei segmenti di memoria allocati sulla GPU durante l'esecuzione del programma, garantendo una gestione ottimale delle risorse ed una maggiore flessibilità nell'adattamento alle esigenze computazionali. Tale configurazione risulta particolarmente vantaggiosa per ottimizzare l'utilizzo complessivo della memoria sulla GPU e per garantire un'efficace gestione delle risorse disponibili. Successivamente, viene eseguita l'operazione di svuotamento della

cache CUDA al fine di liberare la memoria della GPU, garantendo così un ambiente di esecuzione ottimale per le operazioni successive.

Dopo aver configurato l’ambiente di lavoro, si procede specificando il percorso del modello preaddestrato LLaMA2 all’interno del *framework* Hugging Face. Successivamente, si configura il modello, adottando le impostazioni predefinite (fornite da Hugging Face) e stabilendo un limite massimo di nuovi *token* generati, fissato a 300 (poiché l’*output* richiesto non necessita di risposte complesse ma solamente di un *output* limitato alle dimensioni di dizionari JSON). Infine, si inizializza una pipeline per l’esecuzione delle operazioni di generazione di testo, impiegando il modello LLaMA2.

Dopodiché è stata definito un metodo (`construct_llama2_prompt`) in grado di costruire un *prompt* combinando i messaggi di dialogo (ovvero della conversazione tra un utente ed il sistema). All’interno di questo dialogo viene inserito il *prompt* predefinito del sistema come primo messaggio, successivamente viene unito il primo ed il secondo messaggio del dialogo utilizzando i delimitatori di *prompt* di sistema:

- **B_INST, E_INST**: Delimitatori per indicare l’inizio e la fine di un’istanza (istanza di dialogo).
- **B_SYS, E_SYS**: Delimitatori per indicare l’inizio e la fine del *prompt* di sistema.
- **BOS, EOS**: Delimitatori per indicare l’inizio e la fine della sequenza di *token* generata dal modello.

Pertanto, tale metodo costruisce il *prompt* utilizzando la storia della *chat*, inserendo i delimitatori di istanza tra ogni coppia di messaggi ed i delimitatori di inizio e fine della sequenza di *token* generata dal modello, restituendo il *prompt* costruito come una stringa.

La risposta del modello viene fornita attraverso un metodo chiamato `get_output`. In particolare, tale metodo costruisce un *prompt* utilizzando lo *script* precedente (`construct_llama2_prompt`) genera una risposta. Dopo di che, estrae la parte di testo generata che potrebbe contenere oggetti JSON, utilizzando espressioni regolari, e prova a deserializzare ogni stringa *json-like* estratta, aggiungendo, poi, i risultati validi ad una lista di dizionari. Qualora non venisse estratto nessun *output* in formato JSON valido, viene restituita una lista vuota. L’*output* di questo metodo viene passato ad un altro (denominato `extract_answer`) che estrae la risposta generata dal modello LLaMA2 dall’*output* fornito utilizzando il delimitatore `[/INST]`, utilizzato per dividere l’*output* ed ottenere solo la parte di risposta generata dal modello.

4.3.4 Gemini

Per l’utilizzo del modello **Gemini**, poiché lo sviluppo di tale progetto è stato precedente all’avvento di un’API per il contesto italiano ed europeo, come descritto in precedenza nella sottosezione 3.2.4, è stato necessario adottare un approccio non ufficiale [95] per interagire con questo modello. Tale approccio prevede l’utilizzo di *cookie* (piccoli frammenti di dati memorizzati sul dispositivo dell’utente durante l’interazione con un sito *web*) all’interno

del *browser* di **Gemini**, consentendo così l’interazione con il modello mediante *script* automatizzati. Tale soluzione è stata sviluppata da terze parti con l’obiettivo di agevolare i programmatore nella validazione di specifiche funzionalità del modello, compensando il ritardo nell’implementazione e nel rilascio di un’API ufficiale da parte di Google.

Il metodo descritto è reso disponibile attraverso la piattaforma GitHub. All’interno del *repository*, è definita una classe denominata **Bard**, originariamente progettata per interagire con il servizio Google Bard, successivamente aggiornata per il modello **Gemini** a Dicembre 2023 (mantenendo il nome “Bard”) ed infine rinominata come “Gemini” nel Febbraio 2024. Questa classe offre una serie di metodi per comunicare con il servizio di Intelligenza Artificiale ed ottenere risposte a domande in formato di testo o immagini. All’interno di questo lavoro, è stato necessario utilizzare principalmente le funzioni relative all’invio di richieste al servizio ed alla ricezione di risposte. Prima di inviare una richiesta al servizio di Generative AI, il metodo `get_answer` prepara i dati necessari, ad esempio costruendo il testo della domanda e la richiesta HTTP da inviare al servizio. Utilizzando la libreria `requests` di `python`, tale richiesta viene inviata tramite una chiamata POST all’*endpoint* appropriato del servizio. Dopo l’invio della richiesta, il metodo `get_answer` attende la risposta dal modello. Una volta ricevuta, viene elaborata per estrarre le informazioni rilevanti, come il contenuto della risposta, l’ID della conversazione, l’ID della risposta, eventuali *link* o immagini associate, ed altro ancora. Il codice gestisce anche eventuali errori che possono verificarsi durante l’invio delle richieste o l’elaborazione delle risposte, ad esempio i casi in cui la risposta del servizio non è disponibile o non può essere elaborata correttamente, restituendo messaggi di errore appropriati per informare l’utente del problema. Infine, i dati estratti vengono restituiti sottoforma di dizionario JSON.

Per interagire con il modello **Gemini** attraverso questo metodo, è stato necessario inizialmente clonare il *repository* GitHub che contiene l’implementazione dell’API di **Gemini**. Successivamente, è stato richiesta l’acquisizione e l’utilizzo dei *cookie* di autenticazione per il modello, i quali sono stati passati all’interno di un oggetto chiamato **BardCookies** utilizzando il modulo `bardapi`. Tali cookie sono contrassegnati come `HTTPOnly`, una caratteristica di sicurezza che impedisce la lettura o la manipolazione da parte di codice JavaScript lato *client*. Questo assicura che i *cookie* siano accessibili solo dal *server web* che li ha impostati e non possano essere alterati da *script* eseguiti nel *browser* dell’utente. I *cookie* utilizzati per interagire con **Gemini** sono denominati `__Secure-1PSID` e `__Secure-1PSIDTS`, tali valori vengono memorizzati in un dizionario JSON e passati come argomento all’oggetto **BardCookies** durante l’inizializzazione. Questo oggetto gestisce le informazioni dei *cookie* necessarie per l’autenticazione con il modello e permettere la successiva interazione.

Siccome l’utilizzo del modello **Gemini** attraverso il metodo non ufficiale, non prevede una particolare implementazione per comunicare al modello il contesto di utilizzo, l’`instruction_prompt`, descritto all’inizio della sezione 4.3, in questa implementazione è stato semplicemente introdotto all’interno dello `user_prompt` e `check_prompt` forniti al modello per eseguire le richieste.

4.4 Post Processing

La fase di Post Processing ha coinvolto la modellazione degli *output* in linea con i requisiti del progetto. Poiché è stato riscontrato che tutti i modelli, con frequenza variabile, non restituiscono *output* contenenti esclusivamente il dizionario JSON richiesto ma anche ulteriori risposte verbali, è stato necessario sviluppare un metodo dedicato, denominato `extract_json`, capace di ricevere in ingresso una stringa di testo e produrre in uscita una lista di dizionari JSON. In particolare, il metodo sostituisce eventuali singoli apici (') con doppi apici ("") per garantire che il testo sia formattato correttamente come JSON. Successivamente, ricerca tutte le occorrenze di stringhe JSON nel testo utilizzando l'espressione regolare `r'\{\[\^\}]*\}'`, che individua tutte le sottostringhe che iniziano con il simbolo { e terminano con }, indicando un possibile oggetto JSON. Queste sottostringhe vengono quindi raccolte in una lista chiamata `json_like_strings`, e per ciascuna stringa JSON individuata, lo *script* cerca di trasformarla in oggetto JSON utilizzando la funzione `json.loads()`. Se il caricamento ha successo e l'oggetto JSON estratto è un dizionario, questo viene aggiunto ad una lista che viene restituita dal metodo come *output*.

Successivamente, per ogni dizionario estratto attraverso il metodo `extract_json`, viene creato un DataFrame Pandas contenente i dati corrispondenti ed aggiunto ad una lista. Una volta che tutti i DataFrame relativi agli *output* sono stati memorizzati, i risultati vengono concatenati per formare un unico DataFrame. Tale DataFrame viene successivamente esaminato per garantirne la conformità ai requisiti necessari per l'analisi dei risultati. In particolare, viene controllata la presenza di duplicati e, in caso positivo, si procede con l'eliminazione. Inoltre, qualora si verificasse che i dizionari JSON contenessero più colonne rispetto a quelle richieste, vengono mantenute solo le colonne essenziali: `context`, `kpi_name`, `value`, `measure_unit`, `year`. In aggiunta, per garantire l'accuratezza del valore estratto relativo al KPI, vengono verificati i valori delle colonne `measure_unit` e `year`, e, se necessario, vengono corretti. Successivamente, la colonna `value` viene convertita da stringa a `float` (numeri in virgola mobile) e le stringhe che non rappresentano un valore numerico vengono impostate con il valore nullo `null`. Questo provvedimento è indispensabile per la successiva analisi dei risultati, poiché è stato riscontrato che alcuni valori dei KPI all'interno dei *report* sono contrassegnati come “<Not Applicable>”.

4.5 Data Storage

Nella fase conclusiva del processo di estrazione di informazioni, si procede con il salvataggio dei risultati derivati dall'esecuzione di ciascun modello per ogni azienda ed anno considerati. Al fine di automatizzare questo passaggio, è stato sviluppato un metodo denominato `save_from_pandas_to_csv`. Tale metodo, ricevendo come *input* il DataFrame Pandas ottenuto dalla fase precedente (come descritto nella sezione 4.4), consente di archiviare i dati in esso contenuti in un file CSV all'interno del servizio di archiviazione Azure Blob Storage. In particolare, lo *script* sfrutta il metodo `to_csv()` fornito dalla libreria Pandas per convertire il DataFrame in una stringa in formato CSV. È importante notare che viene specificata l'opzione `index=False`, al fine di escludere l'indice del DataFrame nel file CSV risultante. Successivamente, viene creato un oggetto `StringIO`

attorno alla stringa CSV generata nella fase precedente. `StringIO` è un’interfaccia che permette di elaborare una stringa come se fosse un *file*, agevolando l’utilizzo di funzioni e metodi che richiedono un oggetto *file*. Infine, viene eseguito il salvataggio del CSV come *blob*, ossia un oggetto dati binari. Attraverso un *client blob*, ovvero un oggetto che rappresenta una connessione a un singolo *blob* o a un contenitore di *blob*, lo *script* accede al contenitore specifico all’interno del servizio di archiviazione Azure Blob Storage in cui si desidera memorizzare i *file* e carica la stringa CSV, come un *blob*, nel percorso specificato. Tale percorso include il nome del *file* CSV, il quale è composto utilizzando variabili come l’anno di riferimento per il *report* ed il nome dell’azienda. Nel caso in cui un *blob* con la stessa denominazione esista già nel contenitore, viene sollevata un’eccezione di tipo `ResourceExistsError`, la quale è gestita sovra-scrivendo il *blob* esistente con i nuovi dati. In assenza di eccezioni, il *blob* viene caricato nel contenitore senza sovrascrivere eventuali *blob* preesistenti.

4.6 Data Comparison (Ground Truth)

Per condurre uno studio finalizzato alla valutazione e comparazione dell’efficacia dei modelli LLM nell’estrazione di informazioni specifiche, utili per le aziende nel settore energetico, è stato necessario creare un *dataset* contenente i valori effettivi e corretti presenti all’interno dei documenti appartenenti alla lista di compagnie ed anni considerati, come precedentemente illustrato nella sezione 4.1.

Una volta acquisiti i risultati di ciascun modello per ogni iterazione, è stato sviluppato un *notebook* all’interno dell’ambiente Azure Databricks, progettato per confrontare tali risultati con i valori reali estratti nel *dataset* precedente (*ground truth*). Le fasi principali seguite all’interno di questo codice includono:

1. Configurazione delle credenziali e delle variabili di ambiente necessarie per l’accesso ai file contenuti nel Azure Blob Storage.
2. Caricamento del *dataset* appositamente costruito, utilizzato come *ground truth*.
3. Elaborazione di un Dataframe Pandas (`df_model_eval`) contenente le informazioni necessarie per verificare se i valori estratti dai modelli sono corretti. In particolare, è stato implementato un metodo in grado di prelevare dal Azure Blob Storage ciascun *file* CSV relativo all’estrazione di ogni modello per ogni *report* considerato. Successivamente, è stata iterata ogni riga di tale *dataset* e confrontata con la corrispondente riga nel *dataset* del *ground truth*. Il Dataframe `df_model_eval` include le seguenti informazioni:
 - `model_name`: Stringa contenente il nome del modello LLM in esame;
 - `company_name`: Stringa contenente il nome della compagnia a cui fa riferimento il *report*;
 - `year`: Numero intero indicante l’anno di riferimento del *report* considerato;
 - `context`: Stringa contenente il contesto del KPI analizzato;
 - `kpi_name`: Stringa contenente il nome del KPI analizzato;

- **value_model**: Numero decimale indicante il valore estratto dal modello;
- **value_gt**: Numero decimale indicante il valore vero presente all'interno del *report*;
- **in_report**: Valore *booleano* indicante se il KPI analizzato è presente nel *report* considerato;
- **kpi_found**: Valore *booleano* indicante se il KPI analizzato è stato trovato dal modello;
- **is_correct**: Valore *booleano* indicante se il valore estratto dal modello (**value_model**) coincide col valore reale (**value_gt**).

Viene, inoltre, condotto un ulteriore controllo per verificare se sono presenti tutti i KPI richiesti. Qualora il *dataset*, contenente i risultati del modello, non includa tutti i KPI richiesti, i KPI mancanti vengono aggiunti al DataFrame risultante con i valori **kpi_found** e **is_correct** impostati a FALSE.

4. Salvataggio del Dataframe risultante in un *file* CSV all'interno Azure Blob Storage (**df_model_eval.csv**). In particolare, è stata implementato un metodo al fine di controllare se il *file* contenente il confronto sia già presente all'interno dell'archivio Azure Blob Storage. Qualora il **blob** esista, allora il DataFrame già esistente viene recuperato ed unito con il nuovo DataFrame prima di essere salvato. Questo approccio garantisce l'aggregazione di tutti i risultati ottenuti durante il processo di estrazione dei modelli in un unico file di riferimento.

