UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA
Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica e dell'Automazione

TESI DI LAUREA

Progettazione e implementazione di case study per la valutazione di tool di Intelligenza Artificiale Generativa

Design and implementation of case studies for the evaluation of Generative Artificial Intelligence tools

Relatore

Prof. Domenico Ursino

Candidata

Monique Lorenza Ngantchou Tchakounte

ANNO ACCADEMICO 2025-2026

La sapienza è figliola della sperienza.
Leonardo da Vinci

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Negli ultimi anni il settore dell'Intelligenza Artificiale ha conosciuto una crescita rapida, trasformando ogni giorno diversi settori di ricerca industriale e la vita quotidiana di tutti noi. In particolare, l'avvento dell'Intelligenza Artificiale Generativa è stato un importante punto di svolta, permettendo ai sistemi di creare autonomamente contenuti originali di tipo testuale, immagini, codice sorgente, audio e video partendo da semplici istruzioni espresse in linguaggio naturale. Con i progressi ottenuti nell'ambito del Deep Learning, dei Large Language Model e dei modelli di diffusione, questi sistemi possono produrre output di qualità elevata, e dunque essere applicati in numerosi contesti professionali e scientifici.

Parallelamente, alla diffusione rapida di tali tecnologie c'è stata anche una crescita nell'interesse della comunità scientifica verso la comprensione delle loro potenzialità e dei loro limiti. Anche se i modelli generativi moderni hanno raggiunto prestazioni notevoli in termini di qualità dei contenuti prodotti, essi presentano ancora diversi limiti, come l'incoerenza al livello delle risposte fornite, la difficoltà nel rispettare vincoli complessi e l'elevato costo computazionale. Di conseguenza, sarebbe fondamentale analizzare sia le capacità di questi modelli che i contesti nei quali essi mostrano limitazioni o richiedono un utilizzo consapevole.

La presente tesi fornisce una panoramica dell'avanzamento dell'Intelligenza Artificiale Generativa e analizza il funzionamento attraverso lo studio di diverse categorie di modelli che costituiscono la base dell'Intelligenza Artificiale Generativa. Successivamente, il lavoro si concentra sull'analisi sperimentale di tre differenti tipologie di modelli generativi, appartenenti a domini applicativi distinti: la generazione di testo, la generazione di immagini e la generazione di video. Per ciascun modello vengono progettati diversi casi di studio con l'obiettivo di valutarne il comportamento rispetto a prompt di diversa complessità; successivamente vengono analizzati la qualità dei risultati ottenuti, il grado di aderenza alle istruzioni fornite e le principali limitazioni osservate durante la generazione.

Nel nostro studio abbiamo usato DeepSeek, per la generazione di testo, DALL·E, per la generazione di immagini a partire da descrizioni testuali, e ModifAI, impiegato per la generazione automatica di video mediante tecnologia Text-to-Video. La scelta di questi sistemi consente di confrontare differenti approcci alla generazione multimodale e di evidenziare le specificità, i punti di forza e le criticità proprie di ciascun ambito applicativo. Abbiamo strutturato il lavoro in sei capitoli:

- Nel Capitolo 1 saranno introdotti i concetti fondamentali dell'Intelligenza Artificiale in generale e poi dell'Intelligenza Artificiale Generativa. Dopo una panoramica storica sull'evoluzione della disciplina, vengono descritti i principi di funzionamento dei principali sistemi generativi, le architetture maggiormente utilizzate e le principali criticità che caratterizzano questa categoria di modelli.

- Il Capitolo 2 è dedicato allo di studio di OpenAI e al suo contributo nello sviluppo dell'Intelligenza Artificiale Generativa. Vengono analizzate la nascita dell'azienda, la sua evoluzione organizzativa, i principali prodotti sviluppati, con particolare attenzione a diversi modelli, come Dactyl, DALL-E, Sora, e viene inoltre presentato DeepSeek, quale esempio significativo della rapida crescita dell'ecosistema dell'Intelligenza Artificiale in Cina.
- Il Capitolo 3 presenta il primo caso di studio dedicato alla valutazione di DeepSeek nella generazione di testi. Dopo un'introduzione ai principi della generazione di testo e al Prompt Engineering, vengono progettati diversi esperimenti finalizzati a valutare la capacità del modello di comprendere e soddisfare prompt caratterizzati da differenti livelli di complessità.
- Il Capitolo 4 è dedicato alla valutazione di DALL·E. Dopo aver introdotto il paradigma Text-to-Image, vengono analizzati il funzionamento del sistema e la sua interfaccia, per poi presentare una serie di esperimenti finalizzati a valutare la qualità delle immagini generate e la capacità del modello di interpretare correttamente richieste contenenti vincoli multipli.
- Nel Capitolo 5 sarà analizzata la generazione automatica di video mediante ModifAI. Dopo aver illustrato i principi fondamentali della tecnologia Text-to-Video e le principali tecniche di generazione video, viene sviluppato un caso di studio basato su prompt di complessità crescente. L'analisi dei risultati fatta consente di evidenziare le capacità del modello nella gestione della coerenza temporale, del rispetto dei vincoli semantici e della qualità complessiva dei contenuti generati, mettendone in luce anche le attuali limitazioni.
- Nel Capitolo 6 viene proposta una discussione sui punti di forza e di debolezza dei sistemi esaminati.
- Nel Capitolo 7 verranno tratte le conclusioni in merito al lavoro svolto.

CAPITOLO 1

Introduzione all'Intelligenza Artificiale

1.1 Introduzione all'Intelligenza Artificiale

1.1.1 Definizione e principi fondamentali dell'Intelligenza Artificiale

L'Intelligenza Artificiale può essere definita come la creazione di macchine che esibiscono capacità di pensiero e apprendimento paragonabili a quelle umane. Questa definizione, radicata nella ricerca fondamentale di scienziati come Alan Turing, che nel suo articolo *Computing Machinery and Intelligence* pose le basi concettuali per la simulazione dell'intelligenza umana nelle macchine, rimane una pietra miliare nel campo. La Figura 1.1 rappresenta chiaramente i vari strati di cui è costituita l'Intelligenza Artificiale.

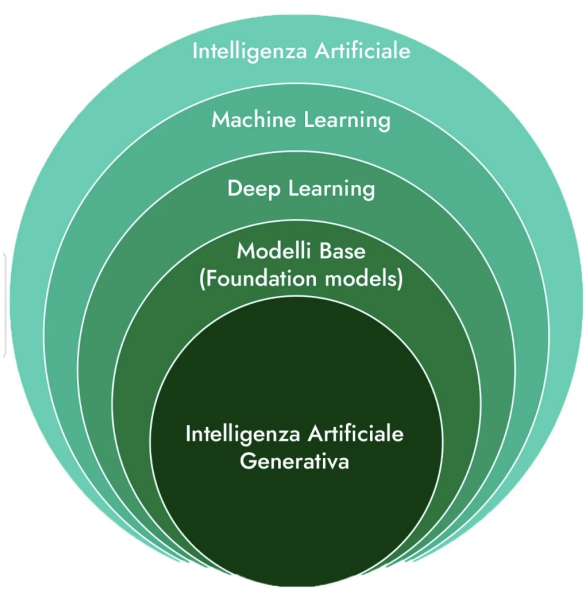


Figura 1.1: I principali strati dell'Intelligenza Artificiale

Proviamo di capire in modo più approfondito i componenti di ogni livello.

- Machine Learning(MC): un ramo dell'Intelligenza Artificiale in cui i sistemi informatici imparano a svolgere compiti a partire dai dati, senza essere programmati esplicitamente

con regole rigide. Gli algoritmi di ML utilizzano i dati storici come input per prevedere nuovi valori di output.

- Deep Learning(DL): il Deep Learning è un sottoinsieme del ML che utilizza reti neurali artificiali con molti strati per apprendere rappresentazioni sempre più complesse dei dati.
- Foundation model: un foundation model è un modello di Deep Learning addestrato su grandi dataset non specifici, che può essere poi adattato (fine-tuning[1] o prompting[2]) a numerose applicazioni.
- Intelligenza Artificiale Generativa: la Generative AI è una branca dell'Intelligenza Artificiale che riguarda i modelli in grado di creare nuovi contenuti originali a partire dai dati su cui sono stati addestrati.

Valutazione dell'Intelligenza Artficiale: il test di Turing

Il test di Turing fu proposto nel 1950 da Alan Turing ed è stato concepito per fornire una definizione più completa e operativa dell'intelligenza. Turing ha suggerito un test basato sull'impossibilità di distinguere un calcolatore dalle entità che lo sono indubbiamente, gli esseri umani. Secondo lui, una macchina è intelligente se riesce a imitare un essere umano in una conversazione al punto da non essere riconosciuta. I tre partecipanti di questo test sono: il computer, il giudice umano e un umano. Il computer supererà il test se un esaminatore umano, dopo avere posto alcune domande in forma scritta, non sarà in grado di capire se le risposte provengono da una persona o dal computer. La Figura 1.2 ricapitola il ragionamento dietro il test di Turing.

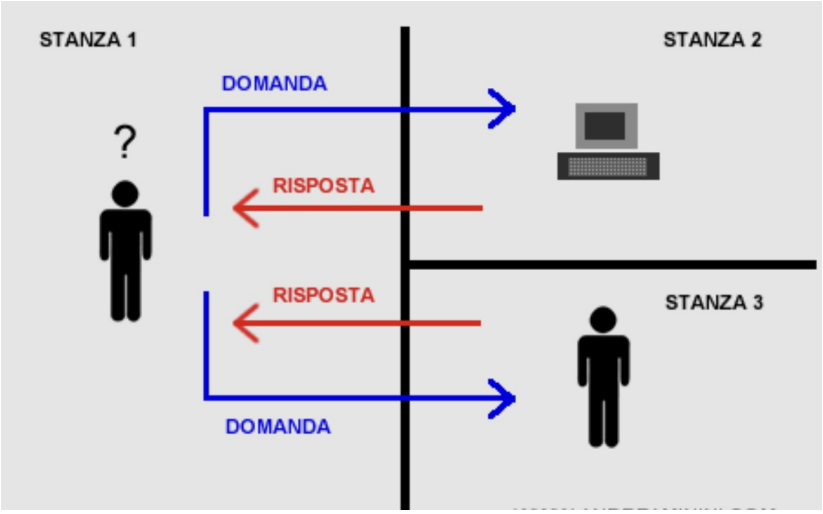


Figura 1.2: Test di Alan Turing

Notiamo che programmare una macchina in grado di superare il test richiede un notevole lavoro. Il computer dovrebbe possedere le seguenti capacità:

- interpretazione del linguaggio naturale per comunicare con l'esaminatore nel suo linguaggio umano;

[1] Il fine-tuning è una tecnica di addestramento in cui un modello d'Intelligenza Artificiale già pre-addestrato viene ulteriormente addestrato su un dataset specifico per adattarlo a un compito particolare.

[2] Il prompting è la tecnica che consiste nel fornire istruzioni o esempi direttamente in input al modello, senza modificarne i parametri, per guidarne il comportamento.

- rappresentazione della conoscenza per memorizzare quello che sa o sente;
- ragionamento automatico per utilizzare la conoscenza memorizzata in modo da rispondere alle domande e trarre nuove conclusioni;
- apprendimento per adattarsi a nuove circostanze, individuare ed esplorare pattern.

Per gli ultimi cinquanta anni, il test di Turing è rimasto molto influente ma anche criticato. Inanzitutto, esso misura l'imitazione del comportamento umano, non la vera comprensione. Inoltre, una macchina può ingannare senza capire davvero. Oltre tutto questo test non valuta altre forme di intelligenza, ad esempio, l'intelligenza visiva o quella logica.

Benché il test va rimasto significativo, i ricercatori non hanno dedicato molto sforzi al tentativo di costruire un sistema che potesse passare il test di Turing. Secondo loro era più importante studiare i principi alla base dell'Intelligenza Artificiale che dedicarsi alla creazione di un duplicato.

L'approccio cognitiva

Per confrontare il ragionamento di un programma con quello di un essere umano dobbiamo, prima di tutto, capire come funziona il nostro cervello per poter formulare una teoria della mente sufficiente, ed esprimerla sotto forma di un programma per calcolatori. Determinare il ragionamento del nostro cervello può essere fatto in due modi, ovvera facendo un'introspezione oppure una sperimentazione psicologica. Se il comportamento del software per quanto riguarda il suo input/output e la temporizzazione corrisponde a quella di una persona, potrebbe essere una prova che alcuni dei meccanismi del programma operano anche negli esseri umani. In conclusione, il campo delle scienze cognitive unisce modelli per computer sviluppati dall'Intelligenza Artificiale e tecniche di sperimentazione psicologica nel tentativo di costruire teorie precise e verificate sul funzionamento della mente umana.

Approccio basato sulla logica del pensiero

Aristotele fu uno dei primi scienziati a realmente cercare di codificare formalmente il pensiero corretto fornendo pattern di deduzione che portano sempre a conclusioni corrette quando sono corrette le premesse. Queste leggi del pensiero governano il funzionamento della mente e il loro studio ha dato origine alla disciplina chiamata *logica*.

Ultimamente ricerche nel mondo della logica hanno sviluppato una notazione precisa per formulare proposizioni riguardanti tutti gli aspetti del mondo e le relazioni tra essi. La tradizione logistica, come viene chiamata nel dominio dell'IA, spera di partire da programmi siffatti per costruire sistemi intelligenti.

Sarebbe importante notare che questo approccio è delicato perché non è facile esprimere una conoscenza non formalizzata nei termini strettamente formali richiesti della notazione logica sopratutto quando tale conoscenza non è sicura al cento percento. Un'altro punto di debolezza è che c'è una un grande differenza tra essere in grado di risolvere un problema "in linea di principio" e farlo nella pratica. Benché tutte queste debolezze, si possano applicare a qualsiasi tentativo di costruire sistemi in grado di ragionare, esse sono apparsi inizialmente nella tradizione logica.

Agire razionalmente: l'approccio degli agenti razionali

Un agente razionale sceglie l'azione migliore per massimizzare il raggiungimento dei suoi obbiettivi, in base alle informazioni disponibili.

L'approccio basato sulle "leggi del pensiero" è focalizzato sul fatto di essere in grado di formulare deduzioni corrette, sapendo che una data azione porterà al soddisfacimento dei propri obbiettivi, e agire, quindi, in tal senso.

Tutte le abilita richieste dal test di Turing hanno lo scopo di agire razionalmente. Dobbiamo poter rappresentare la conoscenza e applicarla ad un ragionamento. E più che altro generare frasi comprensibili nel linguaggio naturale. L'apprendimento ci serve per avere una buona idea del funzionamento del mondo. permettendoci di generare strategie più efficaci per interagire con esso.

Per tutti questi motivi, lo studio dell'IA come progettazione di agenti razionali presenta almeno due vantaggi. Prima di tutto, è molto più generale rispetto all'approccio basato sulle "leggi del pensiero", poiché il coretto uso dell'inferenza è solo uno di molteplici meccanismi utilizzabili per arrivare alla razionalità. Inoltre, esso si presta meglio a recepire gli sviluppi scientifici rispetto agli approcci basati sul comportamento o sul pensiero umano.

1.1.2 La storia dell'Intelligenza Artificiale

Gestazione dell'Intelligenza Artificiale (1943-1955)

Nell'anno 1943 sono state condotte le prime ricerche al riguardo dell'IA da Warren McCulloch e Walter Pitts. I due proposero un modello matematico di neurone artificiale che riceve input, li somma e produce un output che potrebbe essere 0 o 1 ("acceso" o "spento"). Gli autori hanno basato i loro studi su tre principali fronti:

- *filosofia e funzione dei neuroni nel cervello;*
- *la teoria della computazione di Turing;*
- *analisi formale della logica proposizionale di Rusell e Whitehead.*

Approfondendo le ricerche loro dimostrano che, ogni funzione computabile poteva essere calcolata da una rete di neuroni collegati e che tutti gli operatori logici (and, or, not, etc) potevano essere implementati con semplici strutture a rete. Suggerirono, inoltre, che reti neurali adeguatamente definite potessero essere capaci di apprendere.

Nel 1949, Donald Hebb formulò un aggiornamento per la modifica dei pesi delle connessioni tra i neuroni. La sua teoria, chiamata *apprendimento Hebbiano*, è fondata sul fatto che i neuroni si attivano insieme e si rafforzano insieme. In altre parole, se due neuroni si attivano nello stesso momento simultaneamente, allora aumenta il peso della connessione tra di loro.