Dopo aver completato tale fase, ed averla iterata per ciascun *report* considerato all'interno di questo studio, si è proceduto con la fase di formulazione di considerazioni, visualizzazione dei risultati e con la conseguente analisi.

5. Considerazioni e visualizzazione dei risultati

Nel presente capitolo, vengono effettuate le considerazioni del lavoro svolto fino a questo momento. In particolare, si è dedicato maggior interesse all’osservazione ed alla descrizione delle caratteristiche di ciascun modello durante l’esecuzione del *task* richiesto. Dopodiché, a seguito delle considerazioni effettuate, si è proceduto nel tentativo di combinare le prestazioni di due modelli che risultavano avere caratteristiche complementari, ovvero: **Gemini** e **LLAMA**. Il primo ha dimostrato una buona comprensione del testo ma una limitata capacità di strutturare correttamente gli *output*, mentre il secondo ha mostrato una buona capacità di strutturare gli *output* ma una scarsa abilità nell’identificare le informazioni corrette nel testo. Di conseguenza, si è optato per una concatenazione delle risposte ottenute dai modelli. In particolare, a **Gemini** viene richiesto di identificare ed estrarre la porzione di testo relativa al KPI necessario, dopodiché, il testo viene dato in *input* a **LLAMA** al quale viene richiesto di estrarre l’informazione necessaria all’interno di un dizionario in formato JSON. Questo approccio mira a migliorare i modelli esistenti sfruttando le loro rispettive qualità migliori.

Successivamente, si procede con la presentazione delle *dashboard* generate tramite l’utilizzo dello strumento Power BI, le quali offrono un resoconto dettagliato delle prestazioni di ciascun modello. Le visualizzazioni comprendono una matrice di confusione che fornisce una rappresentazione visiva delle *performance* complessive, evidenziando i veri positivi, falsi positivi, veri negativi e falsi negativi. Questa matrice è utile per valutare la capacità del modello di estrarre correttamente i valori richiesti nel *task*. A seguire, viene presentato un grafico ad anello che illustra il tasso di veri positivi in comparazione con il tasso di falsi positivi, fornendo una panoramica immediata dell’efficacia del modello nell’estrazione delle informazioni. Inoltre, viene mostrato anche la visualizzazione denominata *funnel*, che offre una sintesi delle principali metriche di valutazione, tra cui *accuracy*, *precision*, *recall*, *specificity* e *F1 measure*. Questa rappresentazione a gradini permette di valutare rapidamente le prestazioni complessive del modello. Infine, un grafico a barre orizzontali mostra la percentuale di valori correttamente estratti per ogni categoria di KPI definita, consentendo di identificare le aree in cui il modello eccelle e quelle in cui può essere migliorato. Vengono fornite, anche, delle infografiche comparative di tutti i LLM, tra cui un grafico a barre orizzontali che mostra le metriche di *accuracy*, *precision*, *recall*, *specificity* e *F1 measure* per ognuno, permettendo di confrontare direttamente le prestazioni dei modelli in relazione a queste metriche chiave. All’interno di questa *dashboard*, è presente anche il grafico a barre verticali al fine di evidenziare il valore della *F1 measure*, offrendo un’ulteriore valutazione della capacità del modello di bilanciare *precision* e *recall*. Infine, vengono presentati due *scatterplot* che forniscono una rappresentazione grafica delle prestazioni in relazione al rapporto tra *precision* e *recall* ed allo spazio ROC puntiforme, consentendo di valutare il *trade-off* tra True Positive Rate (TPR) e False Positive Rate (FPR), al fine di confrontare l’efficacia dei modelli nell’esecuzione del compito richiesto.

5.1 Considerazioni

Durante la fase di implementazione, è stato possibile studiare come i diversi modelli hanno affrontato il compito di estrazione di specifiche informazioni all'interno di *report* aziendali. In particolare, in questa prima fase si sono potute identificare le seguenti caratteristiche:

- GPT3.5 ha dimostrato una maggiore affidabilità, rispetto alle altre soluzioni esaminate, grazie alla sua capacità di generare in modo coerente il dizionario JSON richiesto. Durante la fase di valutazione del modello, mediante ripetute esecuzioni con la medesima domanda ed il medesimo testo al fine di studiare il miglior *prompt* per tale modello e valutarne le capacità, non si evidenziano significative variazioni negli *output* prodotti. Pertanto, il modello sembra mantenere una stabilità nelle risposte fornite alle richieste. Tuttavia, il modello mostra delle difficoltà nella gestione dei valori nulli, ovvero quando non è presente il valore relativo al KPI di interesse all'interno del *report* analizzato, poiché spesso viene prodotto un *output* in forma testuale anziché un dizionario vuoto come richiesto.
- Dolly è stato il primo modello, tra quelli esaminati, a riscontrare significative difficoltà nell'adempimento del compito assegnato. La prima problematica incontrata, anticipata nella sezione 4.3.2, è emersa con la richiesta di produrre *output* sotto forma di dizionari JSON. Tale modello si è, quindi, dimostrato non in grado di fornire un *output* strutturato in formato JSON rendendo necessario un adattamento del codice alle peculiarità di tale strumento. Di conseguenza, è stato implementato un processo di estrazione dei singoli valori numerici e di costruzione automatica del dizionario. Successivamente, si sono riscontrate considerevoli instabilità durante la messa a punto del modello, evidenziate attraverso l'esecuzione ripetuta della stessa richiesta con il medesimo testo al fine di valutare l'idoneità dei *prompt* e le capacità del modello. In particolare, si è constatato che il modello non restituiva un valore uniforme ma mostrava una tendenza a variare ad ogni esecuzione. Per superare questa problematica, durante l'implementazione sono stati estratti cinque volte i valori del KPI richiesto per ogni singola richiesta e, come *output* finale, è stato selezionato il valore che compariva con maggiore frequenza.
- LLAMA ha dimostrato una maggiore rigidità, in relazione agli altri approcci, nell'estrazione di dizionari JSON come richiesto, anche in presenza di valori nulli. Sebbene talvolta inserisca commenti testuali, risulta essere il più rigoroso nel rispettare il formato richiesto per l'*output*. Tuttavia, la principale sfida riscontrata è legata, principalmente, alla comprensione del testo. Questo implica che, nonostante la sua capacità di generare un *output* strutturato come da richiesta e di compilare il dizionario con le informazioni adeguate, presenta notevoli difficoltà nell'individuare l'informazione corretta, spesso confuso da testi con informazioni simili ed incapace di distinguere i contenuti presenti in modo accurato.
- Gemini, infine, ha dimostrato una notevole capacità di comprensione del testo, consentendogli di individuare ed estrarre l'informazione corretta richiesta dal testo fornito, tuttavia, risulta spesso prolisso. Nella maggior parte dei casi, commenta l'*output* fornito cercando di spiegare la logica di estrazione, anche quando la richiesta esplicita la necessità di avere solamente la produzione di un dizionario JSON

adeguato come *output* del modello. Per questo motivo, il modello non è in grado di fornire un dizionario JSON vuoto nel caso in cui non riesca a trovare l'informazione richiesta all'interno del testo, ma fornisce, invece, una spiegazione testuale dell'accaduto. Inoltre, tende ad estrarre informazioni aggiuntive, non richieste, all'interno dell'*output*, includendo ulteriori dettagli oltre ai campi specificatamente richiesti (`context`, `kpi_name`, `value`, `measure_unit` e `year`).

Per quanto concerne la fase di autocorrezione degli *output* dei modelli, tramite l'applicazione del *prompt* `check_prompt`, precedentemente descritto nella sezione 4.3, si è osservata una significativa inadeguatezza del modello LLAMA. Quest'ultimo ha dimostrato una marcata incapacità nella generazione di un *output* concatenando la richiesta precedente, non fornendo alcuna risposta nonostante svariati tentativi di adattamento del *prompt*. Pertanto, si è optato per considerare esclusivamente il primo *output* generato da tale modello. Per quanto riguarda, invece, i restanti modelli, si sono riscontrate delle risposte quasi identiche a quanto prodotto nella prima richiesta. In particolare, Dolly si è rivelato poco affidabile poiché non è in grado di produrre in modo stabile gli *output*, mentre Gemini e GPT3.5 forniscono degli *output* validi, il primo mantenendo, però, la sua caratteristica di produrre risposte prolixe. Questo passaggio non sembra generare, pertanto, miglioramenti significativi in termini di qualità della risposta, poiché gli *output* sembrano essere piuttosto simili a quelli ottenuti durante la prima estrazione.

Dopo aver condotto un'attenta analisi per ciascun modello, sono state individuate caratteristiche complementari al fine di integrare le capacità osservate e verificare se sia possibile ottenere risultati con prestazioni superiori. In particolare, considerando la propensione di Gemini a generare *output* prolissi ma caratterizzati da una profonda comprensione del testo e la tendenza di LLAMA ad essere rigoroso nelle richieste ma limitato nella comprensione di testi complessi e nell'individuazione delle informazioni rilevanti, si è scelto di provare a concatenare gli *output* al fine di conseguire risultati più consoni alle richieste specifiche. Questa decisione è stata condotta, anche, allo scopo di esplorare la possibilità che l'implementazione di miglioramenti tramite LLM accessibili liberamente possa, eventualmente, raggiungere o avvicinarsi alle prestazioni di un modello a pagamento come GPT3.5, che, all'interno di questo studio, ha riscontrato una maggiore efficacia nell'affrontare e comprendere la richiesta iniziale, molto superiore ai restanti modelli.

Pertanto, una volta configurato l'ambiente Azure Databricks, al fine di ospitare entrambi i modelli, ed implementati i sistemi necessari per l'interazione con essi (in particolare il caricamento del modello LLAMA tramite Hugging Face e l'utilizzo di *cookie* per l'interazione col modello Gemini, come accuratamente descritto nelle sezioni 4.3.3 e 4.3.4), si è proceduto alla creazione dei *prompt* da utilizzare. In primo luogo, si è mantenuto per entrambi l'`instruction_prompt`, precedentemente descritto nella sezione 4.3, al fine di specificare il contesto di ricerca e fornire una specifica guida per affrontare il compito assegnato, come segue:

You are an expert in extracting sustainability Key Performance Indicators (KPI) from reports. I will provide you with some KPIs and your objective is to detect them from a given text, with name exactly as provided and extract their values.

You will output the detected KPIs with their values in a json format, or an empty json if no KPI among the provided ones is detected.

Successivamente, è stato creato un *prompt* specifico per **Gemini** al fine di consentire al modello di identificare la porzione di testo correlata al KPI da estrarre ed indicando come *output* esclusivamente il testo rilevante. Tale *prompt*, denominato **gemini_input**, è stato definito nel seguente modo:

```
gemini_input:  
  I will provide you a text extracted from sustainability reports.  
  You will search the content for all and only the KPIs listed below in the  
  "KPIs and metrics" section structured as follows:  
  - Context name: "context_name_example"; KPI name: "kpi_name_example";  
  Measure Unit: measure_unit_example; Year: year_example".
```

The reporting year of the report is the year before of "Year" in "KPIs and metrics" list.

You have to print only the part of text that contain in the title the "context_name_example" and then the value of the string "kpi_name_example".

Don't provide any comment.

KPIs and Metrics:
{kpi_list}

Input:
{pages_text}

Durante la fase di valutazione dell'efficacia del *prompt* realizzato, è emerso che il modello è in grado di produrre risultati migliori e meno affetti da rumore, ovvero commenti verbali aggiuntivi, come riscontrato nella precedente richiesta. Una volta ottenuto l'*output* da **Gemini**, contenente solo la porzione di testo relativa al KPI da estrarre, viene concatenato all'interno di un ulteriore *prompt* strutturato in modo identico allo **user_input**, precedentemente descritto nella sezione 4.3, e dato al modello **LLAMA** per estrarre il dizionario JSON relativo. In particolare, viene definito come segue:

```
llama_input:  
  I will provide you with individual pages extracted from sustainability  
  reports. You will search the content for all and only the KPIs listed  
  below in the "KPIs and metrics to search for" section structured as  
  follows:  
  - Context name: "context_name_example"; KPI name: "kpi_name_example";  
  Measure Unit: measure_unit_example; Year: year_example".
```

The string of the context is always before the name of the kpi.

Find in the input the value of the KPI contained in the below list "KPIs and Metrics to Search". Store the found KPI names exactly as they appear

```

in the input and their value in a dictionary JSON-like with this structure:
[ {
    "context": "context_name_example",
    "kpi_name": "kpi_name_example",
    "value": value_example,
    "measure_unit": measure_unit_example,
    "year": year_example
}, ... ]

```

If you find a different KPI not listed in "KPIs and Metrics to Search", do not add it to the json. Before adding a KPI to the json, check its value: all the values have to be numerical, otherwise ignore it.

KPIs and Metrics:
{`kpi_list`}

Input:
{`gemini_output`}

Per l'implementazione del *prompt check_prompt*, precedentemente descritto nella sezione 4.3, si è scelto di escluderlo dall'analisi in quanto il modello LLAMA ha riscontrato delle difficoltà nel fornire *output* consecutivi e nell'esecuzione degli altri modelli non si sono riscontrati significativi miglioramenti.

Nella sezione 5.2 vengono esposte le prestazioni di tutti i modelli esaminati, valutate attraverso l'applicazione di metriche specifiche.

5.2 Visualizzazione dei risultati

Attualmente, non esiste una metodologia ampiamente accettata per il confronto dei Large Language Models (LLM), data anche la relativa novità di tali sistemi. Tuttavia, stanno emergendo i primi tentativi di comparazione, tra questi modelli, per la generazione di testo. Le metriche attualmente impiegate per valutare le Intelligenze Artificiali generative non sono ancora completamente esaustive e spesso risultano fuorvianti, poiché non tengono pienamente conto dei numerosi aspetti di complessità di tali modelli. Questo può generare aspettative che non sempre vengono soddisfatte. Un esempio di tale iniziativa è la classificazione dei modelli ospitata sulla piattaforma Hugging Face [97], in cui i LLM vengono confrontati attraverso sette *benchmark* chiave, utilizzando “Eleuther AI Language Model Evaluation Harness” ovvero un *framework* sviluppato da EleutherAI [84] per testare i modelli linguistici generativi su una vasta gamma di compiti di valutazione. È importante notare, però, che i modelli di OpenAI e Google non sono presenti in questa lista poiché i requisiti richiedono che il modello sia disponibile sulla piattaforma.

Di conseguenza, poiché il presente lavoro si è dedicato all'analisi dei modelli per l'estrazione di valori di specifici KPI indicati, si è optato per l'impiego di metriche consuete nel

contesto della valutazione delle prestazioni dei modelli di classificazione. In particolare, viene calcolata la matrice di confusione al fine di ottenere una rappresentazione dell'accuratezza per ogni modello. Questa matrice viene generata considerando tre caratteristiche create durante la fase di Data Comparison (descritta nella sezione 4.6), ovvero:

- `in_report`, booleano indicante se il KPI analizzato è presente nel *report* considerato;
- `kpi_found`, booleano indicante se il KPI analizzato è stato trovato dal modello;
- `is_correct`, booleano indicante se il valore estratto dal modello (`value_model`) coincide col valore reale (`value_gt`).

Per poter calcolare la matrice di confusione ed includere tutte le possibili casistiche riscontrate durante l'estrazione dei KPI, sono stati determinati i seguenti valori:

- Veri Positivi (True Positive, TP), valori dei KPI presenti nel *report*, individuati dal modello e correttamente estratti;
- Falsi Positivi (False Positive, FP), valori dei KPI presenti nel *report*, individuati dal modello ma non correttamente estratti oppure valori dei KPI non presenti nel *report* ma individuati dal modello e, pertanto, non corretti (identificati come allucinazioni);
- Veri Negativi (True Negative, TN), valori dei KPI non presenti nel *report*, non individuati dal modello e quindi non estratti;
- Falsi Negativi (False Negative, FN), valori dei KPI presenti nel *report* ma non individuati dal modello e quindi non estratti.

Nel presente studio, pertanto, si attribuisce particolare importanza alla definizione dei False Positive tramite una prospettiva più rigorosa. Tale definizione comprende sia i casi in cui il KPI è effettivamente presente nel *report* ma il modello non lo estrae correttamente, sia gli scenari in cui il modello allucina, ovvero estrae un KPI che in realtà non esiste all'interno del *report*. Questa definizione, più restrittiva rispetto alla consuetudine, considera i falsi positivi non solo come previsioni positive errate, ma li interpreta anche come risultati che portano ad una penalizzazione maggiore. L'obiettivo è, quindi, è associare un maggiore peso a tutte le estrazioni errate effettuate dal modello al fine di migliorare l'identificazione e la successiva correzione. Nella **Tabella 5.1** vengono mostrate le casistiche appena descritte.

KPI matrice	<code>in_report</code>	<code>kpi_found</code>	<code>is_correct</code>
True Positive (TP)	TRUE	TRUE	TRUE
False Positive (FP)	TRUE	TRUE	FALSE
	FALSE	TRUE	FALSE
True Negative (TN)	FALSE	FALSE	FALSE
False Negative (FN)	TRUE	FALSE	FALSE

Tabella 5.1. Definizione dei TP, FP, TN e FN rispetto alle informazioni raccolte durante l'implementazione. È importante notare che si è considerato come FP un caso più restrittivo, dato dall'unione di due condizioni.

Un’ulteriore osservazione riguarda la mancanza della combinazione FALSE TRUE TRUE nella **Tabella 5.1**. Questa situazione non può verificarsi in alcun modo: se un KPI è stato estratto dal modello ed è corretto, deve essere riportato nel documento; se non è presente, allora non può essere considerato corretto, poiché non ha un valore assegnato. Pertanto, tale terna non può essere inclusa all’interno della tabella.

Una volta definite le condizioni necessarie per definire i veri positivi, falsi positivi, veri negativi e falsi negativi, per ogni modello vengono calcolate le seguenti metriche:

- *Accuracy*, rappresenta la percentuale di predizioni corrette effettuate dal modello rispetto al totale delle predizioni eseguite.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Questa metrica offre una valutazione della correttezza complessiva del modello.

- *Precision*, indica la proporzione di predizioni positive correttamente identificate rispetto al totale delle predizioni positive effettuate dal modello, ovvero le predizioni corrette insieme agli errori ed allucinazioni. Essa viene calcolata come il rapporto tra il numero di veri positivi e la somma dei veri positivi e dei falsi positivi.

$$Precision = \frac{TP}{TP + FP}$$

- *Recall*, anche nota come *sensitivity*, rappresenta la capacità del modello di identificare in modo preciso i casi positivi rispetto al totale dei casi positivi presenti. Pertanto, oltre ai True Positive, ovvero in cui il KPI è presente nel *report* e correttamente estratto dal modello, è importante considerare anche i casi positivi che il modello ignora erroneamente, identificati come Falsi Negativi. In tal caso, però, è rilevante includere anche i casi erroneamente classificati come positivi dal modello, evidenziati nella tabella con la terna TRUE TRUE FALSE. È significativo includere tali casistiche all’interno della formula poiché influenzano molto la capacità del modello di individuare accuratamente i casi positivi. Di conseguenza, questa metrica si esprime come il rapporto tra il numero di veri positivi e la somma dei veri positivi, dei falsi negativi e dei casi estratti erroneamente dal modello (indicato con la denominazione “*real true*”). Tale approccio consente di ottenere una valutazione precisa della capacità del modello di identificare correttamente i casi positivi nell’ambito specifico della ricerca.

$$Recall = \frac{TP}{real\ true}$$

- *Specificity*, conosciuta anche come True Negative Rate, indica la percentuale di predizioni negative correttamente identificate rispetto al totale delle predizioni negative effettuate dal modello. Pertanto, questo indicatore concentra l’attenzione esclusivamente sui casi negativi. Per il calcolo si considerano, quindi, i True Negative insieme alla porzione di False Positive contrassegnati nella **Tabella 5.1** con la terna FALSE TRUE FALSE, ovvero KPI estratti dal modello ma non presenti nel

report. In questo modo viene garantito che vengano considerati esclusivamente i casi negativi. Di conseguenza, la metrica si calcola come il rapporto tra il numero di True Negative e la somma dei True Negative insieme alla porzione di casi generati da allucinazioni del modello, denominata “*real false*” all’interno della formula.

$$Specificity = \frac{TN}{real\ false}$$

- *F1 measure*, definita come la media armonica tra *precision* e *recall*, è particolarmente utile quando si desidera trovare un equilibrio tra queste due metriche.

$$F1\ measure = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Al fine di condurre un’analisi completa relativa alle prestazioni dei diversi modelli utilizzati e facilitare la comprensione dei risultati, attraverso visualizzazioni grafiche intuitive, è stato sviluppato un sistema di visualizzazione interattiva utilizzando il *software* di analisi dati Power BI. Tali infografiche includono una matrice di confusione rappresentante i veri positivi (TP), falsi positivi (FP), veri negativi (TN) e falsi negativi (FN) in percentuale rispetto al totale dei KPI analizzati, fornendo una panoramica completa delle previsioni effettuate dai modelli. In aggiunta, è stato implementato un grafico ad anello che confronta il tasso di falsi positivi, percentuale di casi negativi erroneamente classificati come positivi dal modello (False Positive Rate, FPR), con il tasso di veri positivi, percentuale di casi positivi correttamente identificati dal modello (True Positive Rate, TPR). In particolare:

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN}$$

La visualizzazione di tali metriche è in grado di facilitare la valutazione delle prestazioni del modello in termini di sensibilità e specificità, permettendo di valutare la sua capacità di identificare correttamente i casi positivi rispetto alle false positività. Successivamente, viene utilizzato un *funnel* per rappresentare le diverse metriche di valutazione considerate, come l’*accuracy*, la *precision*, la *recall*, la *specificity* e la *F1 measure*. Tale approccio consente di individuare eventuali punti di forza e di debolezza dei modelli e di valutare quale metrica sia più significativa rispetto al contesto specifico dell’applicazione. Infine, è stato incluso un grafico a barre orizzontali che visualizza la percentuale di KPI estratti correttamente da ciascun modello per ogni categoria di KPI considerata, permettendo una valutazione dettagliata delle capacità predittive nel contesto specifico del dominio di interesse. A seguire vengono mostrate le singole *dashboard* create con le infografiche appena descritte.

All’interno della **Figura 5.1** vengono esaminate le prestazioni del modello GPT3.5 nel contesto di ricerca. Il modello è in grado di identificare correttamente il 74% delle informazioni richieste, mentre estrae il 17% di valori errati o generati da allucinazioni. È in grado di individuare correttamente poco più del 3% delle informazioni non presenti nei *report* analizzati e, allo stesso tempo, estrae erroneamente poco più del 5% dei KPI non presenti. Il grafico ad anello evidenzia le prestazioni del modello nell’individuare i KPI

all'interno dei *report*. Il True Positive Rate (TPR) raggiunge il 52,88%, dimostrando una buona capacità nel rilevare accuratamente i KPI. D'altra parte, il False Positive Rate (FPR) è del 47,12%, suggerendo che vi è margine di miglioramento per la riduzione dei falsi positivi generati dal modello. Il *funnel* mostra un'accuracy complessiva del 78%, indicando una buona capacità di classificare correttamente i KPI. La *precision* raggiunge il valore di 81%, mentre la *recall* si alza al 93%, riflettendo la capacità del modello di riconoscere correttamente i casi positivi tra tutti quelli effettivi. La *specificity* risulta invece più bassa, attestandosi al 17%, indicando una minore capacità del modello di identificare correttamente i casi negativi tra quelli effettivi. La *F1 measure*, considerando sia la *precision* che la *recall*, raggiunge un valore di 0,87, suggerendo un buon bilanciamento tra la precisione e la completezza delle previsioni del modello. Infine, il grafico a barre orizzontali mostra la percentuale di categorie di KPI estratte correttamente, evidenziando una maggiore numerosità nell'estrazione della categoria "Emission in reporting year (Metric Tons CO₂e)" rispetto alle altre categorie presenti. Questi risultati indicano che il modello GPT3.5 ha una buona capacità nell'identificazione e classificazione dei KPI, sebbene vi sia margine di miglioramento, soprattutto per quanto riguarda la riduzione dei falsi positivi e l'incremento della specificità.

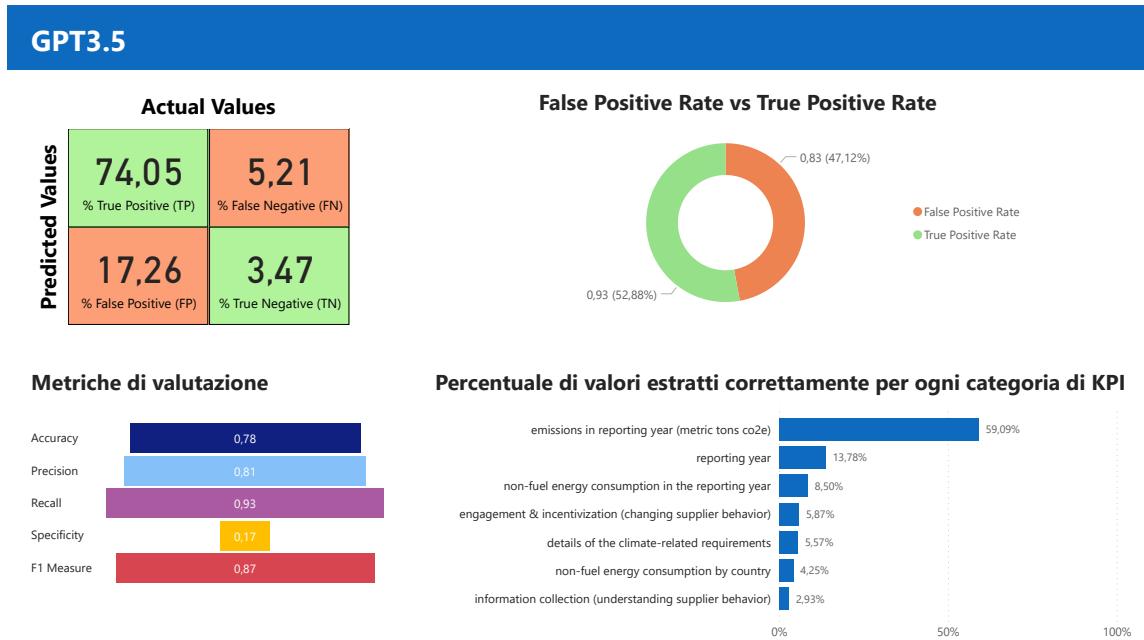


Figura 5.1. Dashboard riassuntiva per la valutazione del modello GPT3.5.

Nella **Figura 5.2**, vengono esaminate le prestazioni del modello Dolly. La matrice di confusione evidenzia che il modello identifica correttamente quasi il 22% delle informazioni richieste, ma estrae erratamente o genera allucinazioni per oltre il 73% dei KPI. Inoltre, non estrae correttamente poco più dell'1% dei KPI presenti nei *report* ma riesce a rilevare correttamente più del 3% dei KPI non presenti. Il grafico ad anello mostra un True Positive Rate (TPR) del 49,81% ed un False Positive Rate (FPR) del 50,19%. Rispetto al modello precedente, si osserva che il secondo modello ha un TPR leggermente inferiore e, di conseguenza, un FPR leggermente superiore. Questo suggerisce che il modello è meno preciso nel rilevare i KPI all'interno dei *report*, identificando una percentuale poco

inferiore di casi positivi e generando un numero poco maggiore di falsi positivi. Dal *funnel* emergono un'accuracy ed una precision molto basse per Dolly, rispettivamente del 25% e del 23%, pertanto, solo una piccola percentuale delle predizioni del modello è corretta rispetto al totale dei casi e delle predizioni positive, solo una piccola percentuale è effettivamente corretta. La recall risulta essere molto più alta, pari al 95%, indicando che il modello è in grado di identificare correttamente la maggior parte dei casi positivi rispetto al totale dei casi effettivi positivi. Tuttavia, questa metrica è accompagnata da una bassa precisione, suggerendo che il modello potrebbe essere troppo incline a predire positivamente. La specificity raggiunge il 5%, suggerendo che il modello ha una bassa capacità di distinguere correttamente i casi negativi tra tutti i casi negativi effettivi e, quindi, indicando che il modello è incline ad identificare falsi allarmi come negativi. Infine, la F1 measure risulta anch'essa molto bassa, con un valore di 0,37, suggerendo un bilanciamento poco soddisfacente tra precision e recall, indicando che il modello ha difficoltà a raggiungere una buona precisione senza compromettere la completezza delle previsioni o viceversa. Riguardo alle categorie di KPI estratte correttamente, si nota che la categoria "Emission in reporting year (Metric Tons CO₂e)" viene estratta più facilmente, anche più frequentemente rispetto al modello GPT3.5. Tuttavia, si evidenzia una percentuale di successo molto più bassa per le altre categorie. In particolare, si nota che due categorie, "Engagement & incentivization (changing supplier behavior)" e "Information collection (understanding supplier behavior)", non vengono mai identificate dal modello e, per questo motivo, non sono presenti all'interno del grafico. Questi risultati suggeriscono che il modello Dolly ha prestazioni inferiori rispetto al modello GPT3.5 in termini di accuratezza e precisione nello svolgimento del compito, evidenziando la necessità di miglioramenti, al fine di rendere il modello più efficace ed affidabile.