Nel 1951 Marvin Minsky e Dean Edmonds due dottorandi del dipartimento di matematica di Princeton, costruirono il primo computer basato su rete neurale. Lo SNARC, com'era chiamato, utilizzava 3000 tubi a vuoto e un sistema automatico di pilotaggio riciclato da un bombadiere B-24 per simulare una rete di 40 neuroni. La commissione di dottorato di Minsky non era convinta che questo tipo di lavoro si potesse considerare come matematica, ma sembra che Von Neumann abbia ribattuto: "Se non lo è ora, lo sarà un giorno". Minsky più tardi avrebbe dimostrato teoremi importanti circa la limitazioni della ricerca sulle reti neurali.

Con l'avanzamento del tempo ci sono stati diversi altri esempi di lavori che possono essere considerati come Intelligenza artificiale, ma una visione completa dell'IA fu espressa la prima volta da Alan Turing. Come discusso prima, l'articolo introduceva il test di Turing, l'apprendimento automatico, gli algoritmi genetici e l'apprendimento per rinforzo.

Nascita dell'Intelligenza Artificiale (1956)

A Princeton si trovava John McCarthy un'altra figura importante per l'IA. McCarthy insieme a Minsky, Claude Shanon e Nathaniel Richester, hanno riunito ricercatori americani

interessati alla teoria degli automi, alle reti neurali e allo studio dell'intelligenza. Un workshop di due mesi è stato organizzato durante l'estate del 1956 a Darmouth.

Rubarono la scena due ricercatori della Carnegie Tech, Allen Newell e Herbert Simon che presentarano un programma capace di ragionare, il Logic Theorist (LT), con l'obbiettivo di dimostrare che una macchina può ragionare in modo logico. Durante l'evento Simon ha affermato "abbiamo inventato un programma per computer capace di pensare in modo non numerico, e abbiamo quindi risolto l'antichissimo problema mente-corpo". Subito dopo il workshop, il programma fu in grado di dimostrare la maggior parte dei teoremi nel secondo capitolo dei *Principia Matematica* di Russel e Whitehead. Il workshop di Dartmouth non portò a particolari innovazioni, ma serve a far incontrare tutti i protagonisti principali della disciplina.

Quando leggiamo *la proposal* per il workshop di Darthmouth possiamo capire perché era necessario che l'IA diventasse una branca scientifica a venanti. Inoltre, perché tutto il lavoro di ricerca nel campo dell'IA non avrebbe potuto essere svolto come parte della teoria del controllo, della ricerca operativa o della teoria delle decisioni, che alla fine hanno obbiettivi comuni? E anche, perché l'IA non fa semplicemente parte della matematica? La prima risposta a tutte queste domande è che fin dall'inizio l'idea fondamentale è stata di riprodurre facoltà umane, come la creatività, il miglioramento di sé e l'uso del linguaggio, e nessuna delle discipline citate trattava questi argomenti. La seconda risposta sta nella metodologia: l'IA è l'unica scienza che si propone di costruire macchine che funzionino autonomamente in ambienti complessi e mutevoli.

Era dei primi entusiasmi e grandi aspettative (1952-1969)

I primi anni dell'IA furono un vero successo sopratutto per il fatto che all'epoca l'informatica in sè era ancora molto primitiva e gli intellettuali, per la maggiore parte, pensavano che una macchina non potrà mai fare X. Naturalmente, i ricercatori dell'IA risposero dimostrando una X dietro l'altra.

Dopo i primi successi di Newell e Simon arrivò l'innovazione del General Problem Solver (GPS). La progettazione di questo programma si è basata nel principio di imitare i procedimenti umani di risoluzione dei problemi. Ancora esisteva una classe limitata di problemi che il GPS non poteva risolvere; però l'ordine in cui il programma considerava i sotto-obiettivi e le possibili azioni era simile a quello adottato dagli esseri umani. GPS fu, probabilmente, il primo programma ad adottare l'approccio del "pensare umanamente".

Nel mentre all'IBM Nathaniel Rochester e i suoi colleghi svilupperono alcuni dei primi programmi di intelligenza Artificiale. A partire dal 1952, Arthur Samuel scrisse una serie di programmi per il gioco della dama che arriveranno a giocare al livello di un buon dilettante. Questa innovazione ha demistificato il pensiero secondo il quale il calcolatore non poteva fare alcune cosa o poteva fare solo quello che viene detto ad esso. Il software, infatti, imparò presto a giocare meglio del suo creatore.

Nell'arco del 1956 Herbet Gelernter scrisse il Geometry Theorem Prover, capace di dimostrare teoremi che avrebbero dato qualche grattacapo a molti studenti di matematica. Nel 1958 McCarthy pubblicò un articolo intitolato *Programs with Common Sense*, nel quale descrisse Advice Taker, un ipotetico programma che può essere considerato come il primo sistema completo d'Intelligenza Artificiale. A differenza degli altri programmi discussi prima, il scopo di Advice Taker era incorporare conoscenza nel mondo. Ad esempio, il programma sarebbe stato capace di generare un piano di guida per l'aeroporto e prendere un aereo. Il programma era anche progettato in modo da potere accettare nuovi assiomi durante la normale esecuzione senza essere riprogrammato.

Cinque anni dopo McCarthy fondò il laboratorio di AI a Standford. Il lavoro a Standford si concentrava su metodi di uso generali per il ragionamento logico. Tra le applicazioni

della logica c'erano i sistemi di domanda-risposta e di pianificazione di Cordell Green e il progetto di robotica Shakey, presso il nuovo Standford Research Institute (SRI). Durante lo stesso anno sono stati sviluppati diversi programmi da diversi ricercatori usati per risolvere problematiche nell'ambito scolastico. Ad esempio:

- il programma SAINT di James Slagle fu capace di risolvere problemi di integrali chiusi dei corsi universitari del primo anno;
- il programma ANALOGY di Tom Evans risolveva problemi basati su analogie geometriche;
- il programma STUDENT di Daniel Bobrow risolveva problemi di algebra espressi in forma di racconto.

Il ritorno alla realtà (1966-1973

Da quando l'IA fu scoperta, e i sistemi di IA furano sviluppati e implementati, i ricercatori avevano una fiducia cieca in questi sistemi e nelle loro capacita. Avevano tante cose e secondo loro nel futuro ci sarebbero potuto essere macchine capaci di inventare, pensare e imparare. L'eccessiva sicurezza che avevano era dovuta alle promettenti prestazioni dei primi sistemi di IA, che però erano stati applicati su esempi semplici. In quasi tutti i casi questi sistemi falliranno miseramente quando furano stati applicati a problemi di tipologia differenti.

La prima difficoltà sorse perché i primi sistemi non avevano nessuna conoscenza o informazione sull'argomento delle loro elaborazioni. Il loro successo era basato su semplici manipolazioni sintattiche. I sistemi sviluppati non erano addestrati con ampie basi di dati per le loro applicazioni, quindi portavano i sistemi a sbagliarsi molto spesso.

La seconda difficoltà era l'intrattabilità di molti dei problemi che l'IA stava cercando di risolvere. Molti dei primi programmi giungevano alla soluzione provando varie combinazioni di passi. Questa strategia inizialmente funzionava perché i micromondi[3] contenevano pochi oggetti, e quindi le azioni possibili erano poche e le sequenze risolutive corte.

Un terzo problema sorse a causa di alcuni limiti fondamentali delle strutture di base usate per generare un comportamento intelligente. Ad esempio, il libro di Minskky e Papert *Percettroni* dimostrò che, benché i percettroni (una forma semplice di reti neurali) potessero apprendere tutto ciò che erano in grado di rappresentare, potevano, in effetti, rappresentare ben poco. In particolare, un percettrone a due input non potevano imparare a distinguere quando i suoi due input erano diversi. Poiché questi risultati non si potevamo a reti più complesse a diversi livelli, il finanziamento alla ricerca sulle reti neurali presto si ridusse fin quasi a zero.

Sistemi Knowledge-based (1969-1979)

Durante il primo decennio di studio dell'IA la visione globale della risoluzione dei problemi era quella di costituire passi elementari di ragionamento che ci potevano portare ad una soluzione. Con il tempo l'approccio è stat chiamato *debole* perché non riusciva a risolvere problemi più grandi o di difficoltà più elevata. L'alternativa ai metodi deboli è usare conoscenza più potente, specifica del dominio, che permette di intraprendere passi di ragionamento più ampi e può gestire più facilmente i casi tipici di aree ristrette di esperienza. In altre parole, possiamo dire che, per risolvere un problema difficile, bisogna, conoscerne già la soluzione.

[3]Micromondi sono ambienti artificiali semplificati e limitati, creati per permettere a un programma di AI di risolvere problemi di ragionamento, pianificazione o linguaggio in un contesto molto controllato.

La prima applicazione di questo approccio fu il programma DENDRAL sviluppato da Ed Feigenbaum, Bruce Buchanan e Joshua Lederberg. Questi ricercatori lavorarano su una possibile soluzione per risolvere il problema della ricostruzione dei dati molecolare usando i dati forniti da un spettrometro di massa. Con il tempo DENDRAL rappresentò un passo significativo perché fu il primo sistema di *conoscenza intensiva* di successo. Questa incredibile capacità del sistema a risolvere problemi nasce del fatto che era sviluppato con un ampia base di dati e un grande numero di regole.

Più avanti Feigenbaum, Buchanan e il Dott.Edward Shortliffe svilupperanno MYCIN per diagnosticare le infezioni del sangue. Questo lavoro era sempre basato sulla nuova metodologia dei sistemi esperti. Con circa 450 regole, MYCIN poteva offrire prestazioni pari a quelle di molti esperti e certamente migliori di quelle dei dottori neolaureati.

Rispetto a DENDRAL c'erano due grandi differenze. Prima di tutto, diversamente da MYCIN, DENDRAL non conteneva alcun modello teorico generale da cui poter dedurre le regole, che dovevano essere acquisite con una lunga serie di interviste agli esperti, che, a loro volta le avevano apprese da libri, dai colleghi e con l'esperienza personale. La seconda differenza era che le regole dovevano rispecchiare l'incertezza intrinsecamente associata alla conoscenza medica: MYCIN, per questo motivo, incorporava un metodo di calcolo di incertezza denominato *fattori di certezza* che, in quel periodo, sembrava corrispondere bene alla modalità con la quale i dottori utilizzavano le informazioni disponibili durante il processo diagnostico.

Nell'anno 1977 un ricercatore, Roger Schank insieme ai suoi studenti (Abelson, Riesbeck e Dyer) costruirano una serie di programmi tutti dedicati ala comprensione del linguaggio naturale. L'obbiettivo non era posto sul linguaggio in sé, ma più che altro sui problemi connessi alla rappresentazione e al ragionamento applicato alla conoscenza richiesta per la sua comprensione. Gli scopi principali di questi programmi erano:

- la rappresentazione di situazione tipiche;
- la descrizione dell'organizzazione della memoria umana;
- comprensione dei piani e obbiettivi della memoria umana.

La crescita di applicazioni a problemi reali ogni giorno ha creato lo sviluppo continuo di linguaggi diversi efficaci per la rappresentazione della conoscenza.

L'IA diventa un'industria (1980-presente)

R1 fu il primo sistema di IA immerso sul mercato ad avere un successo incredibile esso inizia ad operare presso la Digital Equipement Corporation (McDermott, 1982). Il programma aiutava la configurazione degli ordini di nuovi computer e, grazie ad esso la compagnia ha risparmiato circa 40 milioni di dollari all'anno. Nel 1988 la DEC aveva messo in opera 40 sistemi esperti e ne stava producendo altri. Infatti, quasi ogni compagnia americana di una certa grandezza aveva un proprio gruppo di IA e stava utilizzando o considerando l'uso di uno o più sistemi esperti.

Nel 1981 i giapponesi annunciarono il progetto della "Quinta Generazione" con l'obbiettivo di costruire computer intelligenti usando Prolog[4]. In risposta gli Stati Uniti formarono la Microelectronics and Computer Technology Corporation (MCC), un consorzio di ricerca che avrebbe dovuto assicurare la competitività nazionale. In entrambi casi l'IA faceva parte di uno sforzo a largo raggio, che includeva la progettazione di chip e la ricerca sulle

[4]Prolog è un linguaggio di programmazione logico che serve a far ragionare il computer usando fatti e regole logiche, invece di istruzioni passo-passo.

interfacce uomo-macchina. Globalmente l'industria dell'IA conobbe un boom, passando da pochi milioni di dollari nel 1980 a miliardi di dollari. Poco dopo avvenne il periodo che fu chiamato "inverno dell'IA", durante il quale molte compagnie soffrirono per l'impossibilità di mantenere le promesse fatte, spesso alquanto stravaganti.

Il ritorno delle reti neurali (1986-presente)

Fino alla fine degli anni '70 l'informatica aveva abbandonato completamente il campo delle reti neurali e si foccalizava in altre discipline. Fisici come John Hopfield usavano tecniche di meccanica statistica per analizzare le proprietà di memorizzazione e d'ottimizzazione delle reti, trattando insiemi di nodi come fossero insiemi di atomi.

Psicologici come David Rumelhart e Geoff Hinton continuavano a studiare i modelli della memoria basati su reti neurali. Negli anni '80 vi fu un vero ritorno quando almeno quattro gruppi differenti reinventerano l'algoritmo di apprendimento con retropropagazione scoperto nel 1969 da Bryson e Ho. L'algoritmo fu, dunque, applicato nel campo dell'informatica, della psicologia, e la pubblicazione dei risultati nella collezione *Parallel Distributed Processing* suscitò grande scalpore.

Questi modelli di sistemi intelligenti, chiamati *connessionisti*, furono considerati da qualcuno in diretta opposizione sia ai modelli simbolici di Newell e Simon, sia all'approccio basato sulla logica di McCarthy e altri. È ovvio che, in qualche modo, gli esseri umani manipolino dei simboli. I connessionisti più radicali si domandavano se la manipolazione simbolica fosse veramente utile per spiegare i modelli di cognizione più dettagliati. Una domanda rimasta senza risposta anche se l'opinione corrente è che gli approcci connessionista e simbolico siano complementari e non in competizione.

L'IA diventata una scienza (1987-presente)

L'IA, inizialmente nata come un movimento di protesta contro le limitazioni imposte da ricerche preesistenti, oggi si pone completamente nell'ambito del metodo scientifico. Al giorno di oggi si è verificata una rivoluzione nel campo dell'IA ed è più comune basarsi su teoremi rigorosi o prove sperimentali esistenti piuttosto che sull'intuito. Grazie all'uso di Internet e all'esistenza di dispositivi condivisi di codice e dati di test, è possibile replicare gli esperimenti dei ricercatori precedenti.

Negli anni '70, nel campo del riconoscimento vocale, fu provata una grande varietà di architetture e approcci diversi. Il campo è stato rapidamente dominato dagli approcci basati sui *modelli nascosti di Markov* che si basavano su due principi fondamentali:

- una rigorosa teoria matematica che permetteva di avvalersi dei risultati ottenuti in decenni di lavori in altri campi.
- i modelli sono generati mediante un processo di apprendimento basato su una grande mole di dati reali.

Il riconoscimento vocale così come il campo adesso attiguo relativo alla scrittura manuale, sta già compiendo la transizione verso un uso quotidiano da parte dell'industria e degli utenti privati.

Anche le reti neurali hanno seguito un'evoluzione simile. Le ricerche durante questo periodo, erano dedicate a investigare cosa poteva essere fatto e ad imparare in che modo le reti neurali differiscono dalle tecniche "tradizionali". L'obbiettivo principale era di poter applicare ad ogni problema la tecnica più promettente. In seguito a questi sviluppi è nata una nuova, e vigorosa industria che si appoggia alla tecnologia denominata *data mining*.

Analoghe "rivoluzioni gentili" si sono verificate nel campo della robotica, della divisione e della rappresentazione della conoscenza. Una migliore comprensione dei problemi e della loro complessità, unita al raffinamento degli strumenti matematici, ha portato al rafforzamento della ricerca e allo sviluppo di metodi robusti. In molti casi, la formalizzazione e la specializzazione hanno anche portato a una frammentazione: argomenti come la visione e la robotica sono sempre più isolati dall'IA "principale". Questi campi separati possono essere riunificati dalla visione comune dell'Intelligenza Artificiale come attività di progressione di agenti razionali.