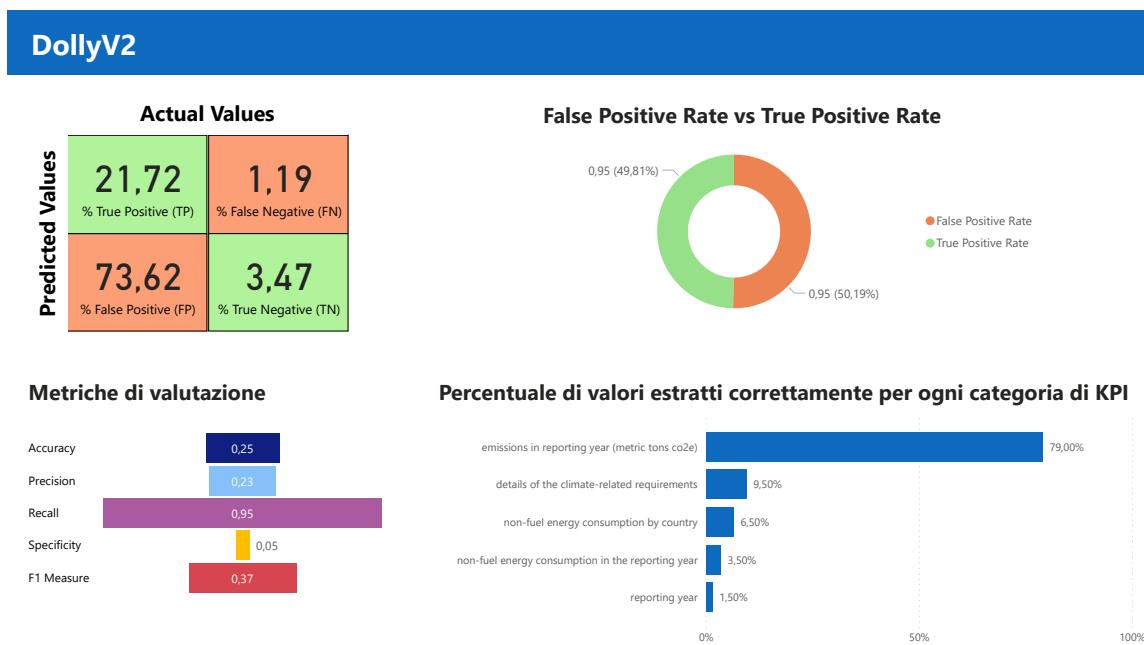


Figura 5.2. Dashboard riassuntiva per la valutazione del modello Dolly.

Il modello LLAMA, analizzato nella **Figura 5.3**, presenta una matrice di confusione composta dal 29% di veri positivi, indicando i KPI estratti correttamente, un 14% di valori

estratti non correttamente (errati o a seguito di allucinazioni del modello), poco più del 3% di valori correttamente non identificati poiché assenti nei *report* e quasi il 54% di informazioni erroneamente non estratte. Il grafico ad anello mostra un True Positive Rate (TPR) del 30,57% ed un False Positive Rate (FPR) del 69,43%. Rispetto ai modelli precedenti, questo terzo modello presenta una prestazione significativamente inferiore nel rilevare correttamente i KPI all'interno dei *report*. Il basso TPR indica che il modello identifica solo una percentuale relativamente piccola di casi positivi, mentre l'alto FPR indica che genera un numero significativamente maggiore di falsi positivi, evidenziando una carenza nella capacità di distinguere efficacemente tra casi positivi e negativi. Il *funnel* mostra un'accuracy del 33%, indicando che il modello classifica correttamente solo un terzo dei casi totali e suggerendo una bassa precisione complessiva delle previsioni. Tuttavia, la *precision* del 68% suggerisce che quando il modello effettua una previsione positiva, ha una probabilità relativamente alta di essere corretta, evidenziando una buona capacità di discriminare i veri positivi dai falsi positivi. La *recall*, invece, raggiunge il 35%, questo suggerisce che il modello identifica solo una piccola frazione dei veri positivi presenti nei dati, indicando una mancanza di sensibilità nel catturare tutti i casi positivi effettivi. La *specificity* del 20% indica che il modello ha una bassa capacità di distinguere correttamente i casi negativi tra quelli effettivi, indicando una tendenza a classificare erroneamente molti casi negativi come positivi. Infine, la metrica *F1 measure*, che combina *precision* e *recall*, raggiunge un valore di 0,46, indicando un compromesso tra la precisione e la completezza delle previsioni del modello. Questo valore suggerisce che il modello ha una prestazione media nel bilanciare l'accuratezza e la sensibilità durante le estrazioni.

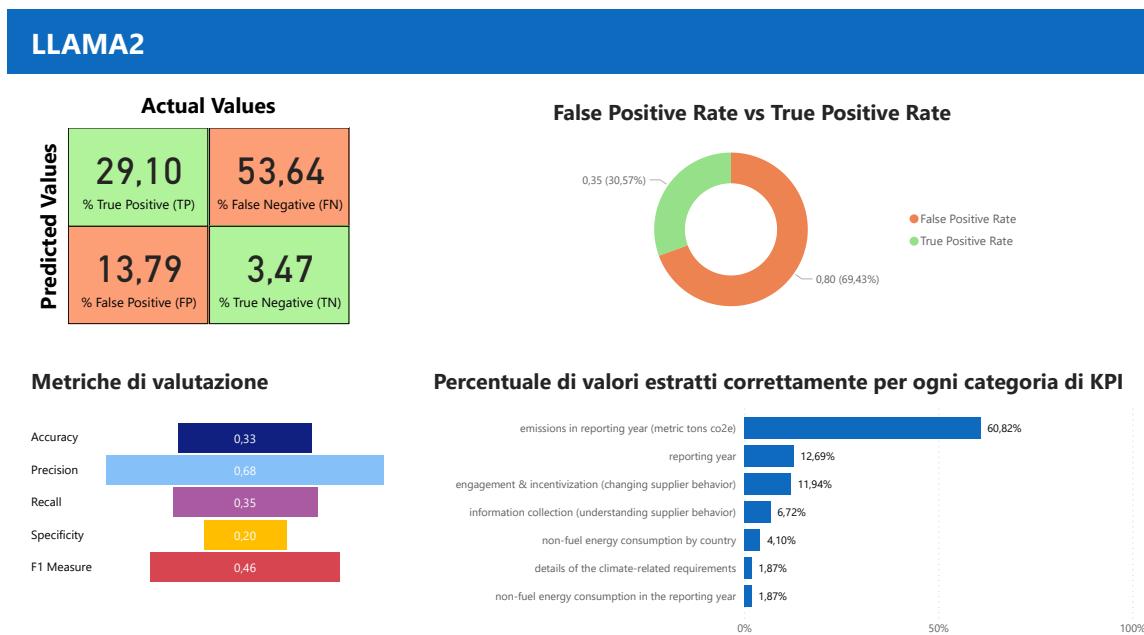


Figura 5.3. Dashboard riassuntiva per la valutazione del modello LLAMA2.

Per quanto riguarda Gemini, nella **Figura 5.4** è possibile osservare una matrice di confusione composta da 43,5% di valori estratti correttamente, poco più del 2% di KPI estratti a seguito di allucinazioni del modello o con valori errati, poco più del 3% correttamente non estratti poiché assenti nei *report* ma una quantità pari al 50,6% di informazioni

estratte erroneamente, poiché non presenti. Il True Positive Rate (TPR) raggiunge il 53,17%, indicando che il modello è in grado di identificare correttamente oltre la metà dei casi positivi di KPI nei *report* analizzati, rappresentando un miglioramento rispetto al terzo modello e suggerendo una maggiore sensibilità nel rilevare i KPI. Tuttavia, il False Positive Rate (FPR) del 46,83% indica che il modello genera ancora un numero significativo di falsi positivi, che potrebbero influenzare negativamente le interpretazioni dei dati e le decisioni aziendali. Il *funnel* evidenzia un'accuracy del 47%, pertanto indica che il modello classifica correttamente meno della metà dei casi totali, suggerendo una precisione complessiva delle previsioni che potrebbe essere migliorata. Tuttavia, la *precision* molto alta, pari al 95%, denota che quando il modello effettua una previsione positiva, ha una probabilità estremamente elevata di essere corretta, indicando una capacità eccezionale nel discriminare i veri positivi dai falsi positivi. Riguardo i valori di *recall* e *specificity*, pari rispettivamente al 46% e al 59%, si evidenzia che il modello ha una capacità equilibrata nel catturare i veri positivi e nel distinguere correttamente i veri negativi tra quelli effettivi. La *recall* del 46% indica che il modello identifica circa la metà dei veri positivi presenti nei dati, mentre la *specificity* del 59% indica che il modello ha una buona capacità di distinguere correttamente i casi negativi tra quelli effettivi. Infine, la metrica *F1 measure* raggiunge il valore di 0,62, indicando un buon equilibrio tra la precisione e la completezza nel modello. In relazione alle categorie di KPI estratte correttamente, si osserva ancora una prevalenza della categoria “Emission in reporting year (Metric Tons CO₂e)”, con un andamento complessivo simile al modello GPT3.5. Questi risultati indicano che il modello mostra una prestazione migliore rispetto al modello LLAMA, ma potrebbe comunque beneficiare di ulteriori ottimizzazioni, specialmente per ridurre il numero di falsi positivi e migliorare ulteriormente la precisione delle estrazioni.

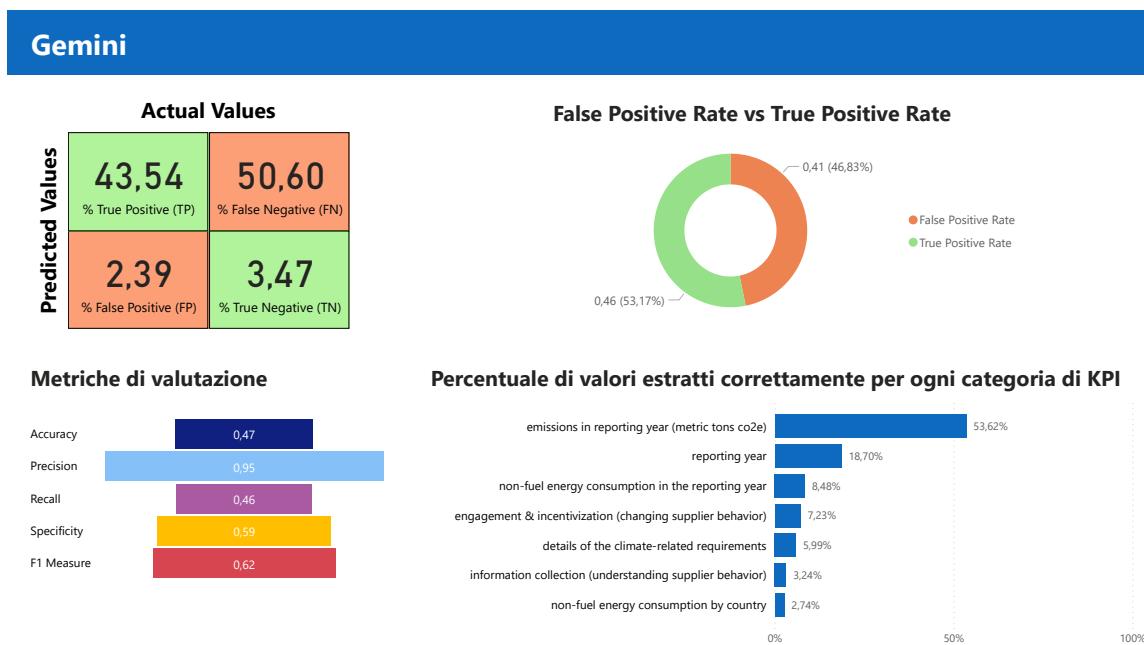


Figura 5.4. Dashboard riassuntiva per la valutazione del modello Gemini.

Infine, viene presentata la *dashboard* relativa al modello composto dalla concatenazione degli *output* di Gemini e LLAMA, visibile nella **Figura 5.5**. La matrice di confusione

mostra circa il 46% di valori correttamente estratti, più del 14% di valori estratti non correttamente, o prevenienti da allucinazioni del modello, più del 3% di valori non estratti poiché assenti nei *report* e quasi il 34% di valori estratti erroneamente poiché non presenti nei *report*. Il grafico ad anello mostra un True Positive Rate (TPR) del 41,14% ed un False Positive Rate (FPR) del 58,86%. Questi risultati indicano un miglioramento rispetto a modelli precedentemente analizzati, come LLAMA, ma evidenziano ancora margine per miglioramenti. Il TPR suggerisce che il modello è in grado di identificare correttamente oltre il 40% dei casi positivi di KPI nei *report* analizzati. Tuttavia, il FPR indica che il modello genera ancora un numero significativo di falsi positivi, indicando la necessità di ulteriori ottimizzazioni per ridurre tale valore. Il *funnel* del modello combinato evidenzia un'accuracy del 50%, evidenziando, quindi, che il modello classifica correttamente la metà dei casi totali, suggerendo una precisione complessiva delle previsioni che potrebbe essere migliorata. La *recall* pari al 56% indica che il modello identifica correttamente oltre la metà dei veri positivi presenti nei dati, evidenziando una buona capacità di catturare i casi positivi effettivi. La *precision* del 76% indica che quando il modello effettua una previsione positiva, ha una probabilità elevata di essere corretta, suggerendo una capacità notevole di discriminare i veri positivi dai falsi positivi. Tuttavia, la *specificity* molto bassa del 19% indica che il modello ha una bassa capacità di distinguere correttamente i casi negativi tra quelli effettivi, indicando una tendenza a classificare erroneamente molti casi negativi come positivi. Infine, la *F1 measure* pari a 0,65 indica un buon equilibrio tra *precision* e *recall* nel modello. Tuttavia, avendo un valore di *specificity* molto bassa è evidente la necessità di ulteriori ottimizzazioni per migliorare la capacità del modello di discriminare correttamente i casi negativi. Per le categorie di KPI estratte, "Emission in reporting year (Metric Tons CO₂e)" risulta essere ancora la maggiormente identificata correttamente, con una percentuale maggiore rispetto a tutti i modelli. Le altre categorie presentano un andamento simile al modello Gemini, ma leggermente più basso.