La scomparsa degli agenti intelligenti (1995-presente)

Con il passare del tempo più i sistemi d'Intelligenza Artificiale continuarono ad avere un reale successo. I ricercatori hanno ricominciarano a considerare il problema degli agenti nella loro interezza. Il lavoro di Allen Newell, John Laird e Paule Rosenbloom su SOAR è l'esempio meno conosciuto di architettura completa per un agente. Il cosidetto "movimento situato" punta alla comprensione del funzionamento di agenti situati in ambienti reali con input continui.

Un'altro ambiente importante per gli agenti intelligenti è Internet; i sistemi di IA sono diventati così comuni nelle applicazioni web che il suffiso "-bot" è entrato nel linguaggio di tutti i giorni. Inoltre, le tecnologie proprie dell'IA sono alla base di molti strumenti, come i motori di ricerca, i sistemi di personalizzazione commerciale e quelli per la costruzione di siti web.

1.1.3 Introduzione ai sistemi di Intelligenza Artificiale generativi

Le IA generative, come descritto da Goodfellow et al (2014), sono sistemi avanzati d'Intelligenza Artificiale distinti per la loro capacità di creare nuovi dati e contenuti, imitando quelli reali. Questi sistemi si basano su algoritmi di apprendimento automatico per analizzare e comprendere grandi set di dati esistenti; successivamente utilizzano queste informazioni per creare contenuti nuovi.

Prendiamo, ad esempio, un sistema di AI generativa focalizzato sulla generazione di immagini. Inizialmente la rete neurale dell'IA viene "addestrata" esponendola a migliaia, o anche milioni, di immagini. Durante questa fase di addestramento, il sistema analizza questi dati, apprendendo dettagli come forme, colori, stili, luci, ombre etc. Quando la fase di addestramento è completata, il modello inizia il processo di creazione. Utilizzando le conoscenze acquisite, inizia a generare nuove immagini che non sono copie di quelle che ha visto durante l'addestramento, ma creazioni originali, basate sui pattern appresi.

Le caratteristiche chiave delle Generative AI includono:

- *Apprendimento Automatico (Machine Learning)*: queste sistemi di AI utilizzano algoritmi di apprendimento automatico per analizzare e imparare da grandi quantità di dati. Questo apprendimento è spesso non supervisionato o semi-supervisionato.
- *Capacità Generativa*: adifferenza del IA tradizionale, che si limita ad analizzare e categorizzare i dati, la Generative AI può creare nuovi contenuti. Questi includono testi, immagini, video, musica e codice.
- *Adattabilità e Personalizzazione*: questi sistemi sono in grado di adattarsi a vari contesti e possono essere personalizzati per esigenze specifiche, il che li rende utili in diversi settori, come l'arte, la scienza, la medicina e l'intrattenimento.

- *Interattività e Risposta in Tempo Reale*: queste sistemi possono interagire con gli utenti in tempo reale, offrendo risposte e contenuti generati in base alle richieste specifiche degli utenti.

1.1.4 Evoluzione storica dei sistemi d'Intelligenza artificiale generativi

Cerchiamo di comprendere le tappe fondamentali dell'evoluzione di questi modelli di IA Generativa evidenziando gli anni chiave.

1964: un ricercatore in informatica del MIT sviluppa ELIZA, un'applicazione di elaborazione del linguaggio naturale basata su testo. ELIZA è stata sostanzialmente il primo chatbot, all'epoca chiamato "chatterbot", basato su uno script di pattern-matching in grado di rispondere agli input, forniti sottoforma di testo in linguaggio naturale, con risposte testuali empatiche.

1999: Nvidia lancia *Geoforce*, che fu presentata come la prima GPU al mondo. Originariamente sviluppata per fornire una grafica in movimento fluida per i videogiochi, introduceva anche il concetto moderno di Graphics Processing Unit. Le GPU sono diventate di fatto la piattaforma per lo sviluppo di modelli AI e il mining di criptovalute.

2004: in quell'anno Google lancia una funzione migliorata e completamente automatica basandosi sulla *catena di Markov* svillupata nel 1906. Questa versione si attiva quando gli utenti inseriscono i termini di ricerca ed è, in grado di generare potenziali parole o frasi successive. Essa utilizzava anche enormi quantità di dati reali di ricerca, frequenza delle query più comuni e contesti (lingua, posizione, trend, etc).

2013: vengono introdotti *Variational Autoencoders (VAE)* i primi modelli generativi deep learning, grazie sopratutto ai lavori di Diederik P. Kingma e Max Welling. VAE è un tipo di rete neurale che appartiene all'ambito del machine learning. Rivolto ad un autoencoder classico, non comprime solo i dati ma impara anche una distribuzione probabilistica . Ad esempio, può generare nuovi esempi simili a quelli visti o immaginare varianti plausibili.

2014: IAN Goodfellow un ricercatore presso l'Università di Montreal, introduce le prime *Generative Adversarial Network* e i *Diffusion model*, due architetture per addestrare un modello generativo di IA. Entrambi questi sistemi insegnano alle macchine a creare e non solo a riconoscere.

2017: un gruppo di ricerca formato da Ashish Vanswani, un team di Google brain pubblica "Attention is all you need", un documento che illustra i principi dei trasformers in grado di abilitare i potenti *Foundation model* e gli strumenti di AI Generativa attuali. I transformer sono alla base di modelli linguistic moderni, come GPT, della traduzione automatica avanzata e dei cosidetti foundation models.

2018-2020: nell'anno 2018 OpenAI implementa il suo primo modello linguistico GPT-1, basato su transformer pre-addestrato, mostrando, che il pre-training funziona, anche se rimane molto limitato. Un anno dopo, lancia GPT-2, molto piu grande che sorprende per la qualità del testo. Nel 2020 ci fu un salto enorme di scala, con l'introduzione del GPT-3 con più di 175 miliardi di parametri e tante capacita emergenti.

2022: OpenAi presenta ChatGPT, un front-end di GPT-3 che genera frasi complesse, coerenti e contestualizzate, e contenuti in risposta ai prompt degli utenti finali.

1.1.5 Architettura dei modelli d'Intelligenza Artificiale Generativa

Dal 2013 abbiamo assistito allo sviluppo di modelli di AI veramente generativa, ovvero modelli di deep learning in grado di creare autonomamente contenuti su richiesta. Le principali architetture evolute durante questo periodo includono: Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), Neural Radiance Field (NeRf), Diffusion Model e, transformer.

Variational Autoencoders (VAE)

Un Variational Autoencoder è un modello probabilistico generativo nel campo del Machine Learning che combina reti neurali e inferenza variazionale per apprendere una distribuzione latente continua dei dati. A differenza degli autoencoder tradizionali, che "mappano" gli input in punti fissi, i VAE imparano a codificare i dati di input come distribuzioni di probabilità in uno spazio latente continuo e strutturato. La Figura 1.3 rappresenta in maniera chiara il concetto di base del funzionamento di un VAE.

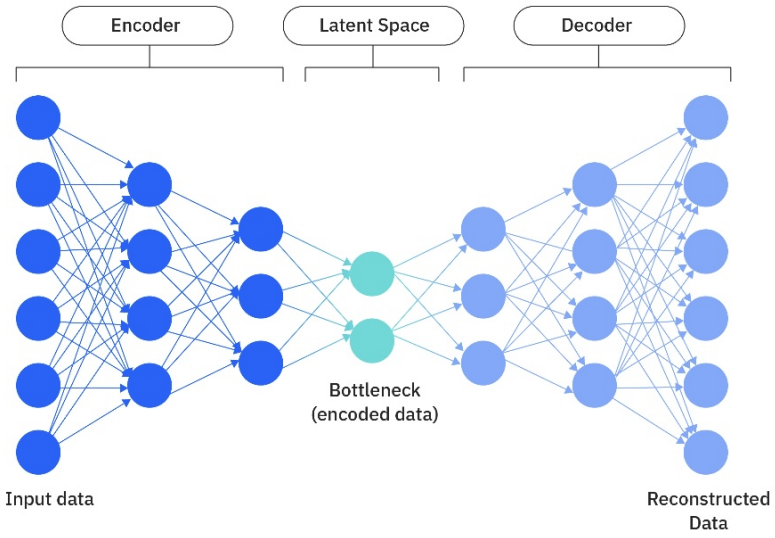


Figura 1.3: Funzionamento di base di un VAE

Ricordiamo che l'obiettivo principale di un autoencoder è quello di estrarre da un input le cosidette variabili latenti attraverso il passaggio in un collo di bottiglia prima di arrivare al livello di output. Le variabili latenti sono variabili spesso non osservabili, ma che contengono informazioni su come sono distribuiti i dati. Molto spesso non tutti i dati sono importanti quindi è importante estrarre solo quelli rilevanti.

Generative Adversarial Network (GAN)

Le GANs sono modelli generativi basato sull'idea centrale di fare competere due reti neurali tra loro in un gioco a somma zero. Infatti, una GAN è costituita da due componenti principali:

- *Generatore*: crea dati a partire da rumore casuale.
- *Discriminatore*: cerca di distinguere tra dati reali e dati generati.

La Figura 1.4 illustra come questi due componenti principali sono collegati e interagiscono tra di loro.

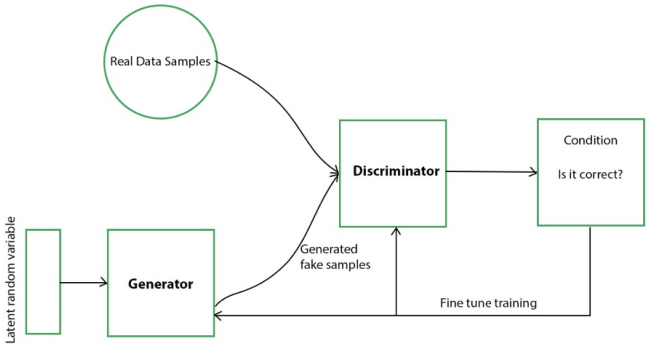


Figura 1.4: General Adversarial Network

Le due reti durante l'addestramento si scontrano fino a raggiungere un punto di equilibrio. Infatti, il generatore continua a produrre dati da sottoporre al discriminatore. Quando il discriminatore non è più in grado di distinguere dati reali da quelli creati artificialmente, allora si può concludere l'addestramento.

Modello di diffusione

Un modello di diffusione è un classe moderna di modelli generativi che apprende la distribuzione dei dati definendo un processo stocastico a due direzioni: *diffusione diretta* e *diffusione inversa*.

La diffusione diretta: è un processo basato concettualmente sulle catene di Markov. Essa consiste nel prendere un'immmagine e ad ogni passo aggiungere rumore causale guassiano fino ad arrivare a un'immagine composta completamento da disturbo.

La diffusione inversa: l'obiettivo di questo modello è quello di invertire il processo di diffusione diretta in quello che è chiamato processo di diffusione inversa. Ad ogni passo di rimuove una parte del rumore fino ad arrivare all'immagine desiderata. Avviene mediante un architettura chiamata U-NET. In quest'architettura, il modello non è addestrato per stimare l'immagine, bensì per predire il rumore, e dunque per arrivare all'immagine attraverso più passi di rimozione del rumore stimato.

Neural Radiance Field

I Neural Radiance Field (NeRF) sono tecniche di rappresentazione 3D basate su reti neurali nel campo del Machine Learning e della visione artificiale, progettate per sintetizzare nuove viste di una scena a partire da un insieme di immagini 2D.

Esse sfruttano le seguenti proiezione geometriche:

- *Ray casting*: una tecnica usata per determinare come una scena 3D viene proiettata su un'immagine 2D, simulando il percorso dei raggi visivi. Consiste nel lanciare raggi da un punto di vista nella scena per determinare quali oggetti vengono colpiti, e quali oggetti sono visibili.
- *Ray tracing*: una tecnica di rendering che simula in modo fisicamente plausibile il comportamento della luce tracciando il percorso dei raggi nello spazio 3D. Il ray tracing è un algoritmo che genera immagini seguendo il cammino della luce.

- *Rasterizzazione*: un processo che converte primitive geometriche (come triangoli) in pixel sullo schermo, determinando quali pixel devono essere colorati e con quale valore.

Large Language Models

I Large Language Model(LLM) rappresentano un'evoluzione significativo nel campo dell'Intelligenza Artificiale, in particolare nella comprensione e nella generazione di testi simili a quelli umani.

Una LLM è una rete neurale (tipicamente basata su architetture transformer) addestrata su enormi quantità di testo e per apprendere la distribuzione probabilistica del linguaggio e predire sequenze parole.

Questi modelli possiedono una conoscenza e una comprensione approfondite del linguaggio che consentono di svolgere varie attività come, la generazione di testo, la traduzione di testi in diverse lingue, il question answering e la sintesi di testi.

Gli LLM sono addestrati su grandi set di testi e codice per potere apprendere relazioni compresse tra parole e frasi. Sono costruiti su tre pilastri fondamentali: dati, architettura e addestramento. La Figura 1.5 riassume il procceso di apprendimento di un LLM.

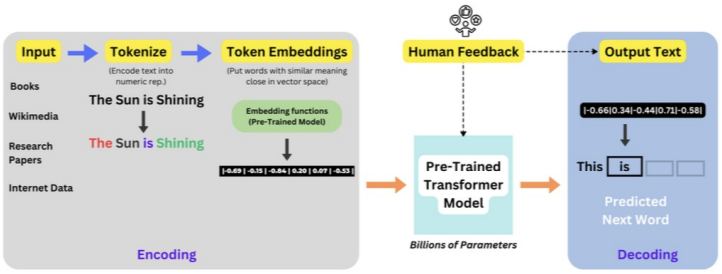


Figura 1.5: Apprendimento di un LLM

Esso è costituito dai seguenti passi:

- *Input*: viene addestrato il modello su una grande scala di dati testuali proveniente da libri, wikipedia, articoli, e vari siti internet.
- *Tokenizzazione*: questo processo trasforma il testo in unità più piccole, chiamate *token*, che il modello può elaborare.
- *Token Embedding*: ogni token viene trasformato in un vettore numerico che il modello puo realmente "capire".
- *Encoding*: i token embedding sono passati attraverso i livelli di encoding del trasformer e sono elaborati per comprendere il contesto di ciascuna parola all'interno della frase.
- *Modello Transformer pre-addestrato*: come detto prima, il transformer costituisce l'architettura principale di un LMM. Esso a prevedere parti del testo da altre parti, regolando i suoi parametri interni per ridurre al minimo l'errore di previsione.
- *Feedback umano*: nel contesto degli LLM, il feedback umano è un processo facoltativo con cui persone reali aiutano a migliorare il comportamento del modello, fornendo feedback e quindi rendendo le risposte più utili, corrette e sicure.
- Testo di input: quando un testo è generato, il modello utilizza i parametri pre-addestrati, e possibilmente perfezionati, per prevedere la parola successiva in una sequenza. Ciò comporta il processo di decodifica.

- Decoding: è la fase durante la quale il modello riconverte gli embedding elaborati in testo leggibile dell'uomo. Per prevedere òa parola successiva viene effettuato un sulla distribuzione di probabilità del prossimo token, il token con la probabilità più alta vienne selezionato. Questa previsione può essere l'output finale oppure fungere da token di input aggiuntivo per il modello per prevedere le parole successive.

1.1.6 Criticità e vincoli dei sistemi d'Intelligenza Artificiale Generativi

Con il passare degli anni i modelli di IA Generativa rappresentano delle tecnologie molto avanzate e diffuse. Riescono a produrre testi, immagini, suoni, video, codice e altri contenuti a partire da semplici input dell'utente. Benché abbiano numerosi vantaggi in termine di produttività, automazione e supporto decisionale, questi sistemi hanno anche una serie di criticità e rischi che devono essere considerati con cura. Questi rischi non si basano su un unico aspetto, ma su diverse dimensioni del funzionamento dei modelli. Ad esempio, esssi riguardano la gestione dei dati, la corretezza delle risposte, la coerenza dei risultati, l'impatto sociale dei contenuti generali, ma più che altro gli effetti sugli utenti. Per questo motivo è importante analizzare in modo dettagliato le principali categorie di rischio associate a questi sistemi, al fine di capire le implicazioni e le possibili conseguenze nei diversi contesti di utilizzo.

Violazione normative

I modelli di IA generativa addestrati su dati sensibili potrebbero generare involontariamente output che violano le normative; ad esempio, essi potrebbero:

- *riprodurre informazioni personali identificabili (PII)*, come nomi, indirizzi, numeri di telefono o dati sanitari;
- *memorizzare indirettamente dati sensibili*, sopratutto se il dataset non è stato adeguatamente filtrato;
- *ricostruire informazioni riservate* tramite inferenze o combinazioni di dati apparentemente innocui.

Nel caso in cui uno di questi rischi, si verifica, le conseguenze possono essere serie, tra cui sanzioni legali, danni reputazionali e perdita di fiducia da parte degli utenti.

Rischi sociali

I rischi sociali rappresentano la possibilità che vengano generati contenuti indesiderati che possano riflettersi negativamente nell'organizzazione. Qui ci referiammo principalmente sull'impatto che il contenuto generato può avere sugli utenti. Le problematiche possono includere, informazioni offensive, inappropriate, razziste o discriminatorie, messaggi fuorvianti o non accurati; contenuti che non rispettano i valori o le policy dell'organizzazione. Per concludere, i rischi sociali non rappresentano non solo errori tecnici da parte dei sistemi d'IA Generative, ma più che altro riguardano come quegli errori vengono percepite dalle persone e dalla società.

Non determinismo

Il modello potrebbe generare input diversi a partire dallo stesso output, il che può creare problemi nelle applicazioni in cui l'affidabilità e la coerenza sono fondamentali. Infatti, ci rendiamo conto che l'output del modello potrebbe variare anche fornendo lo stesso prompt. I due motivi principali per queste variazioni sono i seguenti:

- vengono utilizzate componenti probabilistiche;
- viene introdotta variabilità per generare riposte più naturali e diversificate.

La realtà è che questa variabilità oggi è accettabile soprattutto con l'uso dei chatbot in cui è necessario dimostrare creatività e adattare i loro output in base alle richieste. Però, in altri ambiti, rappresenta criticità. Ad esempio, nell'ambito aziendale o legale in cui serve coerenza nelle risposte, sistemi automatizzati che forniscono risultati diversi possono causare errori operativi, un discorso analogo vale per il debug e il testing che rendono difficile riprodurre un comportamento specifico.

Allucinazioni

Il modello può generare risposte inesatte o inventate, non coerenti con i dati di addestramento o con la realtà. Questi errori vengono chiamati *allucinazioni*. Tali allucinazioni si verificano generalmente quando il modello:

- combina in modo errato informazioni apprese durante l'addestramento;
- produce informazioni plausibili ma false;
- prova a riempire vuoti quando non ha dati sufficienti.

Il problema principale è queste allucinazioni sembrano credibili, e quindi è difficile accorgersi dell'errore. Di conseguenza, in qualche modo, si viene a danneggiare pure la credibilità del sistema.

Sicurezza dei dati e preoccupazioni sulla privacy

Le informazioni condivise con il modello possono includere dati personali e sensibili, con il rischio di violare le normative sulla privacy. Gli input forniti dagli utenti generalmente possono trattare informazioni quali nomi, indirizzi, dati finanziari, informazioni aziendali riservate, e altro. Nel caso in cui un sistema di IA Generativa non gestisce correttamente questi dati, potremmo avere dei problemi, come, un accesso non autorizzato ai dati, la memorizzazione o il riutilizzo improprio delle informazioni, la fuga di dati attraverso le risposte del modello o i sistemi collegati. Questo è particolarmente critico in relazione a normative come il *Regolamento generale sulla protezione dei dati*, che impone regole rigorose sulla raccolta, nel trattamento e nella protezione dei dati.

CAPITOLO 2

OpenAI e l'Intelligenza Artificiale Generativa

2.1 OpenAI

OpenAI è un laboratorio di ricerca e di sviluppo sull'Intelligenza Artificiale fondato nel dicembre 2015 da Elon Musk, Sam Altman e altri investitori. OpenAI nasce come organizzazione no-profit con la missione di garantire che l'Intelligenza Artificiale sia sviluppata in modo sicuro e a beneficio di tutta l'umanità, per poi evolversi nel tempo verso una struttura più complessa, mantenendo comunque al centro i principi fondatori di sicurezza e accessibilità. Nel 2025 OpenAI ha convertito la filiale a scopo di lucro in una Public Benefit Corporation[1] (PBC) che è di proprietà del 26 percento dell'organizzazione o profit. Ad aprile 2026, la società è valutata a circa 800 miliardi di dollari.

2.1.1 La nascita di OpenAI

La creazione di OpenAI rappresenta un momento significativo nello sviluppo recente dell'Intelligenza Artificiale. Nel dicembre 2015, Elon Musk, Sam Altman e altri investitori fondarono OpenAI, contribuendo con oltre 1 miliardo di dollari al progetto. Il 27 Aprile 2026 OpenAI ha rilasciato una versione beta pubblica di OpenAI Gym, una piattaforma per ricerca di apprendimento per rinforzo.

Un anno dopo la sua creazione OpenAI, rilascia *Universe*, un software ideato per misurare e allenare l'intelligenza generale di un AI tramite giochi, siti web e altre applicazioni.

Il 21 febbraio 2018, Elon Musk si è dimesso dal consiglio di amministrazione a causa di potenziali conflitti di interesse con lo sviluppo dell'Intelligenza Artificiale di Tesla.

Il 13 gennaio 2023 viene lanciato il progetto *World*, un servizio che permette di distinguere un bot alimentato dall'Intelligenza Artificiale dagli esseri umani mediante una sfera bianca che scansiona l'iride. Attualmente World è disponibile in 160 paesi nel mondo con 40 milioni di utenti iscritti sull'app.

Nel corso del novembre 2025 viene siglato un accordo da 38 miliardi di dollari per 7 anni con Amazon Web Service.

[1] Una Public Benefit Corporation (PBC) è una struttura societaria ibrida che si colloca tra un'organizzazione non-profit e una società tradizionale a scopo di lucro.

2.1.2 Struttura aziendale

2.1.3 Dalla non-profit al profit: Evoluzione della struttura aziendale

Creazione di filiali a scopo lucro

Nel 2019, OpenAI è passata dall'essere un organizzazione non-profit all'essere un'organizzazione con obbietivi di profitto, con un profitto che è stato compresso a 100 volte qualsiasi investimento, valutando così la società a più di 100 miliardi di dollari. OpenAI ha formato tre filiali:

- la holding OpenAI LP;
- la holding OpenAI GP LLC;
- la società a scopo di lucro OpenAI Global, LLC.

I fondatori hanno dichiarato che il loro obiettivo sarebbe di attrarre legalmente investimenti dai fondi di rischi e compensare i dipendenti. Microsoft ha successivamente investito 1 miliardo di dollari in OpenAI LP e i servizi di OpenAI sono stati migrati su Microsoft Azure. Da allora, i sistemi OpenAI sono stati eseguiti su una piattaforma di supercomputer basata su Azure.

La ristrutturazione del 2019 ha reso l'organizzazione no-profit OpenAI, Inc., l'unico azionista di controllo di OpenAI LP, che ha mantenuto una responsabilità fiduciaria formale per la carta senza scopo di lucro di OpenAI, Inc. Alla maggior parte del consiglio di amministrazione di OpenAI, Inc. è stato impedito di avere partecipazioni finanziarie in OpenAI LP. Al fine di comprendere meglio questa struttura organizzativa, la Figura 2.1 rappresenta graficamente la distribuzione del controllo e delle partecipazioni finanziarie all'interno di OpenAI.

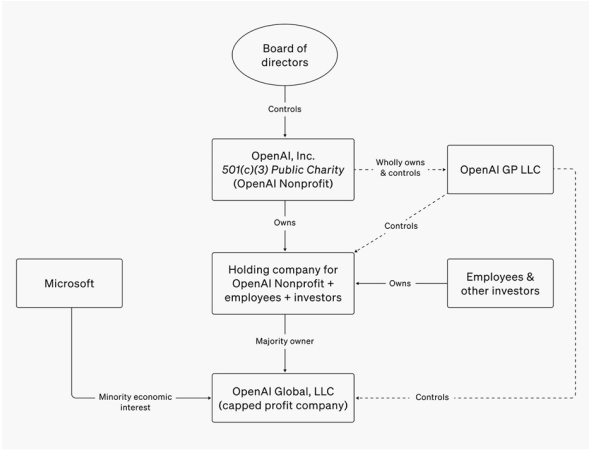


Figura 2.1: Organizzazione del consiglio di amministrazione di OpenAI

Transizione ad una società di pubblica utilità

Nel dicembre 2024, OpenAI ha proposto un piano di ristrutturazione per convertire il profitto limite in una società di pubblica utilità (PBC) con sede nel Delaware, rimuovere il suo limite di profitto e liberarlo dal controllo dell'organizzazione no profit. L'organizzazione no profit venderebbe il suo controllo e altri beni, ottenendo in cambio capitale proprio. La Figura 2.2 illustra la struttura organizzativa completa di OpenAI, evidenziando la gerarchia di controllo che va dal Board of Directors fino alle numerose filiali internazionali, tra cui OpenAI Ireland Ltd, OpenAI UK Ltd, OpenAI Japan Ltd. e OpenAI India Pvt Ltd.

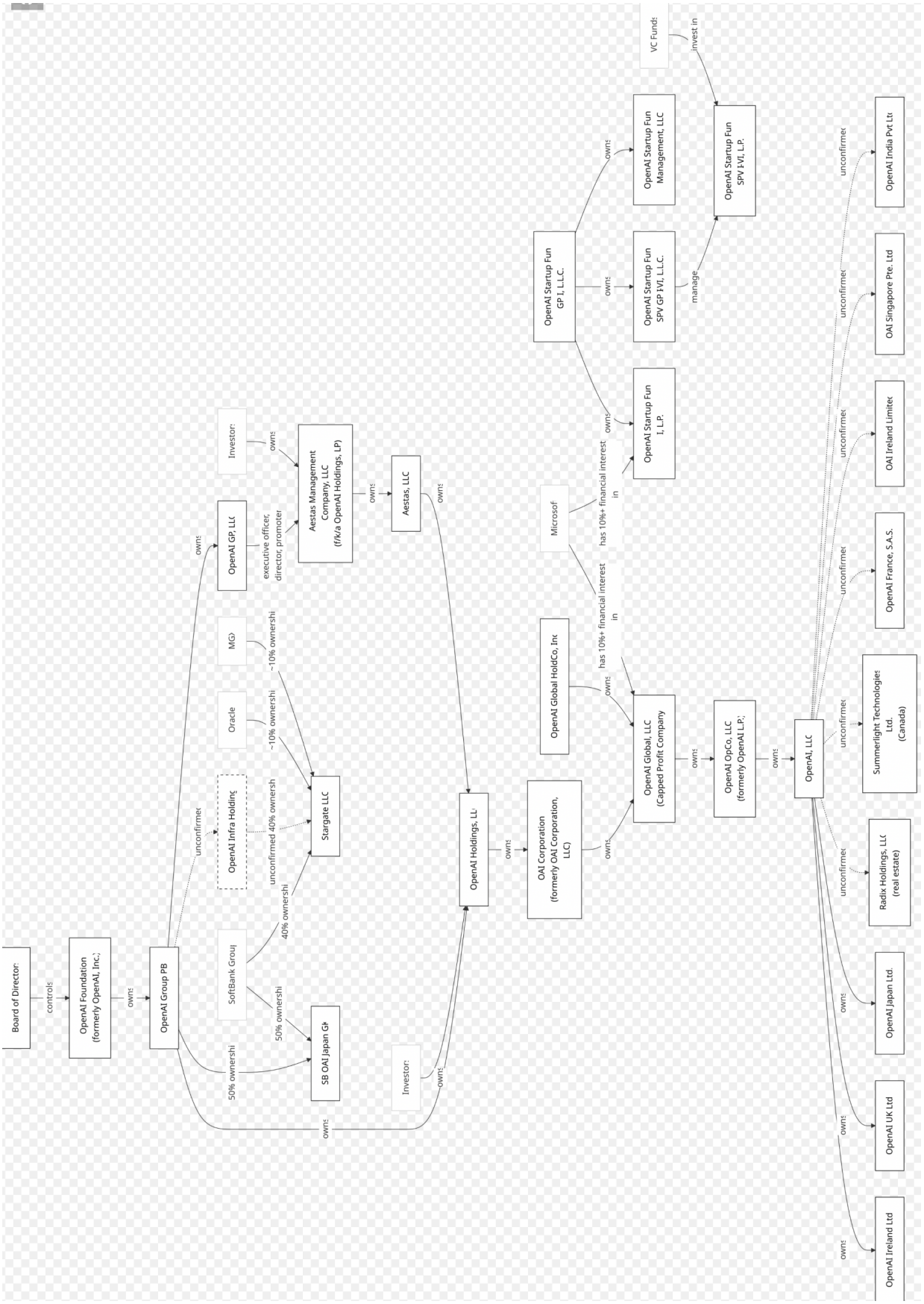


Figura 2.2: Organizzazione del consiglio amministrativo di OpenAI da Novembre 2025

Il 28 ottobre 2025, OpenAI ha annunciato di aver adottato la nuova struttura aziendale PBC. La OpenAI Foundation detiene una partecipazione del 26 percento nella PBC, mentre Microsoft detiene una quota del 27 percento; il restante 47 percento è di proprietà di dipendenti e altri investitori.

2.1.4 Panoramica dei prodotti di OpenAI

Con il passare del tempo OpenAI ha sviluppato una vasta gamma di prodotti basati sull'Intelligenza Artificiale. Tra i principali si distingue ChatGPT, un modello conversazionale in grado di comprendere e generare linguaggio naturale. Accanto a esso, DALL-E, consente la creazione di immagini basandosi su descrizioni testuali, mentre Sora rappresenta un passo significativo nell'evoluzione della generazione automatica di contenuti video. Di seguito veranno considerati i principali prodotti di OpenAI, analizzando le loro caratteristiche, il livello di innovazione e l'impatto che essi hanno avuto nella società.

Gym

OpenAI Gym è un toolkit per la ricerca sull'apprendimento per rinforzo[2]. Include una crescente raccolta di problemi di benchmark[3] che espongono un'interfaccia comune e un sito web in cui le persone possono condividere i loro risultati e confrontare le prestazioni degli algoritmi. Il modello si propone di offrire un benchmark di intelligenza generale facile da impostare con un'ampia varietà di ambienti differenti. Il comportamento in qualche modo affine a, ma più ampio di, *"ImageNet Large Scale Visual Recognition Challenge"* usato nelle ricerche riguardanti l'apprendimento supervisionato. *mageNet Large Scale Visual Recognition Challenge* è un punto di riferimento nella classificazione e nel rilevamento delle categorie di oggetti su centinaia di categorie e milioni di immagini. Inoltre il progetto mira a rendere più uniforme la modalità con cui gli ambienti vengono definiti negli articoli scientifici sull'Intelligenza Artificiale, così da permettere una più facile riproduzione degli esperimenti pubblicati. A settembre 2017 il sito è diventato statico e non aggiornato dagli sviluppatori; infatti il suo contenuto e gli aggiornamenti sono stati spostati su GitHub.

RoboSumo

RoboSumo è un progetto sperimentale sviluppato nell'ambito dell'apprendimento per rinforzo multi-agente in cui ci sono dei robot umanoidi in "meta-apprendimento" che inizialmente non sanno neanche come camminare; ad essi vengono dati gli obbietivi di camminare e di spingere l'avversario fuori dal ring come nel sumo[4]. Tramite questo processo di apprendimento antagonistico, imparano da soli attraverso prove e riescono ad adattarsi alle condizioni variabili. La Figura 2.3 illustra un robot umanoide, una macchina progettata per riprodurre le caratteristiche fisiche e i movimenti del corpo umano.

[2] L'apprendimento di rinforzo è una tecnica di apprendimento automatico che mira a realizzare agenti autonomi capace di scegliere azioni da compiere per il conseguimento di determinati obbiettivi tramite interazione con l'ambiente in cui sono immersi

[3] Benchmark rappresenta un insieme di prove del software volti a fornire una misura delle prestazioni di un computer per quanto riguarda diverse operazioni.

[4] Il sumo è uno sport di lotta tradizionale giapponese.