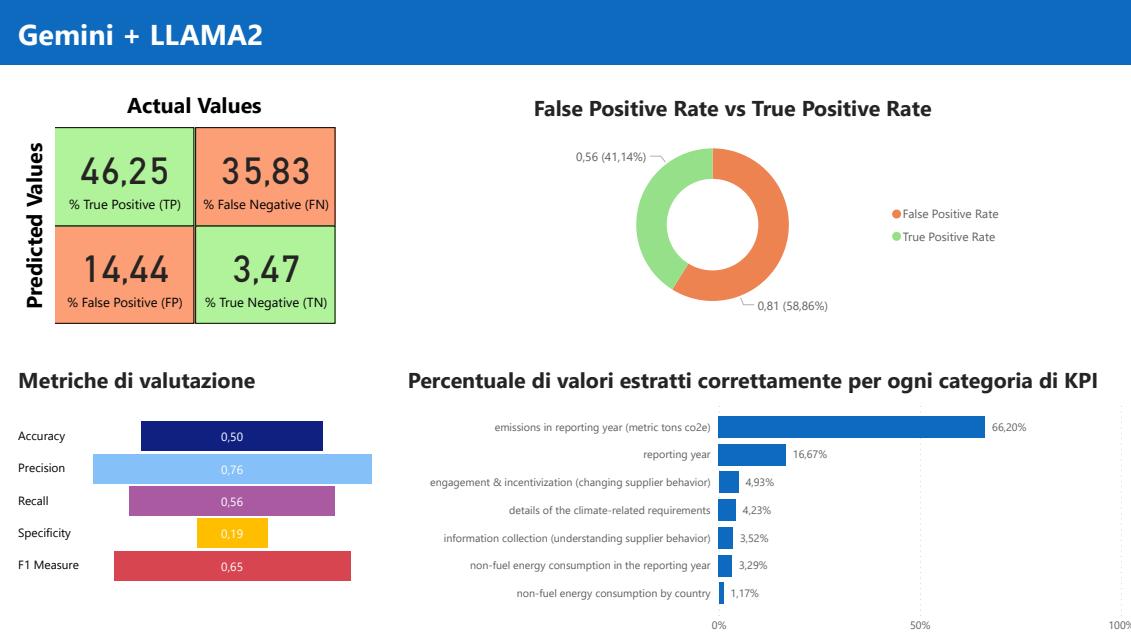


Figura 5.5. Dashboard per la valutazione dei modelli Gemini e LLAMA concatenati.

Questi risultati indicano che il modello combinato di **Gemini** e **LLAMA** mostra un miglioramento in termini di TPR, rispetto al modello **LLAMA** utilizzato singolarmente, ma il FPR rimane ancora molto elevato. Questo suggerisce che il modello potrebbe beneficiare di ulteriori miglioramenti per la precisione e la riduzione di casi falsi positivi.

Nella **Figura 5.6**, viene mostrato il grafico a barre orizzontali che rappresenta i valori delle metriche di *accuracy*, *precision*, *recall* e *specificity* relativi a tutti i modelli considerati nella ricerca. Tale rappresentazione consente una visione complessiva delle prestazioni dei diversi modelli, replicando le analisi precedentemente condotte per consentire un confronto visivo delle caratteristiche dei diversi modelli. In sintesi, il modello **GPT3.5** mostra la più elevata accuratezza (0.78) e specificità (0.93), conferendogli affidabilità per compiti di classificazione, anche se il risultato ottenuto per la *recall*. Il modello **DollyV2**, nonostante presenti la più bassa accuratezza (0.25), registra una *recall* più elevata (0.95), indicando una capacità di identificare positivi veri, anche a fronte di molti falsi positivi, risultando quindi idoneo per *task* in cui la rilevazione dei positivi veri sia cruciale accettando la presenza di falsi positivi. **Gemini** e **LLAMA2**, invece, si collocano tra questi estremi: il primo risulta più adatto per *task* in cui sia importante bilanciare *precision* e *recall*, mentre il secondo potrebbe essere impiegato in situazioni dove la specificità è meno critica e la *recall* assume priorità. La combinazione di **Gemini** e **LLAMA2** conduce a prestazioni bilanciate tra accuratezza, precisione e *recall*, evidenziando un miglioramento della *recall* di **Gemini** e della precisione di **LLAMA2**. Pertanto, questa combinazione potrebbe essere adatta in contesti dove la specificità non sia cruciale.

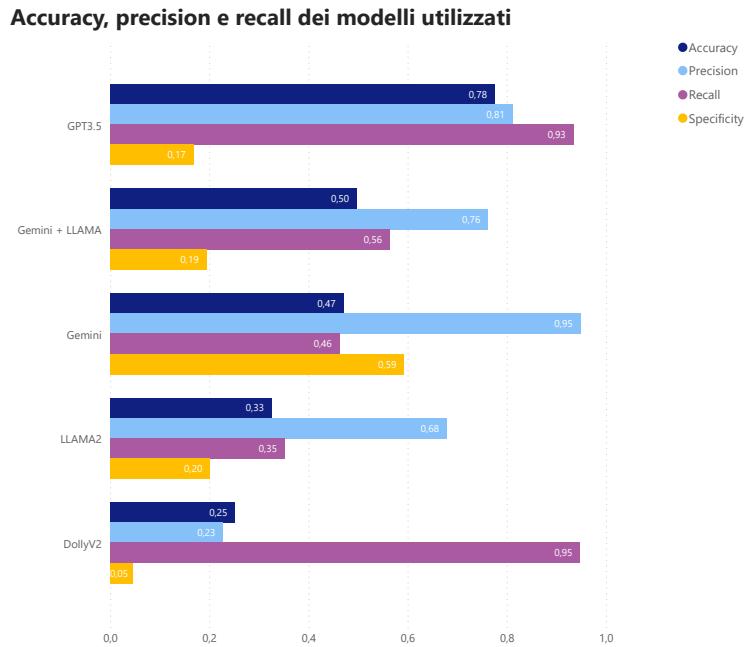


Figura 5.6. Grafico a barre orizzontali raffigurante le metriche di *accuracy*, *precision*, *recall* e *specificity*.

Nella **Figura 5.7**, sono presentati tre grafici che consentono di confrontare i modelli utilizzati all'interno della presente ricerca. Il primo grafico, rappresentato tramite un

grafico a barre verticali, mostra i valori della misura *F1 Measure* ottenuti dai cinque diversi modelli utilizzati per l'identificazione dei KPI ESG all'interno dei *report*. Ogni barra del grafico rappresenta un singolo modello e l'altezza della barra corrisponde al valore di *F1 measure* associato a quel modello. La *F1 measure*, una metrica che tiene conto sia della *precision* che del *recall* del modello, è calcolata come la media armonica tra queste due misure. Essa risulta particolarmente utile quando si desidera bilanciare l'importanza di entrambe le metriche e può variare da 0 a 1, dove un valore di 1 indica un perfetto equilibrio tra *precision* e *recall*. Un valore elevato di *F1 measure* suggerisce quindi un buon equilibrio tra le metriche considerate, indicando una capacità complessiva del modello nell'identificare accuratamente i KPI senza generare un numero eccessivo di falsi positivi o falsi negativi. Analizzando il grafico, emergono differenze nei valori di *F1 measure* tra i vari modelli. Il modello GPT3.5 si distingue per un valore superiore, indicando una maggiore efficacia complessiva nell'estrazione dei KPI rispetto ai modelli alternativi. Al contrario, sia **Gemini** che la combinazione tra **Gemini** e **LLAMA** mostrano valori inferiori rispetto al modello precedente, pur presentando valori simili, intorno al 0.65. **LLAMA** e **Dolly** registrano punteggi ancora più bassi, rispettivamente 0.46 e 0.37, indicando la necessità di miglioramenti per ottenere una maggiore *precision* e *recall*.

Il secondo grafico mostra uno *scatterplot* che confronta le metriche di *precision* e *recall* dei modelli considerati. Ogni simbolo nel grafico rappresenta un singolo modello e la posizione viene, appunto, determinata dai valori di *precision* e *recall* ottenuti durante la fase di valutazione. L'asse delle ordinate corrisponde alla *recall*, indicando la percentuale di veri positivi correttamente identificati rispetto al totale dei casi positivi presenti nei *report*. Più un modello si avvicina all'angolo superiore destro del grafico, maggiore è la sua *recall*, indicando, quindi, una maggiore capacità nel rilevare efficacemente i KPI all'interno dei *report*. L'asse delle ascisse, invece, rappresenta la *precision*, indicando la percentuale di veri positivi identificati correttamente rispetto al totale dei casi positivi individuati dal modello. Maggiore è il posizionamento di un modello verso l'angolo superiore destro del grafico, maggiore è la sua *precision*, evidenziando una capacità più elevata di individuare correttamente i KPI senza generare un eccesso di falsi positivi. Osservando lo *scatterplot*, emergono significative differenze nelle prestazioni dei diversi modelli. **LLAMA** presenta valori bassi sia per la *recall* che per la *precision*. **Dolly** mostra valori molto elevati per la *recall*, leggermente superiori a **GPT3.5**, ma una *precision* molto bassa, a differenza del modello **GPT3.5**. **Gemini** e la combinazione di **Gemini** con **LLAMA** presentano valori simili per la *precision*, con il primo leggermente superiore, ma per la *recall* risulta migliore il secondo. Questo indica che l'unione di **Gemini** con **LLAMA** permette di avere meno allucinazioni (meno falsi positivi), riscontrando una tendenza a migliorare la propria sensibilità nel riconoscere i veri positivi.

Infine, il terzo grafico rappresenta lo *scatterplot* relativo allo spazio ROC, mostrando le prestazioni dei cinque diversi modelli considerati. Ogni punto nel grafico corrisponde ad un singolo modello e la posizione nello spazio è determinata dal tasso di True Positive Rate (TPR) rispetto al tasso di False Positive Rate (FPR). L'asse delle ascisse rappresenta il FPR, indicando la percentuale di casi negativi erroneamente identificati come positivi dal modello. Più un modello si sposta verso sinistra sull'asse delle ascisse, minore è il FPR e migliore è la sua capacità di evitare falsi positivi. L'asse delle ordinate, invece,

rappresenta il TPR, indicando la percentuale di casi positivi correttamente identificati dal modello. Più un modello si sposta verso l'alto sull'asse delle ordinate, maggiore è il TPR e migliore è la sua capacità di individuare correttamente i casi positivi. Analizzando lo *scatterplot*, si notano differenze nelle prestazioni dei vari modelli. I modelli che si trovano più in alto e più a sinistra nel grafico mostrano prestazioni migliori, con un equilibrio tra un alto TPR ed un basso FPR, indicando una maggiore sensibilità e specificità nel rilevare i KPI ESG. Tuttavia, si osserva che i modelli faticano a rimanere nella zona sinistra del grafico. Il modello **Gemini**, sebbene sia il più a sinistra, mostra prestazioni mediocri per la sensibilità. **LLAMA** evidenzia prestazioni basse sia per la sensibilità che per il FPR. **Gemini** unito con **LLAMA** migliora nella sensibilità, suggerendo che tale fusione potrebbe portare a miglioramenti significativi nelle prestazioni complessive, mentre **GPT3.5** conferma le migliori prestazioni per la sensibilità, ma con basse prestazioni per il FPR. **Dolly**, infine, presenta una sensibilità leggermente migliore di **GPT3.5**, ma una pessima *performance* per il FPR. Pertanto, nessun modello raggiunge la condizione ideale di elevata sensibilità e basso tasso di falsi positivi contemporaneamente, suggerendo altrettanti miglioramenti in tutti i modelli con uno specifico focus sulla riduzione dei falsi positivi, senza compromettere eccessivamente la sensibilità.

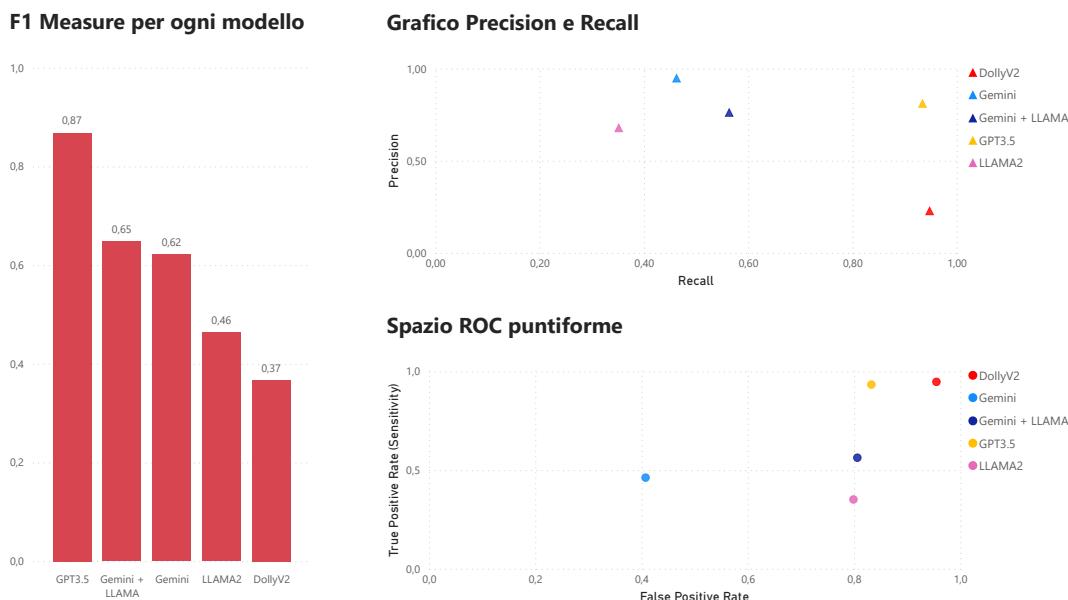


Figura 5.7. Dashboard raffigurante il grafico a barre verticali per la metrica *F1 measure* per i modelli considerati e due *scatterplot* relativi al confronto di *precision* e *recall* e i valori relativi allo spazio ROC per ogni modello considerato.