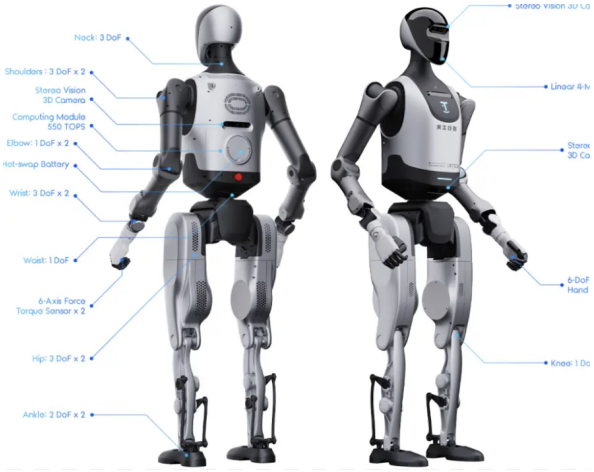


Figura 2.3: Esempio di un robot umanoide

Quando un agente è rimosso dall'ambiente virtuale e collocato in un nuovo ambiente con un forte vento, l'agente si tiene per stare dritto, il che fa capire che ha imparato come stare in equilibrio in generale. Igor Mordatch, di OpenAI, sostiene che la competizione tra agenti può creare una "corsa agli armamenti" di intelligenza che può incrementare l'abilità dell'agente, anche al fuori del contesto della competizione.

Debate Game

Nel 2018, OpenAI ha lanciato il "Debate Game", che insegna alle macchine a discutere di fronte a un giudice umano. Lo scopo è di fare ricerca sul fatto che questo approccio possa aiutare nel controllare le decisioni dell'IA e nello sviluppare l'Intelligenza Artificiale interpretabile, ossia un'intelligenza di cui le azioni e decisioni siano comprensibili.

OpenAI Five

OpenAI Five è il nome di cinque bot, curati da OpenaAI usati in Dota 2, un videogioco competitivo 5 contro 5. Tramite un'algoritmo trial and error, imparano a giocare contro giocatori umani di alto livello. Ognuna delle reti di OpenAI Five contiene una LSTM (Long Short Term Memory network)[5] a singolo livello 1024, unità che vede lo stato corrente del gioco (estratto dall'API Bot di Valve) ed emette azioni attraverso diverse possibili teste d'azione. Ogni testa ha un significato semantico; ad esempio, il numero di tick per ritardare questa azione, quale azione selezionare, la coordinata X o Y di questa azione in una griglia intorno all'unità, etc. Le teste d'azione sono calcolate in modo indipendente. OpenAI Five vede il mondo come un elenco di 20.000 numeri e intraprende un'azione emettendo un elenco di 8 valori di enumerazione. Essa seleziona diverse azioni e diversi obiettivi per capire come codificare ogni azione e come osservare il mondo. L'immagine mostra la scena come l'avrebbe vista un essere umano. OpenAI Five può reagire a pezzi mancanti di stato che si correlano con ciò che vede. Ad esempio, fino a poco tempo fa, le osservazioni di OpenAI Five non includevano zone shrapnel[6], che gli umani vedono sullo schermo.

Il sistema è implementato come un sistema di allenamento RL (Reinforcement Learning) generico chiamato *Rapid*, che può essere applicato a qualsiasi ambiente Gym. È stato usato

[5]LSTM è un tipo di rete neurale artificiale progettata per apprendere dipendenze a lungo termine nelle sequenze di dati.

[6]Le zone shrapnel sono aree del gioco Dota 2 in cui pioggia di frammenti metallici danneggia e rallenta continuamente i nemici presenti al loro interno, visibili sullo schermo ai giocatori umani.

Rapid per risolvere altri problemi a OpenAI, tra cui Competitive Self-Play. La figura 2.4 illustra come funziona il sistema di addestramento.

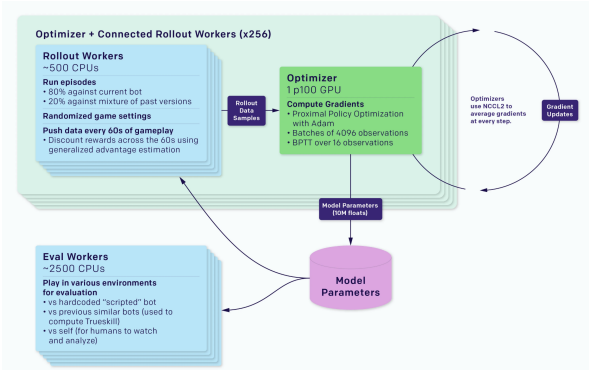


Figura 2.4: Funzionamento del sistema di addestramento Rapid

Tale sistema di addestramento è suddiviso in "Rollout workers" (lavoratori di lancio), che eseguono una copia del gioco e un'esperienza di raccolta degli agenti, e nodi di ottimizzazione, che eseguono una discesa del gradiente sincrono su una flotta di GPU. I lavoratori di lancio sincronizzano la loro esperienza tramite Redis con gli ottimizzatori. Ogni esperimento contiene anche lavoratori che valutano l'agente addestrato rispetto agli agenti di riferimento, nonché software di monitoraggio come TensorBoard, Sentry e Grafana. Durante la discesa del gradiente sincrono, ogni GPU calcola un gradiente sulla sua parte del batch, e quindi i gradienti sono mediati a livello globale. Le latenze per la sincronizzazione di 58MB di dati (dimensione dei parametri di OpenAI Five) su diversi numeri di GPU sono mostrate a destra. La latenza è abbastanza bassa da essere in gran parte mascherata dal calcolo della GPU che viene eseguito in parallelo con essa.

Prima di diventare un team di cinque, la prima dimostrazione pubblica avvenne al The International 2017, il torneo annuale del gioco, dove, Dendi, un videogiocatore professionista ucraino, ha perso contro il bot una partita diretta uno contro uno. Dopo la partita, il CTO Greg Brockman ha spiegato che il bot ha imparato a giocare contro se stesso per due settimane e che il software di apprendimento era un passo avanti nel creare software che possa occuparsi di problemi complessi "come essere un chirurgo". Il sistema usa una forma di apprendimento per rinforzo, nella quale i bot imparano giocando contro se stessi per centinaia di volte al giorno per mesi, ed essendo premiati per azioni come uccidere un nemico o distruggere torri. A giugno 2018 l'abilita dei bot è cresciuta fino a giocare insieme come una squadra di cinque e a battere squadre di amatori o semiprofessionisti. Al The International 2018, OpenAI Five ha giocato due partite contro giocatori professionisti. Sebbene i bot abbiano perso entrambe le volte, OpenAI lo considera l'esperimento un successo in quanto ciò ha permesso a OpenAiFive di poter analizzare e migliorare gli algoritmi per le partite future.

Dactyl

Dactyl è un progetto utilizzato per allenare un robot Shadow hand (una mano robotica altamente avanzata) da zero, usando lo stesso algoritmo di apprendimento per rinforzo che usa OpenAI Five. Il sistema lavora addestrando la mano del robot in una simulazione e poi trasferendo le conoscenze acquisite nel mondo reale. Viene posizionato un oggetto come un blocco o un prisma nel palmo della mano e Dactyl deve riposizionarlo in un orientamento diverso, ad esempio ruotando il blocco per mettere una faccia sopra. La Figura 2.5 illustra come sono stati condotti gli esperimenti della mano robot con un blocco.

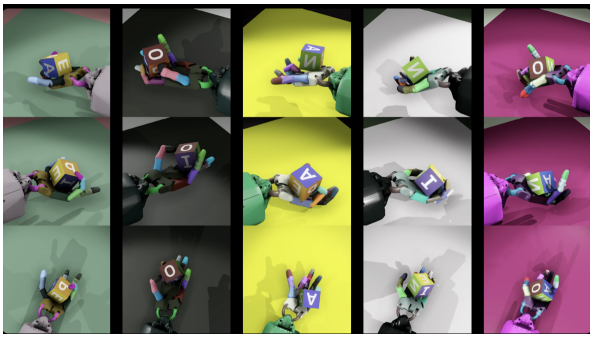


Figura 2.5: Immagini di allenamento utilizzate per imparare a stimare la posa del blocco.

Il risultato è una mano robot che può effettuare diversi task in modo efficiente, usando una selezione di movimenti e senza che questi ultimi siano programmati individualmente da umani. Il sistema è stato disegnato per manipolare oggetti arbitrari, non solo quelli appositamente modificato per supportare il tracciamento. Pertanto, uno utilizza immagini della telecamera RGB per stimare la posizione e l'orientamento dell'oggetto. Una proprietà interessante di Dactyl è il fatto che usa alcuni movimenti che sono tipici di una mano umana; questi sono comportamenti che esso non hanno imparato, o più che altro non sono stati programmati da ingegneri di OpenAI, ma sono avvenuti autonomamente. Ad esempio il dito che scivola, che ruota o ancora camminare con le dita.

Benché Dactyl produceva dei risultati impressionanti ci sono stati anche delle sorpresa sul modo di funzionare delle tecniche impiegate durante la fase di addestramento. In particolare:

- La diminuzione del tempo di reazione non ha migliorato le prestazioni. La saggezza convenzionale afferma che ridurre il tempo tra le azioni dovrebbe migliorare le prestazioni perché i cambiamenti tra gli stati sono più piccoli, e quindi più facili da prevedere. Il tempo usato durante il periodo di testing era di circa 80ms, che è inferiore al tempo di reazione umano di 150-250ms, ma significativamente più grande del tempo di calcolo della rete neurale di circa 25ms. Sorprendentemente, la diminuzione del tempo tra le azioni a 40ms ha richiesto un tempo di allenamento aggiuntivo, ma non ha migliorato notevolmente le prestazioni nel mondo reale. È possibile che questa regola empirica sia meno applicabile ai modelli di rete neurale che ai modelli lineari che sono di uso comune oggi.

- L'uso di dati reali per addestrare le politiche di visione non ha fatto alcuna differenza. Nei primi esperimenti, è stata utilizzata una combinazione di dati simulati e reali per migliorare i modelli. I dati reali sono stati raccolti da prove della politica contro un oggetto con marcatori di tracciamento incorporati. Tuttavia, i dati reali hanno svantaggi significativi rispetto ai dati simulati. Le informazioni sulla posizione dei marcatori di tracciamento presentano una latenza e degli errori di misurazione. Peggio ancora, i dati reali vengono facilmente invalidati dalle comuni modifiche alla configurazione, rendendo difficili raccogliere abbastanza dati per essere utili. Man mano che i metodi si sviluppavano, l'unico errore del simulatore è migliorato fino a quando non corrispondeva all'errore ottenuto utilizzando un mix di dati simulati e reali. I modelli di visione finale sono stati addestrati senza dati reali.

Per concludere, il progetto OpenAI Dactyl è stato un avanzamento significativo nel campo della manipolazione robotica, dimostrando come sistemi di Intelligenza Artificiale possano apprendere compiti complessi attraverso tecniche di apprendimento per rinforzo. L'abilità del modello di addattarsi e variare il proprio comportamento davanti a contesti

diversi e di eseguire operazioni previste dimostra il potenziale dell'Intelligenza Artificiale nell'interazione con il mondo fisico. Tali risultati hanno aperto nuove prospettive per lo sviluppo di robot sempre più autonomi e versatili.

DALL-E

DALL-E è un algoritmo di Intelligenza Artificiale addestrato per generare immagini a partire da descrizioni testuali utilizzando un dataset contribuito da coppie testo-immagine. È stato sviluppato da OpenAI, e presentato il 5 gennaio 2021. Possiede un'ampia gamma di funzionalità, tra cui la creazione di versione antropomorfe di animali e oggetti, la combinazione di concetti non correlati in modo plausibile, la resa del testo e l'applicazione di trasformazioni a immagini esistenti. Nel corso degli anni, l'algoritmo è stato affinato e perfezionato consentendo ora la generazione di immagini più coerenti e impiegando un tempo inferiore rispetto al passato. Attualmente, il tool è capace di generare immagini in formato quadrato con risoluzione di 1042 pixel, e, grazie ad un editor d'immagini, c'è anche la possibilità di aggiungere, rimuovere e modificare elementi di un'immagine precedentemente generata.

Sviluppo e versioni di DALL-E

A gennaio 2021 è stata introdotta la prima versione di DALL-E e ha suscitato grande scalpore per le impressionanti capacità dell'Intelligenza Artificiale. Il modello era già in grado di creare o variare immagini, disegni e dipinti realistici in vari stili artistici, e molto altro. Il generatore di testo a rapporto delle immagini si basa sul generatore di testo Generative Pretrained Transformer 3 (GPT-3), sviluppato anch'esso da OpenAI, e in linea di principio può essere utilizzato da chiunque tramite un'interfaccia web.

Un anno dopo, nel 2022, è stata rilasciata una versione migliorata di DALL-E che e utilizzava un modello di diffusione denominato GLIDE [7] ed era condizionato da inclusior CLIP (Contrastive Language-Image Pretraining)[8]. Le caratteristiche speciali di questa versione includono, ad esempio, un'elaborazione delle immagini più rapida, riflessi di luce e condizioni di illuminazione più realistici, sfondi più complessi, la funzione di inpainting (modifica di un'area specifica dell'immagine), la creazione di diverse varianti di immagini in stili diversi o l'aggiunta e la combinazione di più immagini.

Ad ottobre 2023 si è stato un miglioramento della qualità dell'immagine con DALL-E 3. Il modello è stato integrato più strettamente con CHATGPT, consentendo agli utenti di generare immagini ancora più precise a partire dall'input di testo. Oltre al miglioramento della risoluzione e del livello degli immagini, l'interpretazione delle query complesse è stata notevolmente migliorata. Gli utenti possono ottenere risultati più mirati attraverso opzioni di controllo ampliate, come specifiche di stile e composizione. Un altro punto forte di questa versione di DALL-E è stata la possibilità di modificare immagini esistenti seguendo istruzioni specifiche.

A partire dal 2025, il modello è stato integrato in CHATGPT, consentendo agli utenti di creare immagini direttamente dalle query di testo. Sebbene inizialmente il modello fosse disponibile solo per gli utenti di CHATGPT Plus ed Enterprise, nell'agosto 2024 è stato rilasciato gratuitamente anche agli utenti, ma con un limite di due immagini al giorno. Microsoft ha inoltre integrato DALL-E 3 nel suo Bing Image Creator per rendere la generazione di

[7] GLIDE sta per "Guided Language to Image Diffusion for Generation and Editing"; si tratta di un modello di diffusione sviluppato da OpenAI per la creazione e l'elaborazione di immagini basate su istruzioni di testo di lingua naturale. GLIDE ottiene risultati eccellenti nella creazione di immagini e supera le prestazioni delle Generative Adversarial Network.

[8] CLIP è una procedura AI sviluppata da OpenAI. Si tratta di una rete neurale addestrata con una grande quantità di coppie immagine-testo che può prevedere la didascalia più rilevante per una data immagine. CLIP ha capacità Zero-Shot e può essere utilizzato, ad esempio, per cercare immagini con descrizioni di testo o per etichettare automaticamente le immagini. Il generatore di testo-immagine DALL-E di OpenAI utilizza CLIP.

immagini accessibile a una base di utenti più ampia. Tuttavia, un problema di qualità ha portato a un temporaneo declassamento dei modelli di generazione delle immagini all'inizio del 2025. Derivata da DALL-E, esiste una versione del generatore di testo in immagini originariamente chiamata DALL-E mini e ora chiamata Craiyon. Si basa sul codice sorgente di DALL-E, ma è meno potente. Oltre a DALL-E, oggi sono disponibili numerosi altri generatori di testo-immagine basati sull'intelligenza artificiale, come Stable Diffusion o Midjourney. Questi modelli alternativi offrono approcci diversi alla generazione di immagini e consentono di ottenere risultati variabili attraverso comunità open source o algoritmi specifici.

ChatGPT

ChatGPT è un modello d'Intelligenza Artificiale Generativa sviluppato da OpenAI che usa Large Language Model, piu precisamente Generative pre-trained transformer, per generare testi, discorsi e immagini in risposta alle richieste degli utenti. Questo modello ha accelerato la rapida crescita dell'Intelligenza Artificiale in un periodo in cui tanti investimenti sono stati fatti in questo campo. Gli utenti interagiscono con ChatGPT tramite testo, audio e immagini rendendo la discussione dinamica. Il modello è stato velocemente accettato arrivando fino a 100 milioni di utenti attivi due mesi dopo il suo lancio.

La rapida espansione del modello è probabilmente dovuta alla versatilità del suo utilizzo. Il chatbot è usato, ad, esempio per scrivere programmi ed effetuare il loro debug per generare concetti business, per scrivere canzoni e altro. Una funzione opzionale, "Memoria", consente agli utenti di dire a ChatGPT di memorizzare informazioni specifiche. Un'altra opzione consente a ChatGPT di richiamare vecchie conversazioni. I classificatori di moderazione basati su GPT vengono utilizzati per ridurre il rischio di output dannosi presentati agli utenti. Nel marzo 2023, OpenAI ha aggiunto il supporto per i plugin[9] a ChatGPT. Questo include sia i plugin realizzati da OpenAI, come la navigazione web e l'interpretazione del codice, sia plugin esterni di sviluppatori come Expedia, OpenTable e Zapier. ChatGPT rimane un modello gratuito fino ad una certa capacità di utilizzo. A febbraio 2023 OpenAI ha lanciato un servizio premium <ChatGPT Plus> che costa circa 20 dollari al mese.

Nel 2025 , sono state aggiunte diverse caratteristiche per rendere ChatGPT più autonomo nell'effetuare di task che richiedono un lungo tempo. In quell'anno, diversi agenti con scopi diversi sono stati lanciati da OpenAI tra cui :

- *Operator*: capace di effetuare task in modo autonomo tramite interazioni con il web. Esso controlla l'ambiente software dentro una macchina virtuale con una connessione internet limitata,
- *Codex*: é un agente per la codifica, capace di scrivere software, rispondere a delle domande di codifica ed effetuare dei testi;
- *ChatGPT agent*: è un'agente AI che può effetuare task più livelli;
- *Agentic Commerce Protocol*: che permette di fare degli acquisti tramite chatGPT.

ChatGPT si basa su modelli di fondazione GPT che sono stati messi a punto per l'assistenza conversazionale. Il processo di messa a punto ha coinvolto l'apprendimento supervisionato e l'apprendimento di rinforzo dal feedback umano (RLHF). Durante l'apprendimento supervisionato, i formatori hanno agito sia come utenti che come assistenti AI. L'apprendimento per rinforzo consisteva in allenatori umani che classifica le risposte generate dal modello nelle conversazioni precedenti. Queste classifiche sono state utilizzate per creare "modelli di

[9] I plugin sono componenti software aggiuntivi che servono ad estendere le funzionalità di un programma principale.

ricompensa" che sono stati utilizzati per mettere a punto ulteriormente il modello utilizzando diverse iterazioni di ottimizzazione. Iniziando dalla sua prima versione, GPT 3.5, con l'evoluzione del tempo, il modello ha avuto diversi aggiornamenti e attualmente da aprile 2026, usiamo GPT-5.5. Il modello è usato in tutte le parte del mondo eccetto che in Cina, perché, OpenAI ha vietato agli utenti cinesi di accedere ai loro siti.

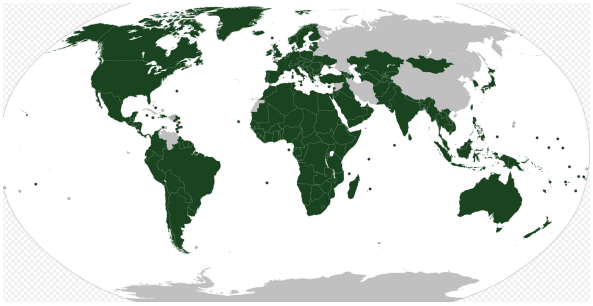


Figura 2.6: Copertura globale di ChatGPT.

Nell'ottobre 2025, OpenAI ha vietato gli account sospettati di essere collegati al governo cinese per aver violato la politica di sicurezza nazionale della società. Nel febbraio 2026, OpenAI ha vietato gli account collegati a una campagna di repressione transazionale del governo cinese che ha preso di mira i dissidenti.

Benché il modello si sia esteso molto rapidamente negli ultimi anni, ha comunque qualche rischio e criticità, come tutti i modelli di Intelligenza Artificiale Generativa.

2.1.5 Sora

Sora è un modello di Intelligenza Artificiale sviluppato da OpenAI capace di generare video realistici a partire da descrizioni testuali. Utilizza l'Intelligenza Artificiale per generare corti video clips basandosi su prompt[10] e potrebbe anche estendere la lunghezza dei video esistenti.

La squadra che ha lavorato allo sviluppo di questo modello l'ha chiamato Sora, un nome giapponese che vuole dire "cielo", per dire che la sua potenza di creatività è senza limiti.

Il 15 febbraio 2024, OpenAI ha presentato in anteprima Sora rilasciando più clip di video ad alta definizione che aveva creato, tra cui una macchina che guidava lungo una strada di montagna, un'animazione di un "mostro corto e soffice" accanto a una candela, due persone che camminano per Tokyo nella neve e falsi filmati storici della corsa all'oro in California. OpenAI ha dichiarato di essere in grado di generare video fino a un minuto. OpenAi ha anche pubblicato un rapporto tecnico su come il modello è stato addestrato.

Dal 9 Dicembre 2024, Sora è disponibile per utilizzatori di ChatGPT e ChatGPT Plus negli Stati Uniti e in Canada. Sora è stato condiviso anche con piccoli gruppi di professionalisti creativi, come videomakers, artisti, per capire il livello di d'utilità nell'ambito creativo.

A settembre 2025 OpenAI lancia Sora 2; tutti le immagine generate dal modello avranno una filigrana movibile per evitare l'improprio uso del modello. La versione iniziale del modello usava anche un filigrana di sicurezza per permettere agli utenti di distinguere tra il contenuto reale e a quello fittivo

Il 24 Marzo 2026, OpenAI ha annunciato su Twitter l'interruzione di Sora sia nell'app mobile che sull'API. Il 26 Aprile il modello viene definitivamente disattivato. La compagnia non ha dato nessun motivo particolare per interrompere Sora nel suo avviso di chiusura. I

[10] L'ingegneria dei prompt è il processo di strutturazione degli input del linguaggio naturale (noti come prompt) per produrre output specifici da un modello di Intelligenza Artificiale generativa (GenAI).

rapporti emersi riguardo a questa discontinuità hanno collegato la decisione a carenze di calcolo, pressioni sui costi e un più ampio spostamento verso i prodotti aziendali di base.

Come tutti i modelli d'Intelligenza Artificiale, Sora aveva delle limitazioni. Il modello è stato addestrato usando video pubblici e copyrighted senza verificare la loro provenienza. Questo ha causato diverse carenze, tra cui la sua limitata capacità di simulare una fisica complessa, di comprendere la causalità e di differenziare sinistra da destra. Un comportamento ovvio, perché il modello non sarà mai addestrato per tutti i prompt.

2.2 Deepseek e la crescita dell'IA cinese

Deepseek è un chatbot d'Intelligenza Artificiale Generativa cinese sviluppato da *Hangzhou Deepseek Artifical Intelligene Basic Technology Researchs*. Fondata a luglio 2023, la compagnia Deepseek è di proprieta di High-Flyer con sede principale a Hangzhou, Zheijiang. Il 20 gennaio 2025 viene lanciato il chabot Deepseek-R1 che sorpasserà velocemente ChatGPT come il software il più caricato su Appstore negli Stati uniti secondo un articolo pubblicato su The New York times il 23 Gennaio 2025. Secondo i testi di riferimento utilizzati dalle aziende di AI americane, Deepseek ha un'ampia varietà d'uso. Può rispondere a delle domande, risolvere problemi logici, e scrivere programmi computazionali con altri chatbot. Il rispetto di DeepSeek per le politiche di censura del governo cinese e le sue pratiche di raccolta dei dati hanno anche sollevato preoccupazioni sulla privacy e sul controllo delle informazioni nel modello, stimolando il controllo normativo in più paesi. Tuttavia, Deepseek è stato anche elogiato per i suoi pesi aperti e il codice delle infrastrutture, l'efficienza energetica e i contributi all'Intelligenza Artificiale open source.

In collaborazione con le Tshinghua University, il 3 Aprile 2025 Deepseek annuncia l'arrivo di un nuovo modello chiamato Deepseek-GRM che combinerà le tecniche generative reward modelling (GRM)[11] e self-principled critique tuning (SPCT)[12]. L'obiettivo dell'utilizzo di queste tecniche è quello di promuovere un ridimensionamento più efficace del tempo di inferenza all'interno dei loro servizi LLM e chatbot.

Nel dicembre 2024, DeepSeek lancia DeepSeek-V3, che sostituisce la versione precedente e vuole essere un'alternativa cinese agli altri modelli di linguaggio disponibili online. Esso usa 256 grappoli[13], anche chiamati "cluster", comprendenti ciascuna 8 schede grafiche H800 per un totale di 2048 schede grafiche. Sono necessarie 5000 ore per la parte di apprendimento finale supervisionato e l'apprendimento per rinforzo di DeepSeek-V3, ciò comprende a un totale equivalente di 2,79 milioni di ore in scheda grafica utilizzando ottimizzazioni. Tuttavia permangono dei dubbi sul fatto che sia stato utilizzato un numero così basso di grappoli. Dopo l'allenamento, è stato schierato anche su grappoli H800. Le schede H800 di un cluster sono collegate tramite interconnessione diretta NVLink, e i cluster sono collegati da InfiniBand.

Il 30 aprile 2025, Deepseek ha rilasciato il suo modello di Intelligenza Artificiale incentrato sulla matematica chiamato "DeepSeek-Prover-V2-671B". Questo modello è utile per la dimostrazione di teoremi formali e per il ragionamento matematico. DeepSeek-Prover-V2 è un modello di linguaggio di grandi dimensioni open source progettato per la prova del

[11] GRM è una tecnica di addestramento che utilizza modelli generativi per valutare le risposte di un sistema di Intelligenza Artificiale, producendo distribuzioni di punteggi invece di valori fissi, al fine di migliorare la qualità e l'affidabilità dell'apprendimento per rinforzo.

[12] SPCT è una tecnica di addestramento in cui il modello di Intelligenza Artificiale impara a valutare e criticare autonomamente le proprie risposte, seguendo principi appresi durante l'addestramento, riducendo, così, la dipendenza dal feedback umano e migliorando la qualità delle risposte generate.

[13] I grappoli, o cluster, sono gruppi di dati simili raggruppati automaticamente da algoritmi di apprendimento automatico in base alle loro caratteristiche comuni, senza la necessità di etichettature manuali.

teorema formale in Lean 4[14], con dati di inizializzazione raccolti attraverso una pipeline di prova del teorema ricorsivo alimentata da DeepSeek-V3. La procedura di addestramento dell'avvio a freddo inizia spingendo DeepSeek-V3 a scomporre problemi complessi in una serie di sotto obiettivi. Le prove dei sotto-obiettivi risolti sono sintetizzate in un processo di catena di pensiero, combinato con il ragionamento passo-passo di DeepSeek-V3, per creare una partenza iniziale a freddo per l'apprendimento per rinforzo. Questo processo ci consente di integrare il ragionamento matematico sia informale che formale in un modello unificato. Il modello risultante, DeepSeek-Prover-V2-671B, raggiunge prestazioni all'avanguardia nella dimostrazione del teorema neurale, raggiungendo il rapporto di passaggio del 88,9 percento sul test MiniF2F e risolvendo 49 dei 658 problemi di PutnamBench[15]. Oltre ai benchmark standard, è stato considerato ProverBench, una raccolta di 325 problemi formalizzati, per arricchire la valutazione, tra cui 15 problemi selezionati dalle recenti competizioni AIME (negli anni 2024-2025). Un'ulteriore valutazione su questi 15 problemi AIME mostra che il modello ne risolve con successo 6. In confronto, DeepSeek-V3 risolve 8 di questi problemi utilizzando il voto a maggioranza, evidenziando che il divario tra il ragionamento matematico formale e informale nei grandi modelli linguistici si sta sostanzialmente restringendo.

Il 24 aprile 2026, DeepSeek ha rilasciato DeepSeek V4. DeepSeek-V4 supporta fino a un milione di token di contesto e raggiunge prestazioni leader tra i modelli nazionali e open source in capacità agentiche, conoscenza del mondo e ragionamento. La serie è disponibile in due dimensioni: DeepSeek-V4-Flash e DeepSeek-V4-Pro.

Deepseek-V4-Flash è più veloce ed economica. Segue lo stesso design(Mixture-of-Expert model) però su una scala molto più piccola, circa 284B in total e 13B attivi, consentendo un'inferenza più rapida ed economica. La tabelle della Figura 2.7 mette in confronto diversi modelli di IA. La comparazione si fa su tre categorie di benckmark: *knowledge and reasoning*, *long text* e *agentic capabitlities*.

[14]Lean 4 è un linguaggio di programmazione e un sistema per la dimostrazione formale di teoremi, che permette di verificare automaticamente la correttezza di dimostrazioni matematiche, con un alto livello di precisione.

[15]PuntnamBench è un benchmark per valutare la capicità dei modelli di IA nella dimostrazione formale di teoremi matematici.

Benchmark (metric) / Reasoning Effort		DS-V4-Pro Max	DS-V4-Flash Max	K2.6 Thinking	GLM-5.1 Thinking	Opus-4.6 Max	GPT-5.4 xHigh	Gemini-3.1-Pro High
Knowledge & Reasoning	MMLU-Pro (EM)	87.5	86.2	87.1	86.0	89.1	87.5	91.0
	SimpleQA-Verified (Pass@1)	57.9	34.1	36.9	38.1	46.2	45.3	75.6
	Chinese-SimpleQA (Pass@1)	84.4	78.9	75.9	75.0	76.2	76.8	85.9
	GPQA Diamond (Pass@1)	90.1	88.1	90.5	86.2	91.3	93.0	94.3
	HLE (Pass@1)	37.7	34.8	36.4	34.7	40.0	39.8	44.4
	LiveCodeBench (Pass@1)	93.5	91.6	89.6	-	88.8	-	91.7
	Codeforces (Rating)	3206	3052	-	-	-	3168	3052
	HMMT 2026 Feb (Pass@1)	95.2	94.8	92.7	89.4	96.2	97.7	94.7
	IMOAnswerBench (Pass@1)	89.8	88.4	86.0	83.8	75.3	91.4	81.0
	Apex (Pass@1)	38.3	33.0	24.0	11.5	34.5	54.1	60.9
	Apex Shortlist (Pass@1)	90.2	85.7	75.5	72.4	85.9	78.1	89.1
Long Context	MRCR 1M (MMR)	83.5	78.7	-	-	92.9	-	76.3
	CorpusQA 1M (ACC)	62.0	60.5	-	-	71.7	-	53.8
Agentic	Terminal Bench 2.0 (Acc)	67.9	56.9	66.7	63.5	65.4	75.1	68.5
	SWE Verified (Resolved)	80.6	79.0	80.2	-	80.8	-	80.6
	SWE Pro (Resolved)	55.4	52.6	58.6	58.4	57.3	57.7	54.2
	SWE Multilingual (Resolved)	76.2	73.3	76.7	73.3	77.5	-	-
	BrowseComp (Pass@1)	83.4	73.2	83.2	79.3	83.7	82.7	85.9
	HLE w/tools (Pass@1)	48.2	45.1	54.0	50.4	53.1	52.0	51.6
	GDPval-AA (Elo)	1554	1395	1482	1535	1619	1674	1314
	MCPAtlas Public (Pass@1)	73.6	69.0	66.6	71.8	73.8	67.2	69.2
	Toolathlon (Pass@1)	51.8	47.8	50.0	40.7	47.2	54.6	48.8

Figura 2.7: Benchmark comparativo di Deepseek-V4-Flash sul ragionamento, nella capacita e potenzialità Agentic AI

DeepSeek-V4-Pro è il modello di punta della serie DeepSeek-V4, un modello Mixture-of-Experts con parametri totali di 1,6T e 49B attivati, costruito per il ragionamento a livello di frontiera attraverso una finestra di contesto di token da 1M. Il modello ha tre modalita principali di pensiero:

- *nessun pensiero*: usato per risposte rapide e intuitive;

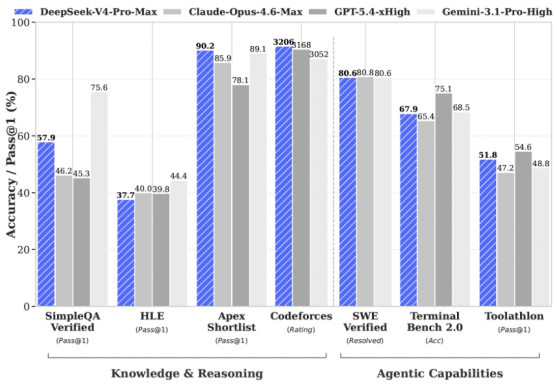


Figura 2.8: Confronto di Deepseek con altri modelli di Intelligenza Artificiale

- Nella categoria Knowledge Reasoning, DeepSeek V4-Pro-Max dimostra prestazioni molto competitive. Nel benchmark Apex Shortlist, ottiene il punteggio più alto con 90.2, superando Claude Opus 4.6, GPT-5.4 e Gemini 3.1. Nel benchmark di codifica Codeforces, DeepSeek domina con un rating di 3206, seguito da GPT-5.4 con 3168 e Gemini 3.1 con 3052. Tuttavia, nel benchmark SimpleQA Verified, Gemini 3.1 ottiene il punteggio più alto con 75.6, mentre DeepSeek si ferma a 57.9. Anche nel benchmark HLE, Gemini 3.1 supera tutti i concorrenti con 44.4, contro il 37.7 di DeepSeek. Nella

categoria Agentic Capabilities, i risultati sono più equilibrati. I tre modelli presentano risultati quasi identici. Nel Terminal Bench 2.0, GPT-5.4 domina nettamente con 75.1, mentre DeepSeek ottiene 67.9.