6. Conclusioni e sviluppi futuri

Il presente studio ha focalizzato la propria attenzione sull’analisi delle capacità offerte dalle nuove tecnologie emergenti nel campo tecnologico, in particolare sui modelli di Intelligenza Artificiale generativa, i quali hanno recentemente registrato un diffuso interesse ed un significativo progresso. L’obiettivo primario è stato valutare l’applicabilità di tali tecnologie all’interno dell’ambiente lavorativo. In particolare, si è analizzata la loro efficacia nell’ambito dell’analisi di mercato focalizzata sugli aspetti ambientali, sociali e di *governance* (ESG) per il settore energetico. Al fine di garantire la solidità di tale ricerca, è stato adottato un approccio comparativo che ha coinvolto diversi modelli di Intelligenza Artificiale, ciascuno caratterizzato da specifiche peculiarità, quali ad esempio l’utilizzo a pagamento oppure la modalità *open source*, l’utilizzo mediante il caricamento in locale del modello con i rispettivi parametri o l’integrazione diretta tramite API REST. In particolare, la fase di implementazione ha coinvolto i modelli GPT3.5 della società OpenAI, disponibile mediante API REST a pagamento, DollyV2-7b prodotto dalla società Databricks e disponibile tramite la piattaforma Hugging Face, LLAMA2-7b sviluppato dalla società Meta e disponibile anch’esso tramite la piattaforma Hugging Face, infine Gemini creato da Google ed utilizzato mediante l’invio di richieste HTTP con un approccio basato su *cookie* per l’autenticazione presso il servizio esterno.

Le diverse implementazioni condotte hanno portato a considerazioni preliminari riguardanti le peculiarità dei modelli analizzati. Si è osservato che il modello GPT3.5 dimostra una maggiore stabilità nelle risposte rispetto ai restanti modelli. Tuttavia, Dolly si è rivelato inadeguato per il compito assegnato, mentre LLAMA è risultato capace di rispettare scrupolosamente le richieste ma con limitate capacità di comprensione del testo. Infine, Gemini è stato in grado di comprendere il testo ma ha fornito risposte proliose che non rispecchiavano interamente le richieste. Date le precedenti osservazioni, si è deciso di effettuare un ulteriore approfondimento utilizzando un approccio che combinasse le prestazioni di Gemini e LLAMA, con l’obiettivo di migliorare le prestazioni individuali e confrontarle con il modello più avanzato identificato fino a quel momento, ovvero GPT3.5. Tale approccio è stato realizzato concatenando gli *output* dei due modelli nel seguente modo: prima il testo è stato sottoposto a Gemini per l’estrazione della porzione relativa al KPI richiesto, successivamente l’*output* è stato fornito a LLAMA per l’estrazione del dizionario in formato JSON contenente le informazioni desiderate. In questo modo è stato possibile combinare le potenzialità dei modelli, portando ad un significativo miglioramento rispetto alle alternative *open source*, sebbene non abbia raggiunto il livello di GPT3.5 come sperato.

Nel complesso, l’analisi delle prestazioni dei cinque modelli considerati mostra una varietà di risultati. Il modello GPT3.5 emerge come il più affidabile in termini di sensibilità nell’individuare correttamente i casi positivi, sebbene presenti delle sfide nel controllo dei falsi positivi. Dolly dimostra una sensibilità elevata a discapito di un alto tasso di falsi positivi. Gemini e LLAMA, sia singolarmente che in combinazione, offrono prestazioni moderate, con margini di miglioramento sia per la sensibilità che per il tasso di falsi positivi. L’unione dei modelli sembra rappresentare, però, un miglioramento per le criticità

riscontrate nei singoli modelli. Tale miglioramento suggerisce che l'uso combinato di diversi modelli, insieme ad un'analisi attenta del contesto, possa condurre ad avanzamenti sostanziali per le prestazioni complessive.

L'analisi ha, pertanto, mirato ad individuare le condizioni ottimali per l'impiego di tali strumenti nel contesto aziendale, con l'obiettivo di fornire un valido supporto per successive analisi di mercato e permettendo alle aziende di mantenere una posizione competitiva nel settore. Durante l'implementazione, però, sono state riscontrate alcune debolezze e considerazioni utili da poter studiare ed approfondire per gli sviluppi ed utilizzi futuri.

Inizialmente, si è proceduto con un'analisi della tecnologia GPT4, evidenziando notevoli miglioramenti rispetto al suo predecessore, GPT3.5. Tra le maggiorie riscontrate, si segnala una maggiore conformità alle richieste formulate e la generazione di *output* basati esclusivamente su dizionari JSON, senza alcuna aggiunta di commenti. Tuttavia, l'adozione di GPT4 all'interno della ricerca è stata limitata dal suo elevato costo di utilizzo, rilevato dopo un'analisi preliminare. È importante notare, anche, che la gestione sia della tecnologia ChatGPT che del modello Gemini è affidata esclusivamente ad organizzazioni esterne al contesto alla ricerca, comportando pertanto un'esposizione dei dati immessi a terze parti. Questo sottolinea l'importanza di valutare attentamente la condivisione dei dati ed il livello di riservatezza associato a queste piattaforme. Nel contesto specifico di questa ricerca, i dati utilizzati sono completamente pubblici, conformemente alla direttiva 2022/2464 [1], pertanto non si sono riscontrate restrizioni particolari riguardanti l'accesso e l'utilizzo dei dati.

Come descritto precedentemente nella sezione 4.3, all'interno del presente studio si è richiesto come *output* dei modelli specificatamente il formato JSON. Tale decisione è stata preceduta da studi preliminari che hanno coinvolto anche l'analisi del formato tabellare (CSV). Durante questa fase, però, è emersa una significativa instabilità nei modelli per la produzione di *output* in formato CSV, mentre si è osservata una maggiore affidabilità nell'utilizzo del formato JSON. Questa scelta è stata avvalorata dal fatto che l'azienda cliente, che ha permesso la realizzazione della presente ricerca, ha espresso una preferenza per lo sviluppo dedicato allo strumento GPT ed è stato verificato che la versione utilizzata, GPT-3.5, è stata appositamente addestrata per generare *output* in formato JSON [98]. Di conseguenza, l'utilizzo dei successivi modelli è stato condotto mantenendo lo stesso approccio metodologico.

Nell'attuale panorama della ricerca, si delinea un quadro in cui le aziende clienti, interessate a questa tipologia di lavoro, forniscono una lista di KPI che viene regolarmente aggiornata ogni anno in base alla struttura del questionario ESG utilizzato per la compilazione dei *report*. Un'eventuale evoluzione futura potrebbe consistere nell'esplorare l'adattabilità dei modelli rispetto alle richieste informative. Tale approccio implica che le informazioni specifiche come il nome e il contesto, già presenti nel testo, non vengano fornite in maniera specifica ma venga data un'indicazione generica relativa al KPI da estrarre. Tuttavia, ciò comporta la necessità di gestire situazioni in cui i KPI si presentano con forme e nomenclature molto simili all'interno del testo. A titolo di esempio, durante le fasi iniziali di questa ricerca è stata adottata una tecnica basata sulla similarità

per individuare le pagine potenzialmente contenenti il KPI desiderato. Tuttavia, questo approccio ha portato ad una sovrabbondanza di pagine, rendendo difficile distinguere le informazioni necessarie durante la fase di ricerca ed estrazione dei KPI. Un’alternativa da considerare potrebbe essere quella di far sì che i modelli di Intelligenza Artificiale generativa cercassero direttamente la porzione di testo relativa al KPI. Tuttavia, è necessario tenere in considerazione i costi computazionali, inclusi quelli relativi all’ambiente di elaborazione ed alla capacità di memoria disponibile, nonché i costi effettivi del servizio, specialmente nel caso in cui si ricorra a metodologie soggette a pagamento.

Per quanto riguarda, invece, la creazione di un modello personalizzato in grado di sfruttare le tecnologie già esistenti nel panorama attuale, al fine di combinare le loro potenzialità e creare uno strumento ottimizzato, si è optato per la concatenazione degli output dei modelli **Gemini** e **LLAMA**. Questa scelta mirava ad integrare le capacità di entrambi i modelli (con particolare enfasi sulla capacità di comprensione del testo di **Gemini** e sulla capacità di strutturazione dell’*output* in formato JSON, come richiesto, di **LLAMA**). L’approccio adottato è risultato significativo poiché, durante la fase di visualizzazione dei risultati, presente nella sezione 5.2, si è osservato un miglioramento effettivo rispetto all’utilizzo dei singoli modelli. Tale risultato riveste un’importanza rilevante in quanto dimostra che, combinando le capacità dei modelli *open source* e raffinandone le caratteristiche, è possibile avvicinarsi alle prestazioni di modelli a pagamento, come **GPT3.5**. Attualmente, sebbene **Gemini** sia disponibile tramite un’API REST a pagamento, è stato utilizzato gratuitamente in questo contesto grazie all’impiego di *cookie* per l’autenticazione al servizio. Tuttavia, l’approccio adottato suggerisce che con i modelli attualmente disponibili si possano ottenere risultati comparabili a quelli dei modelli più performanti e disponibili a pagamento. Oltre ad esaminare la concatenazione degli *output* di due modelli, si è investigata, anche, la fattibilità di poter combinare due modelli LLM attraverso un processo di fusione dei *layer* disponibili. Attraverso lo studio condotto nell’articolo intitolato “Merge Large Language Models with mergekit” di Hugging Face [99], è emerso che questa tecnica innovativa consente di combinare i punti di forza di più modelli pre-addestrati in un unico potenziato. Tale tecnica offre la flessibilità di interpolare i modelli attraverso l’implementazione di diverse tecniche illustrate nell’articolo, anche permettendo di determinare a quale modello dar maggior peso rispetto ad altri. Tuttavia, in questa ricerca non è stato possibile sperimentare direttamente questo metodo poiché al momento la piattaforma Hugging Face offre il supporto della libreria **mergekit** solo per modelli presenti sulla piattaforma, con architetture simili ed un egual numero di parametri. Di conseguenza, con i modelli impiegati per questa ricerca, non è stato possibile utilizzare questa tecnica. Tuttavia, per futuri sviluppi, sarebbe auspicabile esplorare le potenzialità di questa libreria attraverso l’implementazione di modelli aggiuntivi verificandone la validità all’interno del contesto di ricerca.

Per arricchire ulteriormente le analisi condotte, sarebbe vantaggioso, inoltre, esaminare l’effetto dell’utilizzo di tecniche avanzate di raffinamento sui modelli. Tra queste, è opportuno considerare il *fine tuning*, già introdotto nella sezione 2.1.3, e la tecnica RAG (*Retrieval Augmented Generation*) [96], discussa nella sezione 4.3. Il *fine-tuning* consiste nell’utilizzare un modello pre-addestrato su un vasto set di dati e adattarlo ad un compito specifico o ad un insieme di dati specifico. Durante questo processo, i pesi del

modello vengono regolati per migliorare le prestazioni su un particolare compito o set di dati, spesso utilizzando un dataset di addestramento annotato. Questa tecnica è utile per personalizzare un modello pre-addestrato e, quindi, soddisfare le esigenze specifiche di un'applicazione o di un problema. D'altra parte, la tecnica RAG permette di arricchire un LLM con un database esterno da cui può recuperare dinamicamente informazioni per influenzare i propri *output*. Questo non solo amplia la conoscenza del modello, ma cambia radicalmente il modo in cui risponde alle richieste. L'approccio implica l'utilizzo di un generatore di dati casuale per produrre diverse varianti di *input*, che vengono poi valutate ed utilizzate per aggiornare il modello, migliorando così la qualità e la diversità dei suoi *output*.

Uno degli ostacoli principali connessi all'utilizzo di tali metodi per il miglioramento dei modelli, risiede nella disponibilità di dati di addestramento che siano rappresentativi ed ampi, indispensabili per garantire la capacità di generalizzazione dei modelli su nuovi compiti e contesti. Nel caso della presente ricerca questo significa reperire manualmente i dati necessari per un grande quantitativo di *report* a disposizione e specifici per le informazioni richieste. Inoltre, la necessità di evitare l'*overfitting* durante il *fine-tuning* richiede un'attenta gestione della complessità del modello ed una selezione oculata delle tecniche di regolarizzazione. La complessità computazionale e la richiesta di risorse *hardware* notevoli, come potenti unità di elaborazione grafica e tempo di calcolo esteso, aggiungono ulteriori sfide, specialmente considerando le dimensioni e la complessità dei modelli in questione.

La ricerca si è, quindi, concentrata sulla valutazione di quattro modelli presenti nel panorama tecnologico, al fine di confrontare alcuni dei più rinomati. Tuttavia, sarebbe auspicabile ampliare ulteriormente lo studio includendo ulteriori modelli e valutando le loro prestazioni rispetto a quelli considerati. Inoltre, sarebbe opportuno considerare l'espansione del set di dati utilizzato per questo specifico caso d'uso, al fine di ottenere una valutazione più completa ed accurata delle qualità dei modelli nel contesto di riferimento. All'interno di questo studio, sono state condotte analisi su ulteriori *report* accessibili tramite la piattaforma Climate Disclosure Platform [74]. Tuttavia, si è riscontrato un problema riguardante i KPI da estrarre, poiché i questionari compilati dalle aziende presentano variazioni nelle informazioni fornite rispetto ai questionari utilizzati all'interno di questa ricerca, ovvero relativo al settore energetico. Di conseguenza, non essendo possibile distinguere le aziende per settore sulla piattaforma ed essendo necessaria la selezione manuale di aziende in grado di soddisfare tali criteri, non è stato possibile completare questa operazione a causa di vincoli legati alle tempistiche del progetto.

In conclusione, il presente studio rivela un'importante potenzialità nell'applicazione della Generative AI all'interno delle pratiche aziendali, specialmente per la comprensione di testi tecnici focalizzati sulla sostenibilità. Questa tecnologia offre un'opportunità per accelerare il processo di reperimento delle informazioni richieste, consentendo alle aziende di adattarsi rapidamente alle mutevoli esigenze del mercato e di mantenere una posizione competitiva rispetto ai concorrenti. Dall'analisi condotta emerge che il modello GPT3.5 si dimostra particolarmente adatto per il compito assegnato, evidenziando un bilanciamiento soddisfacente tra le metriche utilizzate per la valutazione (sezione 5.2). L'efficacia del lavoro svolto è stata confermata anche dall'azienda promotrice di questa ricerca, che

attualmente utilizza tale strumento per generare *benchmark* e condurre analisi interne mirate al miglioramento del rendimento della società. Tuttavia, al fine di ottenere risultati ancora più precisi e significativi, è necessario intraprendere ulteriori studi ed apportare miglioramenti mirati al modello stesso. Questo potrebbe comportare l'ottimizzazione dei parametri di addestramento, l'aggiornamento del *corpus* di dati di riferimento o l'implementazione di tecniche di *fine-tuning* specifiche per il dominio di interesse. Inoltre, un risultato di notevole interesse, emerso dalla presente ricerca, è l'osservazione che la combinazione delle caratteristiche di modelli diversi possa generare una sinergia positiva. Questo suggerisce la possibilità di integrare le specificità di diversi modelli al fine di mitigare eventuali punti deboli individuali e massimizzare le prestazioni globali del sistema. Tale progetto, inoltre, può essere agevolmente standardizzato e, quindi, rapidamente adattato per essere utilizzato da aziende differenti, rendendo più agevole l'espansione dei KPI di interesse e rendendo la procedura adattabile non solo ai dati presenti sulla piattaforma CDP [74], ma ampliando l'utilizzo per altri sistemi informativi. Questo approccio offre prospettive promettenti per lo sviluppo di soluzioni nel campo della Generative AI che sono non solo più efficaci, ma anche più flessibili ed adattabili alle esigenze specifiche dei contesti aziendali e delle sfide affrontate. Pertanto, questo studio non solo evidenzia il valore della Generative AI nell'ambito aziendale, ma offre anche un punto di partenza per ulteriori ricerche ed applicazioni pratiche. Attraverso un'analisi approfondita delle prestazioni dei modelli ed una comprensione più profonda delle loro caratteristiche, sarà possibile sviluppare sistemi sempre più sofisticati ed adattabili, in grado di soddisfare le esigenze specifiche delle aziende e contribuire attivamente alla loro crescita e competitività nel mercato globale.