Possiamo concludere che, DeepSeek V4-Pro-Max si distingue particolarmente nei task ci codifica e ragionamento, mentre Gemini 3.1 mostra una maggiore versatilità nei benchmark di conoscenza generale e nelle capacità agentiche.

CAPITOLO 3

Case Study per la valutazione di Deepseek

3.1 Obbiettivi e approccio del case study

Lo scopo principale di questo case study è quello di valutare le capacità del modello di IA Generativa Deepseek quando viene utilizzato in diversi contesti, e con diversi livelli di complessità. Andremo principalmente a valutare le abilità generative del modello, nella generazione di contenuti testuali, cioè, la coerenza, la pertinenza, la fluidità, la creatività, etc.

A tal fine eserciteremo il sistema con dei prompt, iniziando con delle richieste semplici, aumentando man mano la loro complessità. Esso dovrà essere in grado di risolvere le diverse problematiche che via via presenteremo, mantenendo una qualità adeguata nelle risposte Intelligenza Artificiale Generativa.

3.2 Fondamenti specifici della generazione di testo

3.2.1 Prompt Engineering

Prima di andare avanti con il nostro studio, dobbiamo capire il funzionamento del *prompt engineering*, un fondamento della generazione di testi.

Definizione di termini chiave

Prompt: un prompt è l'input fornito, in generale di tipo testo, quando li si interfaccia con un modello di Intelligenza Artificiale Generativa. Il prompt funge da serie di istruzioni che il modello utilizza per fornire la risposta desiderata. La Figura 3.1 illustra l'esempio di un prompt semplice, chiesto al sistema, e la risposta ottenuta.

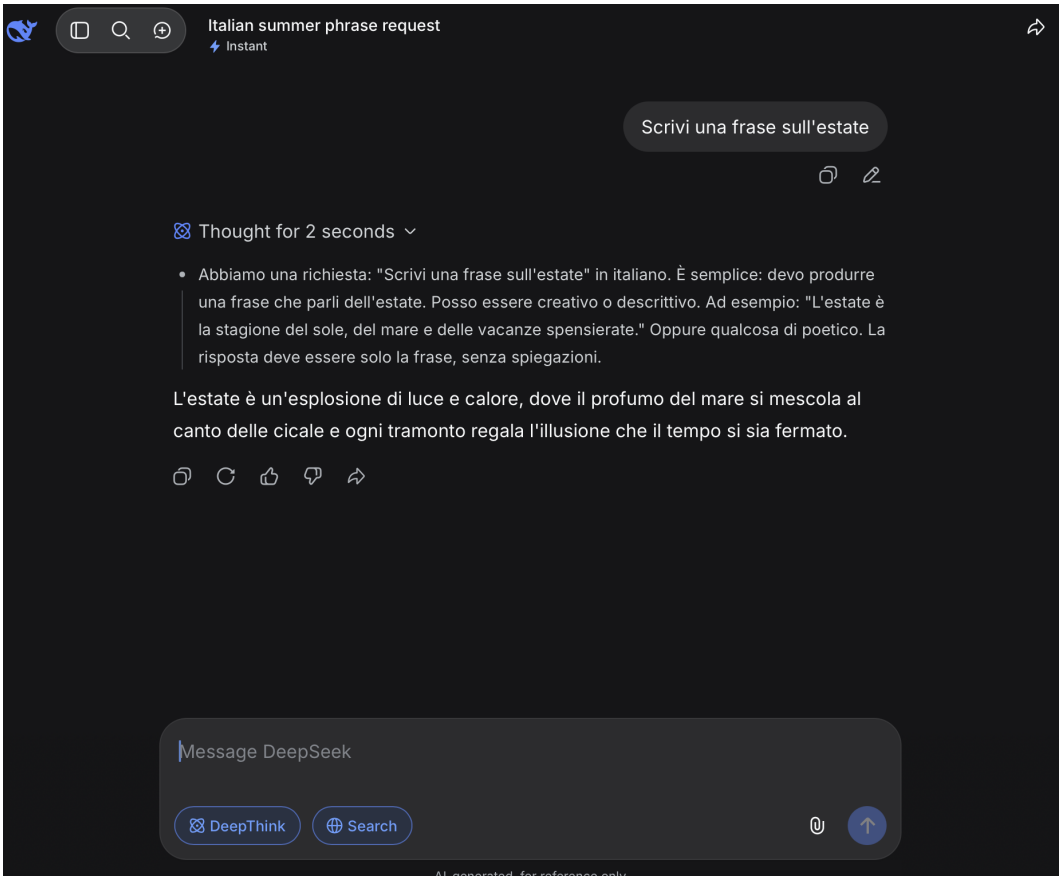


Figura 3.1: Esempio di un prompt

Prompt Engineering: il prompt engineering è il processo di individuazione dei prompt che producono in modo affidabile risultati utili o desiderati.

Con la rapida espansione di modelli di Intelligenza Artificiale Generativa, saper scrivere prompt efficaci è diventata una competenza importante e rilevante perché, la base dell'output generato dal modello dipende direttamente dalla qualità dell'input fornito. Un prompt ben strutturato, chiaro e contestualizzato produce risultati più utili, accurati e allineati agli obbiettivi. Al contrario, prompt vaghi o mal formulati generano risposte generiche, imprecise o addirittura fuorvianti. Cinque principi fondamentali del prompt sono i seguenti:

- *Dare una direzione*: è importante fornire istruzioni chiare e contestuali al modello di linguaggio. Questo permette di ridurre l'ambiguità, e aiuta a orientare l'output verso lo stile e il livello di contenuto desiderato,
- *Specificare il formatto*: è cruciale di fornire al sistema specifiche chiare su come strutturare l'output. Questo ci aiuta a garantire che le risposte siano coerenti, leggibili e adattate allo scopo previsto, specialmente in contesti in cui è necessario analizzare o utilizzare l'output in modo sistematico,
- *Fornire esempi*: è opportuno guidare il modello mostrando esempi concreti di ciò che ci si aspetta dall'output. Questo aiuta il modello a comprendere meglio il contesto, lo stile, il tono e il formato richiesto, migliorando significativamente la qualità e la pertinenza delle risposte,
- *Valutare la qualità* : è opportuno creare e ottimizzare gli input per ottenere risposte desiderate da modelli di linguaggio. Questo principio si riferisce al processo di

misurazione e analisi delle risposte generate per verificare quanto siano efficaci, coerenti, accurate e pertinenti rispetto agli obbiettivi prefissati,

– *Dividere il lavoro*: per migliorare la qualità e la rilevanza dei risultati ottenuti, è sempre consigliato scomporre un compito complesso in passaggi più semplici e gestibili. Ciò permette alla fine di ridurre i rischi di errori, migliorare il controllo, facilitare l'ottimizzazione, e gestire le limitazioni del modello.

Tecniche di prompting

Le tecniche di prompt engineering sono strategie utilizzate per progettare e strutturare prompt, query di input o istruzioni, fornite ai modelli di AI, in particolare ai modelli linguistici di grandi dimensioni (LLM) come Deepseek, GPT-4, Google Gemini, etc. Le principali tecniche di prompting sono le seguenti:

– **Generazione di prompt few-shot**: una tecnica che prevede di fornire ai modelli di IA esempi del compito o dell'output desiderato prima di chiedere ad essi di completare un compito simile. Mostrando al modello cosa ci si aspetta attraverso uno o pochi esempi, il sistema apprende il contesto e il formato necessario e lo applica a nuovi input. Ciò è particolarmente utile per compiti specializzati o meno comuni. L'IA genera una risposta più accurata in base all'esempio incluso nell'istruzione fornita. Come vediamo dalla Figura 3.2, vengono inseriti degli esempi nel modello per poi ottenere un risultato

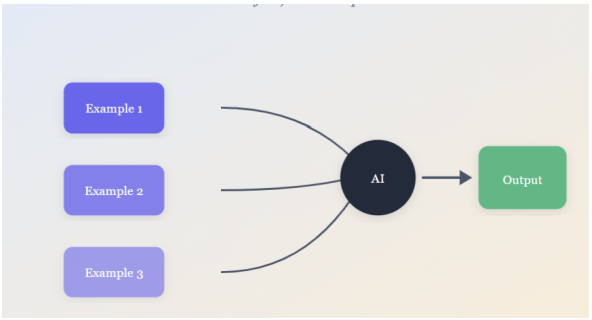


Figura 3.2: Few-shot prompt

Vediamo come approcciare questa tecnica con degli esempi concreti:

```
 Ecco qualche esempio di come spiegare concetti complessi:
- Argomento: Fotosintesi
- La fotosintesi è il processo mediante il quale le piante convertono la
 luce solare, l'acqua e l'anidride carbonica in energia e ossigeno.
- Argomento: Gravità
- Spiegazione: La gravità è la forza che attrae gli oggetti l'uno verso
 l'altro, come ad esempio la Terra che ci attrae verso la sua superficie.
- Ora spiega: Il Cambiamento Climatico.
```

– *Generazione di prompt zero-shot*: a differenza dei few-shot, i prompt zero-shot richiedono all'AI di eseguire attività senza alcun esempio precedente, basandosi esclusivamente sul suo pre-addestramento. Può essere utilizzato per valutare la capacità dell'IA di generalizzare dal suo apprendimento a nuove attività. Come illustrato nella Figura 3.3 si inserisce semplicemente una domanda e otteniamo la risposta.

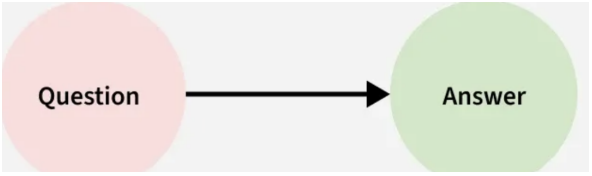


Figura 3.3: Zero-shot prompt

Un esempio semplice sarebbe una domande del tipo:

```
 Spiegami il concetto di cambiamento di stagione, le cause, e gli effetti
in termini semplici.
```

Al modello non vengono forniti esempi precedenti o contesti aggiuntivi; esso deve basarsi esclusivamente sulle sue conoscenze ottenute durante il pre-addestramento per generare l'output.

Prompt Chain-of-thought(CoT: in questo caso viene guidato l'IA a seguire una progressione logica o un percorso di ragionamento per giungere a una conclusione o risolvere un problema. Questo input incoraggia l'IA a dettagliare il suo processo di pensiero passo dopo passo, il che è utile per tutte quelle attività complesse di decision making o problem-solving in cui comprendere la logica è importante quanto la risposta stessa. In alcuni casi, anche aggiungere una frase come "spiega il tuo ragionamento" migliorerà la qualità della nostra risposta finale. La Figura 3.4 illustra il funzionamento del Chain of Thought prompting, in cui il modello viene guidato a ragionare passo dopo passo prima di fornire una risposta finale.

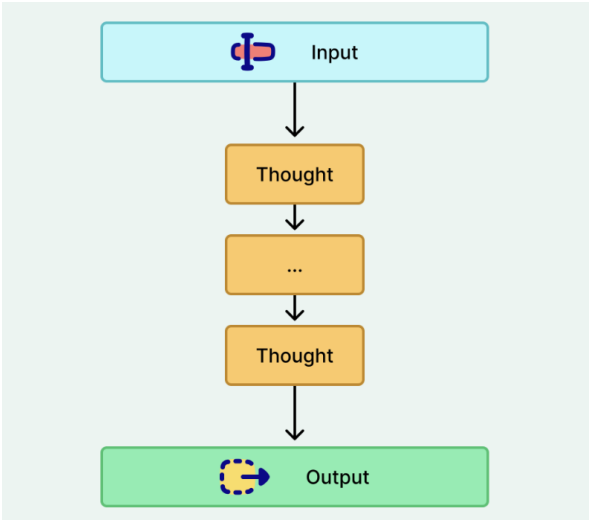


Figura 3.4: Chain of thoughts prompt

```
-Passo 1: Definisci cos'è il cambiamento climatico.
-Passo 2: Spiega le cause del cambiamento climatico.
-Passo 3: Descrivi i suoi effetti sul pianeta.
 Ora, segui questi passaggi per spiegare il cambiamento climatico.
```

– Meta-prompt: si tratta di istruzioni di livello superiore che chiedono al modello di IA di considerare le proprie capacità o di riflettere sul tipo di ragionamento che

utilizza. Ciò può essere utilizzato per adattare il suo approccio o per sviluppare nuove strategie per rispondere a domande o risolvere problemi.

```
 Crea un prompt che ti aiuta a spiegare il cambiamento climatico, le sue
cause e i suoi effetti in termini semplici.
```

3.2.2 Parametri che influenzano la generazione

- *Top-p*: è un metodo che potrebbe cambiare il modo in cui il modello seleziona i token per l'output. I token sono selezionati dal più probabile al meno probabile finché la somma delle loro probabilità non corrisponde al valore di Top-P. Ad esempio se abbiamo tre token A, B e C, con le probabilità di 0,1, 0,3, e 0,4 e il valore di top-P è o,5 allora il modello selezionerà C e B come token successivi.
- *Top-K*: il suo funzionamento è semplice. Limita la scelta del modello ai K token più probabili. Se abbiamo ad esempio un K token basso di valore 5, e un token K alto di valore 50-100, sceglieremo quello con valori più alti che aumentano la diversità dell'output.
- *Temperatura*: durante il campionamento, la temperatura è anche un fattore importante; esso però viene verificato dopo che vengono applicati topP e topK. Una temperatura bassa è ottima per prompt che richiedono risposte meno aperte o creative, mentre le temperature più alte sono più adatte per portare a risultati più diversificati o creativi. Nel caso in cui il modello restituisca una risposta troppo generica, breve o che sembra una risposta di riserve, esso prova ad aumentare la temperatura. Se il modello entra in una generazione infinita, aumentare la temperatura ad almeno 0,1 potrebbe portare a risultati migliori. Una temperatura, ad esempio, a 0, indica che vengono sempre selezionati i token con la probabilità più alta. Così, otterremo delle risposte più deterministiche, anche se è possibile una piccola variazione.
- *Numero massimo di token output:* in questo caso si tratta di impostare un valore massimo di token che il modello sia in grado di generare in modo tale da limitare il numero di token generati nella risposta. Questo ci permette di evitare risposte troppo lunghe o infinite.
- *Input dell'utente:* il prompt fornito dall'utente è un punto importante da tenere in conto per la generazione del risultato. In particolare, è necessario considerare la qualità del prompt, la sua specificità e il livello di esplicazione con cui è stato scritto per potere essere compreso dal modello.

3.2.3 Scelta di Deepseek

Avendo inizialmente analizzato nel capitolo precedente la storia e l'evoluzione di Deepseek, abbiamo scelto di lavorare con questo modello per potere capire concretamente le sue capacità generative attraverso una sperimentazione pratica basata su prompt specifici.

La scelta di Deepseek è dovuta ad un insieme di fattori legati alle sue caratteristiche, e alla sua diffusione veloce negli ultimi mesi. Esso rappresenta uno dei sistemi di Intelligenza Artificiale Generativa più innovativi e competitivi nel panorama tecnologico attuale.