Glossario

Apache Spark Apache Spark è un sistema di calcolo distribuito *open-source* che fornisce un’interfaccia per la programmazione di interi cluster con parallelismo dei dati implicito e tolleranza ai guasti. È stato sviluppato per affrontare le limitazioni del paradigma di calcolo MapReduce, offrendo prestazioni migliorate attraverso il calcolo in memoria e un’API più versatile..

API REST “Application Programming Interface”, è un insieme di regole e protocolli che consentono a diversi software di comunicare e interagire tra loro. Pertanto, tale strumento permette a diverse applicazioni di scambiare dati e funzionalità in modo efficiente e standardizzato, senza necessariamente conoscere i dettagli interni di ciascuna applicazione coinvolta. In questo modo, le API REST semplificano lo sviluppo software..

Azure Blob Storage Servizio di archiviazione di oggetti scalabile offerto da Microsoft Azure, progettato per archiviare grandi quantità di dati non strutturati, come file multimediali o dati di *backup*..

Azure Databricks Servizio di analisi dei dati basato su Apache Spark. Fornisce un ambiente collaborativo basato sul *cloud* per esplorare, analizzare e condividere grandi quantità di dati..

Azure OpenAI Azure OpenAI è una piattaforma *cloud* fornita da Microsoft che offre servizi e strumenti per l’integrazione e lo sviluppo di soluzioni di intelligenza artificiale, inclusi modelli di linguaggio come GPT (*Generative Pre-trained Transformer*)..

BeautifulSoup Libreria Python per estrarre dati da file HTML e XML. È comunemente utilizzata per il *web scraping*..

benchmark Un insieme di test standardizzati utilizzati per valutare le prestazioni di un sistema o di un componente rispetto a altri sistemi o componenti simili. .

Chrome DevTools Protocol Protocollo di comunicazione a basso livello utilizzato dagli strumenti di sviluppo di Chrome e da altri strumenti di *debug* per comunicare con il browser Chrome o qualsiasi altro browser compatibile. Consente agli sviluppatori di interagire con il browser in modo programmato, controllarne il comportamento, ispezionare e modificare il suo stato e automatizzare compiti..

GitHub Piattaforma web utilizzata per il controllo delle versioni del codice e altri progetti tramite Git..

GPU “Graphics Processing Unit”, è un processore specializzato progettato per accelerare il *rendering* grafico e le operazioni parallele..

Hugging Face Hugging Face è un’azienda e una comunità *open source* che sviluppa e supporta una vasta gamma di strumenti e modelli per il *Natural Language Processing* (NLP). È nota soprattutto per il suo *hub* di modelli, che offre accesso a una vasta collezione di modelli preaddestrati per l’elaborazione del linguaggio naturale. Hugging Face è all’avanguardia nella ricerca e nello sviluppo nel campo del NLP e fornisce risorse essenziali agli sviluppatori e ai ricercatori interessati all’intelligenza artificiale e al *machine learning*..

KPI “Key Performance Indicator”, ovvero indicatore chiave di prestazione. È una misura quantificabile che riflette il successo di un’organizzazione nel raggiungere gli obiettivi strategici e operativi. I KPI vengono utilizzati per valutare le performance di un’azienda, un progetto o un processo, fornendo un quadro chiaro e misurabile del suo stato e dei suoi progressi nel tempo..

MapReduce Modello di programmazione e un’implementazione di elaborazione distribuita sviluppata da Google [100]. Consiste in due fasi principali: la fase di “map” e la fase di “reduce”. Nella fase di “map”, i dati vengono trasformati da una serie di coppie chiave-valore in un’altra serie di coppie chiave-valore. Nella fase di “reduce”, i risultati intermedi della fase di map vengono combinati e aggregati in un unico risultato..

OpenAI OpenAI è una società di ricerca sull’intelligenza artificiale con sede negli Stati Uniti e nota per lo sviluppo di modelli di linguaggio avanzati come GPT (*Generative Pre-trained Transformer*)..

Pandas Pandas è una libreria Python che fornisce strutture dati e strumenti di analisi dati facili da usare. È ampiamente utilizzato per la manipolazione e l’analisi dei dati in Python..

Power BI Power BI è una piattaforma di *business intelligence* (BI) sviluppata da Microsoft. Consente agli utenti di visualizzare e condividere dati in modo interattivo, creare *report* e *dashboard* personalizzati per l’analisi dei dati aziendali..

PyTorch *Framework open source* di *machine learning* utilizzato principalmente per lo sviluppo e l’addestramento di reti neurali profonde. È noto per la sua flessibilità e facilità d’uso, nonché per la sua architettura dinamica che facilita la creazione e la modifica dei modelli. Offre un’ampia gamma di funzionalità per il calcolo scientifico, l’ottimizzazione, la manipolazione dei tensori e l’implementazione di algoritmi di apprendimento automatico..

Sandbox Ambiente di esecuzione isolato e protetto in cui è possibile eseguire software in modo sicuro, limitando l’accesso ai sistemi e alle risorse del computer ospite. Neibrowser *web*, è un meccanismo di sicurezza che limita l’accesso dei processi del browser alle risorse del sistema, riducendo così il rischio di danni causati da *malware* o attacchi informatici..

Scala Scala è un linguaggio di programmazione multiparadigma e staticamente tipizzato, progettato per esprimere i modelli di programmazione comuni in modo chiaro e conciso..

Selenium WebDriver Framework di automazione del *browser web* utilizzato per testare applicazioni *web* automatizzando le interazioni del browser. Consente agli sviluppatori di scrivere script in diversi linguaggi di programmazione, inclusi Python, Java, JavaScript, Ruby, C, per simulare azioni utente come il clic su pulsanti, l'inserimento di dati nei moduli e la navigazione attraverso le pagine *web*. È ampiamente utilizzato per il *testing* automatizzato di applicazioni *web* e per l'automazione di processi *web*..

Vertex AI Vertex AI è una piattaforma di intelligenza artificiale (IA) fornita da Google Cloud che offre un insieme di strumenti e servizi per lo sviluppo, l'addestramento e il deployment di modelli di machine learning in modo semplice e scalabile. In sostanza, Vertex AI semplifica il processo di creazione e gestione di modelli di intelligenza artificiale, consentendo agli sviluppatori di concentrarsi sulle logiche dell'applicazione senza dover gestire l'infrastruttura sottostante..

Bibliografia e sitografia

- [1] Parlamento Europeo e Consiglio dell'Unione Europea. *DIRETTIVA (UE) 2022/2464 DEL PARLAMENTO EUROPEO E DEL CONSIGLIO del 14 dicembre 2022 che modifica il regolamento (UE) n. 537/2014, la direttiva 2004/109/CE, la direttiva 2006/43/CE e la direttiva 2013/34/UE per quanto riguarda la rendicontazione societaria di sostenibilità.* 2022. URL: <https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32022L2464>.
- [2] Gokul Yenduri et al. «Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions». In: *arXiv* (2023). URL: <https://doi.org/10.48550/arXiv.2305.10435>.
- [3] Mike Conover et al. *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*. 2023. URL: <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
- [4] Hugo Touvron et al. «Llama 2: Open Foundation and Fine-Tuned Chat Models». In: *arXiv* (2023). URL: <https://doi.org/10.48550/arXiv.2307.09288>.
- [5] Gemini Team Google et al. «Gemini: A Family of Highly Capable Multimodal Models». In: *arXiv* (2023). URL: <https://doi.org/10.48550/arXiv.2312.11805>.
- [6] Power Reply Srl Site. URL: <https://www.reply.com/power-reply/it/>.
- [7] Ministero dello Sviluppo Economico. *Bilancio Gas Naturale*. 2023. URL: <https://dgsaie.mise.gov.it/bilancio-gas-naturale>.
- [8] Repubblica Italiana. *Decreto Legislativo 23 maggio 2000, n. 164*. 2000. URL: https://www.fire-italia.org/prova/wp-content/uploads/2015/04/D.Lgs_.164-2000.pdf.
- [9] ARERA - Autorità di Regolazione per Energia Reti e Ambiente. *I NUMERI DEI SERVIZI PUBBLICI on line i volumi della Relazione Annuale dell'Autorità. I dati 2022 per elettricità, gas, acqua, rifiuti*. 2023. URL: https://www.arera.it/it/com_stampa/23/230711cs.htm#3.
- [10] Reinalda Start Heleen Ekker. «Meer sterfgevallen door Groningse aardbevingsstress». In: *NOS Nieuws* (2022). URL: <https://nos.nl/collectie/13902/artikel/2432133-meer-sterfgevallen-door-groningse-aardbevingsstress>.
- [11] Jannik Fischbach et al. *Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool*. 2022. arXiv: 2212.06540 [cs.IR]. URL: <https://arxiv.org/pdf/2212.06540.pdf>.
- [12] Tahir Islam et al. «The impact of corporate social responsibility on customer loyalty: The mediating role of corporate reputation, customer satisfaction, and trust». In: *Sustainable Production and Consumption* 25 (2021), pp. 123–135. ISSN: 2352-5509. URL: <https://doi.org/10.1016/j.spc.2020.07.019>.

- [13] Starks L.T. Gillan S.L. Koch A. «Firms and social responsibility: A review of esg and csr research in corporate finance.» In: *Journal of Corporate Finance* (2021), p. 101889. ISSN: 2352-5509. URL: https://www.sustainablebusiness.pitt.edu/sites/default/files/gillan_koch_starks_2020_-_working_paper_-_csr_review_1.pdf.
- [14] Redazione Innolva. «Finanza sostenibile e Rating ESG». In: (2022). URL: <https://www.innolva.it/InSight/investimenti-sostenibili/2022-07--Finanza-sostenibile-e-Rating-ESG>.
- [15] Club of Rome. *The Limits to Growth*. Potomac Associates - Universe Books, 1972. URL: <https://www.clubofrome.org/publication/the-limits-to-growth/>.
- [16] Dichiariazione delle Nazioni Unite. *Dichiariazione di Stoccolma*. 1972. URL: https://www.mase.gov.it/sites/default/files/archivio/allegati/educazione_ambientale/stoccolma.pdf.
- [17] World Commission on Environment e Development. *Our Common Future*. Oxford: Oxford University Press, 1987. URL: <https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf>.
- [18] United Nations Global Compact. «Who Cares Wins: Connecting Financial Markets to a Changing World». In: 2005. URL: https://www.unepfi.org/fileadmin/events/2004/stocks/who_cares_wins_global_compact_2004.pdf.
- [19] Commissione Europea. *Piano d'azione per finanziare la crescita sostenibile della Commissione Europea*. 2018. URL: <https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:52018DC0097&rid=4>.
- [20] Stati membri della Convenzione quadro delle Nazioni Unite. *Accordo di Parigi*. 2015. URL: [https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:22016A1019\(01\)#document1](https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX:22016A1019(01)#document1).
- [21] *Trasformare il nostro mondo: l'Agenda 2030 per lo Sviluppo Sostenibile*. 2015. URL: <https://unric.org/it/wp-content/uploads/sites/3/2019/11/Agenda-2030-Onu-italia.pdf>.
- [22] Alice Orecchio. *ESG Scoring: che cos'è e come si calcola*. FinScience. 2022. URL: <https://finscience.com/blog/esg/esg-scoring-che-cose-e-come-si-calcola/>.
- [23] MSCI ESG Research LLC. *ESG Ratings Methodology*. 2023. URL: <https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology.pdf>.
- [24] *FTSE ESG Index Series Ground Rules*. FTSE Russell. 2023. URL: https://www.lseg.com/content/dam/ftse-russell/en_us/documents/ground-rules/ftse-esg-index-series-ground-rules.pdf.
- [25] Nexio Projects. *The Ultimate Guide to Your EcoVadis Assessment*. 2022. URL: https://info.nexioprojects.com/hubfs/Ebooks/The_Ultimate_Guide_To_Your_EcoVadis_Assessment_2022_Update_NexioProjects.pdf?hsLang=en.

- [26] *Environmental, Social and Governance Scores from Refinitiv*. Refinitiv. 2022. URL: https://www.lseg.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf.
- [27] *Guide to Standard Ethics Rating - 2020*. Standard Ethics. 2020. URL: https://www.standardethics.eu/images/1._Sustainability_Rating_definitions_Guide_2020_final.pdf.
- [28] Leite I. et al. Vinuesa R. Azizpour H. «The role of artificial intelligence in achieving the Sustainable Development Goals.» In: *Nat Commun* 11 (2020). URL: <https://doi.org/10.1038/s41467-019-14108-y>.
- [29] Ross Girshick Joseph Redmon Santosh Divvala e Ali Farhadi. «You Only Look Once: Unified, Real-Time Object Detection». In: *arXiv* (2016). URL: <https://arxiv.org/pdf/1506.02640v5.pdf>.
- [30] Wei Liu et al. «SSD: Single Shot MultiBox Detector». In: *ECCV* (2016). URL: <https://arxiv.org/pdf/1512.02325.pdf>.
- [31] Shadrin D. et al. Koldasbayeva D. Tregubova P. «Large-scale forecasting of Heracleum sosnowskyi habitat suitability under the climate change on publicly available data.» In: *Sci Rep* 12 (2022). URL: <https://doi.org/10.1038/s41598-022-09953-9>.
- [32] Sarah L. Gilhespy et al. «First 20 years of DNDC (DeNitrification DeComposition): Model evolution». In: *Ecological Modelling* 292 (2014), pp. 51–62. URL: <https://doi.org/10.1016/j.ecolmodel.2014.09.004>.
- [33] Evgeny Burnaev et al. «Practical AI Cases for Solving ESG Challenges». In: *Sustainability* 15.17 (2023). URL: <https://www.mdpi.com/2071-1050/15/17/12731>.
- [34] Efrat A. Krechetov M. Esmaieeli Sikaroudi A.M. e et al. «Prediction and Prevention of Pandemics via Graphical Model Inference and Convex Programming». In: *Sci Rep* 12 (2022), p. 7599. URL: <https://doi.org/10.1038/s41598-022-11705-8>.
- [35] Amundi Asset Management. *Artificial Intelligence Solutions to Support Environmental, Social, and Governance Integration in Emerging Markets*. 2022. URL: <https://research-center.amundi.com/article/artificial-intelligence-solutions-support-environmental-social-and-governance-integration-emerging>.
- [36] WMP van der Aalst, M. Bichler e A. Heinzl. «Automazione dei processi robotici». In: *Bus Inf Syst Eng* 60 (2018), pp. 269–272. URL: <https://doi.org/10.1007/s12599-018-0542-4>.
- [37] European Banking Authority. *Eba report on management and supervision of ESG risks for credit institutions and investment firms*. 2021. URL: <https://www.eba.europa.eu/sites/default/documents/files/document-library/Publications/Reports/2021/1015656/EBA%20Report%20on%20ESG%20risks%20management%20and%20supervision.pdf>.

- [38] Energy Efficient Mortgages Initiative. *TranspArEEEns: portare una nuova prospettiva all'EEMI ponendo l'accesso delle PMI ai finanziamenti al centro di un ecosistema ESG*. European Mortgage Federation- European Covered Bond Council. 2021. URL: <https://energyefficientmortgages.eu/transpareens-bringing-a-new-perspective-to-the-eemi-by-putting-sme-access-to-finance-at-the-heart-of-an-esg-ecosystem/>.
- [39] M. Minkkinen, A. Niukkanen e M. Mäntymäki. «What about investors? ESG analyses as tools for ethics-based AI auditing». In: *AI & Soc* (2022). DOI: 10.1007/s00146-022-01415-0.
- [40] A. Jobin, M. Ienca e E. Vayena. «The global landscape of AI ethics guidelines». In: *Nat Mach Intell* 1 (2019), pp. 389–399. URL: 10.1038/s42256-019-0088-2.
- [41] V.-D. Păvăloaia e S.-C. Necula. «Artificial Intelligence as a Disruptive Technology—A Systematic Literature Review». In: *Electronics* 12 (2023). URL: 10.3390/electronics12051102.
- [42] J. Nagler, J. van den Hoven e D. Helbing. «An Extension of Asimov's Robotics Laws». In: *Towards Digital Enlightenment*. A cura di D. Helbing. Springer, Cham, 2019. URL: 10.1007/978-3-319-90869-4_5.
- [43] Xinyi Zhou e Reza Zafarani. «A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities». In: *ACM Comput. Surv.* 53.5 (2020). URL: <https://doi.org/10.1145/3395046>.
- [44] B. Kim et al. «A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions». In: *PLoS ONE* 16.12 (2021). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260080>.
- [45] M. Westerlund. *The Emergence of Deepfake Technology: A Review*. URL: <https://timreview.ca/article/1282>.
- [46] R. Wang et al. «Fakespotter: A simple baseline for spotting ai-synthesized fake faces». In: *arXiv* (2019). URL: <https://arxiv.org/abs/1909.06122>.
- [47] S. Barannikov et al. «Representation Topology Divergence: A Method for Comparing Neural Network Representations». In: *Proceedings of the Thirty-ninth International Conference on Machine Learning (ICML)*. URL: <https://arxiv.org/abs/2201.00058>.
- [48] B. Lund e T. Wang. «Chatting about ChatGPT: How may AI and GPT impact academia and libraries?» In: *Libr. Hi Tech News* 40 (2023). URL: <https://www.emerald.com/insight/content/doi/10.1108/LHTN-01-2023-0009/full/html#abstract>.
- [49] Klaić Aleksandar. *Overview of the state and trends in the contemporary information security policy and information security management methodologies*. Rapp. tecn. 2010, pp. 1203–1208. URL: https://ieeexplore.ieee.org/abstract/document/5533647?casa_token=nTxIDSjnGTUAAAAAA:Ulj5GBV25fVR_AtBUF9tLT7oH1PkjlSDL2kbp7EQ9r8b9eaX5VoNCK_hJBViEb3MDGW--nszqxc.
- [50] *ChatGPT Enterprise*. OpenAI. 2023. URL: <https://openai.com/enterprise>.

- [51] N. Jones. «How to stop data centres from gobbling up the world's electricity». In: *Nature* 561 (2018). URL: <https://complexityexplorer.s3.amazonaws.com/Computation+in+CS/SFI+1.4b.pdf>.
- [52] CNET. *Why Tech Pollution's Going Global*. URL: <https://www.cnet.com/tech/tech-industry/why-tech-pollution-going-global/>.
- [53] J. Gusak et al. «Survey on Efficient Training of Large Neural Networks». In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*. URL: <https://www.ijcai.org/proceedings/2022/0769.pdf>.
- [54] Commissione Europea. *Proposta di regolamento del Parlamento Europeo e del Consiglio che stabilisce regole armonizzate sull'intelligenza artificiale (legge sull'intelligenza artificiale) e modifica alcuni atti legislativi dell'unione*. COM COM(2021) 206 final. Commissione Europea, 21 apr. 2021. URL: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0006.02/DOC_1&format=PDF.
- [55] Jaume Duch Guillot. «EU AI Act: first regulation on artificial intelligence». In: (8 giu. 2023). URL: <https://www.europarl.europa.eu/news/en/headlines/society/20230601ST093804/eu-ai-act-first-regulation-on-artificial-intelligence>.
- [56] Milena Di Nenno. «AI Act: l'UE pioniera nella regolamentazione dell'intelligenza artificiale». In: *Affari Internazionali* (2023). URL: <https://www.affarinternazionali.it/ai-act-ue-pioniera-regolamentazione-intelligenza-artificiale/>.
- [57] CED (Committee for Economic Development). *Policy Brief: Proposal for a US-EU AI Code of Conduct*. 2023. URL: <https://www.ced.org/reports/policy-brief-proposal-for-a-us-eu-ai-code-of-conduct>.
- [58] M. Eisenstein. «AI-enhanced protein design makes proteins that have never existed». In: *Nat Biotechnol* (2023). URL: <https://doi.org/10.1038/s41587-023-01705-y>.
- [59] Elizabeth Bell. «IA generativa e modelli di linguaggio di grandi dimensioni: Qual è la differenza?» In: *Appian* (2023). URL: <https://appian.com/it/blog/acp/process-automation/generative-ai-vs-large-language-models.html>.
- [60] Nitish Mittal, Nisha Krishan e Vaani Sharma. *Generative AI – Revolutionizing the Creative Design and Development Process*. Everest Group. 2023. URL: <https://www2.everestgrp.com/reportaction/EGR-2023-64-R-5941/Marketing>.
- [61] Alan M Turing. «Computing machinery and intelligence». In: *Mind* 59.236 (1950), pp. 433–460.

- [62] Tomas Mikolov et al. «Distributed Representations of Words and Phrases and their Compositionality». In: *Advances in Neural Information Processing Systems* 26 (2013), pp. 3111–3119. URL: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>.
- [63] Ralf C. Staudemeyer e Eric Rothstein Morris. «Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks». In: *arXiv* (2019). URL: <https://arxiv.org/abs/1909.09586>.
- [64] Junyoung Chung et al. «Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling». In: *arXiv preprint arXiv:1412.3555* (2014). URL: <https://arxiv.org/abs/1412.3555>.
- [65] Ilya Sutskever, Oriol Vinyals e Quoc V. Le. «Sequence to Sequence Learning with Neural Networks». In: *arXiv* (2014). URL: <https://arxiv.org/abs/1409.3215>.
- [66] Long Ouyang et al. «Training language models to follow instructions with human feedback». In: *arXiv* (2022). DOI: 10.48550/arXiv.2203.02155. URL: <https://arxiv.org/abs/2203.02155>.
- [67] Colin Raffel et al. «Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer». In: *arXiv* (2019). DOI: 10.48550/arXiv.1910.10683. URL: <https://arxiv.org/abs/1910.10683>.
- [68] Zhilin Yang et al. «XLNet: Generalized Autoregressive Pretraining for Language Understanding». In: *arXiv* (2019). DOI: 10.48550/arXiv.1906.08237. URL: <https://arxiv.org/abs/1906.08237>.
- [69] Jacob Devlin et al. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». In: *arXiv* (2018). URL: <https://arxiv.org/abs/1810.04805>.
- [70] Yinhan Liu et al. «RoBERTa: A Robustly Optimized BERT Pretraining Approach». In: *arXiv* (2019). [Submitted on 26 Jul 2019]. DOI: 10.48550/arXiv.1907.11692. URL: <https://arxiv.org/abs/1907.11692>.
- [71] Zhenzhong Lan et al. «ALBERT: A Lite BERT for Self-supervised Learning of Language Representations». In: *arXiv* (2019). DOI: 10.48550/arXiv.1909.11942. URL: <https://arxiv.org/abs/1909.11942>.
- [72] Alec Radford et al. «Learning Transferable Visual Models From Natural Language Supervision». In: *arXiv* (2021). DOI: 10.48550/arXiv.2103.00020. URL: <https://arxiv.org/abs/2103.00020>.
- [73] Giuliano Guerriero. *Vantaggi e Rischi del Large Language Model nel Cloud - Palo Alto Networks*. Florence Consulting Group. URL: <https://cio.florence-consulting.it/palo-alto-networks-vantaggi-rischi-llm-cloud>.
- [74] *Carbon Disclosure Project Site*. URL: <https://www.cdp.net/en/>.
- [75] Ashish Vaswani et al. *Attention Is All You Need*. 2017. URL: <https://arxiv.org/pdf/1706.03762.pdf>.

- [76] Alec Radford et al. «Improving Language Understanding by Generative Pre-Training». In: *OpenAI* (2018). URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf?ref=diariodunanalista.it.
- [77] Denis Paperno et al. «The LAMBADA dataset: Word prediction requiring a broad discourse context». In: *arXiv* (2016). URL: <https://arxiv.org/abs/1606.06031>.
- [78] Alex Wang et al. «GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding». In: *arXiv* (2018). URL: <https://arxiv.org/abs/1804.07461>.
- [79] Pranav Rajpurkar et al. «SQuAD: 100,000+ Questions for Machine Comprehension of Text». In: *arXiv* (2016). URL: <https://arxiv.org/abs/1606.05250>.
- [80] Paul Christiano et al. «Deep reinforcement learning from human preferences». In: *arXiv* (2017). URL: <https://arxiv.org/abs/1706.03741>.
- [81] Stephanie Lin, Jacob Hilton e Owain Evans. «TruthfulQA: Measuring How Models Mimic Human Falsehoods». In: *arXiv* (2021). URL: <https://arxiv.org/abs/2109.07958>.
- [82] Long Ouyang et al. «Training language models to follow instructions with human feedback». In: *arXiv* (2022). URL: <https://arxiv.org/pdf/2203.02155.pdf>.
- [83] Stella Biderman et al. «Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling». In: *arXiv* (2023). URL: <https://arxiv.org/pdf/2304.01373.pdf>.
- [84] Jason Phang et al. «EleutherAI: Going Beyond "Open Science" to "Science in the Open"». In: (2022). URL: <https://arxiv.org/pdf/2210.06413.pdf>.
- [85] T. Brown et al. «Language models are few-shot learners». In: *Advances in Neural Information Processing Systems*. 2020.
- [86] T. Dao et al. «Flash Attention: Fast and memory-efficient exact attention with I/O-awareness». In: *arXiv* (2022).
- [87] Biao Zhang e Rico Sennrich. «Root Mean Square Layer Normalization». In: *arXiv* (2019).
- [88] Jianlin Su et al. «RoFormer: Enhanced Transformer with Rotary Position Embedding». In: *arXiv* (2021).
- [89] Ilya Loshchilov e Frank Hutter. «Decoupled Weight Decay Regularization». In: *arXiv* (2017).
- [90] Joshua Ainslie et al. «GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints». In: *Google Research* (2023). URL: <https://arxiv.org/pdf/2305.13245.pdf>.
- [91] Thibault Lahire. «Importance Sampling for Stochastic Gradient Descent in Deep Neural Networks». In: *arXiv* (2023). URL: <https://arxiv.org/abs/2303.16529>.

- [92] John Schulman et al. «Proximal Policy Optimization Algorithms». In: *arXiv* (2017). URL: <https://arxiv.org/pdf/1707.06347.pdf>.
- [93] Romal Thoppilan et al. «LaMDA: Language Models for Dialog Applications». In: *arXiv* (2022). URL: <https://arxiv.org/abs/2201.08239>.
- [94] Check Point Research. *Lowering the Bar(d)? Check Point Research's security analysis spurs concerns over Google Bard's limitations*. Lug. 2023. URL: <https://research.checkpoint.com/2023/lowering-the-bard/>.
- [95] Daniel Park. *Bard-API*. 2023. URL: <https://github.com/dsdanielpark/Bard-API?tab=readme-ov-file>.
- [96] Yunfan Gao et al. «Retrieval-Augmented Generation for Large Language Models: A Survey». In: *arXiv* (2023). URL: <https://arxiv.org/pdf/2312.10997.pdf>.
- [97] Edward Beeching et al. *Open LLM Leaderboard*. 2023. URL: https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.
- [98] *Function calling - OpenAI Documentation*. URL: <https://platform.openai.com/docs/guides/function-calling>.
- [99] Maxime Labonne. *Merge Large Language Models with mergekit*. 2024. URL: <https://huggingface.co/blog/mlabonne/merge-models#merge-large-language-models-with-mergekit>.
- [100] Jeffrey Dean e Sanjay Ghemawat. «MapReduce: Simplified Data Processing on Large Clusters». In: *Communications of the ACM* (2008). URL: <https://static.googleusercontent.com/media/research.google.com/it/archive/mapreduce-osdi04.pdf>.
- [101] Jannik Fischbach et al. «Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool». In: *ArXiv* abs/2212.06540 (2022). URL: <https://arxiv.org/pdf/2212.06540.pdf>.
- [102] Jeffrey Dean e Sanjay Ghemawat. «MapReduce: Simplified Data Processing on Large Clusters». In: *Commun. ACM* 51 (2008), pp. 107–113. URL: <https://doi.org/10.1145/1327452.1327492>.
- [103] CRIF. *ESG Outlook CRIF: Sostenibilità per Imprese, Individui e Immobili 2023*. 2023. URL: <https://www.crif.it/ricerche-e-academy/ricerche/market-outlook/esg-outlook-crif-sostenibilita-imprese-individui-immobili-2023/> (visitato il 26/09/2023).