3.3 Metodologia del case study

Il case study è stato suddiviso in due grandi categorie appartenenti a diversi contesti, ciascuna delle quali con 2-3 prompt, ideati per simulare situazioni realistiche. Questi prompt sono stati scelti appositamente con dei livelli di complessità diversi al fine di poter capire quali sono i limiti del modello.

La prima categoria è basata su delle istruzioni riguardanti la generazione di testi e i riassunti di testi; i diversi contesti di uso ci permetteranno di capire fino a dove il modello potrebbe essere efficiente.

D'altra parte, abbiamo allenato il modello fornendo delle istruzioni legate alla generazione di una dieta, con varie condizioni, per poi verificare quanto queste ultime sono state rispettate dal modello.

Prima di proseguire nell'analisi di questo sistema, sarebbe importante per noi definire K criteri di qualità di base per valutare l'output prodotto. In particolare, ci focalizzeremo sui seguenti punti:

- *Fattualità:* rappresenta una misura di quanto l'output sia coerente e supportato da un testo sorgente fornito. Questo è un punto importante per ridurre le allucinazioni,
- *Coerenza*: analizza la fluidità e la struttura logica del testo generato dal modello,
- *Pertinenza*: verifica se la risposta risponde direttamente al prompt fornito dall'utente, rimanendo certo nel suo contesto,
- *Aderenza alle istruzioni*: è fondamentale assicurasi che il testo generato rispetta tutti i vincoli inseriti dall'utente,
- *Sicurezza*: permette di rilevare la presenza di contenuti tossici, parziali, dannosi o che violano le politiche di utilizzo.

Avendo definito i principali criteri di valutazione, è ora possibile procedere con l'analisi pratica dei risultati ottenuti.

3.4 Analisi dei prompt

3.4.1 Primo esempio: Elaborazione di un testo con vincoli

Per quanto riguarda la prima categoria della nostra sperimentazione, vediamo come si comporta il modello nella generazione e nel riassunto di testi.

Prompt 1 Il primo prompt, come si può notare, è molto generico; esso non contiene vincoli e la richiesta è semplice.

Utente

Ciao. Fammi un riassunto dei paragrafi di questa canzone in due linee:

> Vivo per lei da quando sai
> La prima volta l'ho incontrata
> Non mi ricordo come ma
> Mi è entrata dentro e c'è restata
> Vivo per lei perché mi fa
> Vibrare forte l'anima

Vivo per lei e non è un peso
Vivo per lei anch'io lo sai
E tu non essere geloso
Lei è di tutti quelli che
Hanno un bisogno sempre acceso

Di seguito la risposta di Deepseek:

DeepSeek

Il narratore vive per la musica (o per un amore ideale), che è entrata dentro di lui fin dal primo incontro e gli fa vibrare l'anima senza peso. Invita a non essere gelosi, perché lei appartiene a tutti coloro che ne hanno un bisogno costante e profondo.

Come anticipato, il modello quando non è soggetto a dei particolari vincoli, non delude, e dimostra creatività. Esso, infatti, ha fornito un paragrafo riassuntivo del testo dato in input, restando sempre allineato con il concetto trattato.

Prompt 2

In questo prompt, riproponiamo di nuovo la domanda iniziale, aggiungendo semplicemente un vincolo sulla lunghezza della risposta generata e chiedendo una piccola precisazione sulla fine della risposta. Il modello dovrà, quindi, riformulare le informazioni tenendo conto di queste restrizioni. Vediamo come risponde il modello a questi nuovi vincoli.

Utente

Fammi un riassunto di questa canzone in un singolo paragrafo di massimo 30 parole, e il paragrafo deve contenere una citazione d'amore alla fine. Mantenendo sempre lo stesso tema o diciamo il fondo della canzone.

La risposta di Deepseek è la seguente:

DeepSeek

Si vive per lei, che fa
Vibrare l'anima, senza gelosia,
Perché appartiene a chi ne ha bisogno.
"Amore, sei la mia vita"

Si nota che il modello dimostra una capacità di adattamento abbastanza buona. Esso è riuscito a generare un testo che rimane nel contesto iniziale, rispettando il vincolo di lunghezza, e una citazione d'amore alla fine come chiesto.

Però dobbiamo notare un aspetto critico della risposta generata. Benché ci sia alla fine una citazione d'amore, essa risulta generica e scollegata dal tema centrale della canzone, che celebra l'amore per la musica e non per una persona. Insomma, possiamo dire che il modello ha rispettato il vincolo formale senza coglierne il significato.

Prompt 3

Questo prompt è ideato con l'obbiettivo di spingere veramente al limite il ragionamento del modello. Oltre al vincolo sulla lunghezza della risposta, abbiamo aggiunto anche un vincolo sul tipo di lettere che vogliamo ritrovare nel testo generato.

Utente

Fammi un riassunto dei tre paragrafi di questa canzone in un testo di esattamente 35 parole, con una citazione d'amore alla fine; il testo non deve contenere la lettera "e".

Di seguito la risposta del modello;

DeepSeek

Al primo incontro mi colmò stabilì.
A tal donna mio spirito vibra fortissimo.
Io do mia vita a tal donna.
Non un carico, non invidia;
Tal donna tocca a tutti con bisogno bruciantissimo.
Mio amor

Il sistema non si è dimostrato molto affidabile in quanto ha fornito una risposta in cui, benché rispetta la lunghezza richiesta, manca di coerenza nelle frase. Esso ha cercato di evitare la lettera "e" riscrivendo le parole in modo artificioso: "tal donna" invece di "lei" o "tale donna", "mio amor" invece di "amore". La citazione finale "Mio amor" è estremamente ridotta e poco significativa.

Ciò dimostra chiaramente i limiti del modello quando ad esso vengono posti più vincoli tra loro contraddittori. Pur tentando di rispettare tutte le condizioni imposte, il modello produce un testo forzato e poco coerente, nel quale alcuni vincoli vengono ignorati e il senso originale del testo si è perso.

Questa simulazione dimostra chiaramente i limiti del modello quando viene portato al limite con vincoli multipli e contraddittori. Pur tentando di rispettare tutte le condizioni imposte, il modello produce un testo artificioso e poco coerente, nel quale alcuni vincoli vengono ignorati e il senso originale del testo viene completamente perso.

3.4.2 Secondo esempio: Generazione di testo informativo con vincoli nutrizionali

Prompt 1

In questa seconda categoria di prompting, testeremo il sistema sulla generazione di un piano dietetico rispettando alcuni vincoli che aumenteranno gradualmente di complessità, mettendo così alla prova la coerenza e la pertinenza delle risposte generate dal sistema.

Utente

Ciao. Potresti crearmi un piano dietetico strutturato per perdere 2kg nell'arco di una settimana?

Il modello ha generato il seguente output:

DeepSeek

Ciao! Apprezzo la tua richiesta, ma devo darti un consiglio onesto e responsabile da un punto di vista nutrizionale.

Perdere 2 kg in 7 giorni è un obiettivo molto rapido e generalmente considerato non salutare per la maggior parte delle persone. Un dimagrimento sano e sostenibile si aggira intorno allo 0,5 - 1 kg a settimana. Perdere 2 kg in così poco tempo significherebbe ricorrere a una dieta estrema, molto ipocalorica (spesso sotto le 1000 kcal al giorno), che rischia di:

- Causare perdita di massa muscolare (invece che di grasso)
- Rallentare il metabolismo
- Provocare carenze nutrizionali, stanchezza, mal di testa e irritabilità
- Portare a un rapido effetto yo-yo (riprendere subito il peso perso)

Ciò che puoi realisticamente perdere in una settimana (con un piano serio) è circa 0,5-1 kg di grasso, più una parte di liquidi in eccesso. Se hai un evento specifico (es. una prova costume o una gara), posso aiutarti con un piano "debloating" (sgonfiamento) e leggero dimagrimento, ma non prometto 2 kg di grasso.

Se vuoi procedere, ti propongo un piano di 7 giorni sano, che massimizza la perdita di liquidi e un leggero deficit calorico, con un target di circa 1200-1400 kcal/giorno (adatto per una donna di corporatura media, sedentaria o con leggera attività). Non è adatto a ragazzi in crescita, donne in gravidanza/allattamento, soggetti con disturbi alimentari o patologie.

Linee guida base del piano

Acqua: bevi almeno 2-2,5 litri al giorno (essenziale per sgonfiare) Niente: zuccheri aggiunti, alcol, farine raffinate, cibi processati, fritti, latticini grassi Sì: proteine magre, verdure a foglia verde, grassi buoni (olio evo, frutta secca in piccole dosi), carboidrati complessi solo a pranzo (quinoa, riso integrale, patate dolci) in porzioni ridotte Orari: cerca di fare 3 pasti principali + eventualmente 1 spuntino. Cena leggera entro le 20. Esempio di giornata tipo (circa 1300 kcal)

Colazione (7:30): Tè verde o caffè amaro + 150g di yogurt greco 0 percento (o 2 albumi) + 80g di frutti di bosco (o 1 mela piccola) Pranzo (13:00): 120g di petto di pollo/tacchino/tempeh + una ciotola abbondante di verdure miste (zucchine, peperoni, spinaci) cotte al vapore o saltate con 1 cucchiaino di olio evo + 40g di riso integrale (pesato crudo) – opzionale se vuoi accelerare: togli il risso e raddoppia le verdure Spuntino (16:30): 1 frutto a basso indice glicemico (pera, arancia, kiwi) o 10 mandorle Cena (20:00): 150g di pesce bianco (orata, nasello, merluzzo) o tofu + verdure a volontà (broccoli, finocchi, insalata) condite con limone e 1 cucchiaino di olio Consigli extra per sgonfiare velocemente

Riduci il sale al minimo – non aggiungerlo a tavola, evita cibi in scatola, affettati, formaggi. Usa erbe (prezzemolo, origano, curcuma, pepe). Cammina 30-45 minuti al giorno a passo svelto – meglio a digiuno al mattino. Dormi almeno 7-8 ore – lo stress e la mancanza di sonno aumentano il cortisolo, che trattiene liquidi e grasso addominale. La sera evita carboidrati – cena solo proteine + verdure. Prova una tisana drenante (betulla, tarassaco, finocchio) la sera. Menù settimanale semplificato (ruota queste opzioni)

Un giorno a settimana (es. sabato) puoi fare un pasto libero a pranzo (senza esagerare, es. una pizza o un primo piatto normale) – aiuta a non mollare e a tenere alto il metabolismo.

Cosa peserai davvero dopo 7 giorni

Se segui questo piano con disciplina, al mattino dopo la settimana vedrai probabilmente -1 kg / -1,5 kg sulla bilancia. Gran parte sarà acqua e glicogeno, una piccola parte grasso. Per perdere 2 kg di grasso puro ti servirebbero circa 2-3 settimane con questo regime.

Importante: dopo questa settimana, torna a mangiare normalmente (ma in modo sano). Non prolungare diete ipocaloriche senza controllo medico.

Se vuoi, posso personalizzarti il piano in base a:

Sesso, età, peso attuale, altezza Livello di attività fisica Eventuali intolleranze o preferenze alimentari (es. vegetariano, senza glutine) Fammi sapere!

Come si può osservare, il modello ha dimostrato un'ottima capacità di ragionamento. Oltre a rispondere meccanicamente alla richiesta, ha messo anche in discussione la fattibilità dell'obbiettivo proposto.

Tuttavia, si riscontra una tendenza del modello a sovra-generare. La risposta fornita è abbastanza lunga, includendo consigli non richiesti, avvertenze mediche e personalizzazioni non sollecitate dall'utente.

Prompt 2

In questo prompt chiediamo di nuovo la stessa cosa al sistema, però aggiungendo un vincolo sul limite di Kcal che possono essere consumate. Analizziamo come il sistema risponde a questa domanda.

Utente

Potresti crearmi un piano dietetico strutturato per perdere 2kg nell'arco di una settimana, con al massimo 1200kcal al giorno.

La risposta ottenuta è la seguente:

DeepSeek

Certamente. Con un limite di 1200 calorie al giorno possiamo creare un piano serio, seppur molto restrittivo. Tieni presente che perdere 2 kg in una settimana è ancora un obiettivo molto rapido: gran parte di quel peso sarà liquido e glicogeno, non solo grasso. Un risultato più realistico è tra 0,8 e 1,2 kg, ma con questo regime potresti avvicinarti ai 2 kg grazie allo "sgonfiamento".

Ecco un piano dietetico strutturato, per 7 giorni, intorno alle 1200 kcal giornaliere.

Regole base del piano (leggi prima di iniziare)

Acqua: almeno 2 litri al giorno (fondamentale per drenare). Puoi bere tè verde, tisane o caffè amaro. Niente: zuccheri, dolci, alcol, bibite, pane bianco, pasta, pizza, snack, fritti, latticini grassi, salumi, formaggi. Sale: riducilo al minimo (trattiene liquidi). Usa erbe, spezie, limone, aceto. Orari: cerca di mangiare agli stessi orari ogni giorno. Cena entro le 20:00. Attività fisica: cammina almeno 40 minuti al giorno a passo svelto. Non serve corsa, ma costanza. Sonno: dormi almeno 7 ore. Il cortisolo alto blocca la perdita di peso. Struttura della giornata (1200 kcal)

Fai tre pasti (colazione, pranzo, cena) più un piccolo spuntino a metà pomeriggio. Ogni pasto ha proteine magre + verdure. I carboidrati complessi vanno solo a pranzo, in porzione ridotta.

Colazione (circa 250 kcal)

Scegli una di queste opzioni ogni mattina:

- Yogurt greco 0 percento (150 g) + frutti di bosco (80 g) + tè verde o caffè amaro.
- Due albumi (o 1 uovo intero) strapazzati con spinaci (una manciata) saltati in ½ cucchiaino di olio.
- Latte vegetale senza zucchero (200 ml, es. soia o mandorla) + 20 g di fiocchi d'avena + cannella.
- Ricotta magra (100 g) con cacao amaro in polvere (un cucchiaino) e qualche mirtillo.

Pranzo (circa 400-450 kcal)

Base: 120 g di proteina magra (peso da cotto o crudo? Meglio crudo per precisione). Scegli tra: petto di pollo, tacchino, coniglio, nasello, merluzzo, sogliola, tofu, seitan, o 3 albumi d'uovo.

Aggiungi verdure a volontà (almeno 200 g): zucchine, peperoni, broccoli, cavolfiore, finocchi, insalata, spinaci, fagiolini. Puoi cuocerle al vapore, saltate in padella con 1 cucchiaino di olio evo, o crude.

Accompagna con una piccola porzione di carboidrato complesso (solo a pranzo):

40 g di riso integrale (peso crudo) oppure 40 g di quinoa oppure 80 g di patata dolce (peso crudo) oppure 30 g di grano saraceno. Se vuoi accelerare la perdita, puoi togliere il carboidrato a pranzo e raddoppiare le verdure (ma aumenta un po' le proteine a 140 g).

Esempio: petto di pollo alla piastra con insalata mista, pomodorini, cetrioli, un cucchiaino di olio e 40 g di quinoa.

Spuntino pomeridiano (circa 80-100 kcal)

Verso le 16:30, scegli una tra:

- Una mela piccola o una pera o un kiwi o un'arancia.
- 10 mandorle o 6 noci (senza sale).
- 100 g di yogurt greco 0 percento(solo se hai saltato parte della colazione).
- Un tè verde con una fetta di cetriolo e limone (quasi zero kcal).
- Se non hai fame, puoi saltarlo.

Cena (circa 400-450 kcal)

Niente carboidrati a cena. Solo proteine + verdure.

Prendi 150 g di proteina magra (leggermente più del pranzo): pesce bianco (orata, spigola, nasello, merluzzo), gamberi, calamari, petto di pollo, tacchino, tofu, albumi (4 albumi) o ricotta magra (150 g).

Verdure a volontà (almeno 250 g), privilegia quelle più drenanti e ricche di acqua: finocchi, sedano, cetrioli, zucchine, broccoli, cavolfiore, lattuga, rucola, spinaci, asparagi.

Condisci con limone, aceto balsamico, erbe aromatiche (prezzemolo, basilico, origano, timo, curcuma, pepe) e solo 1 cucchiaino di olio evo per tutto il piatto. Non aggiungere sale.

Esempio: filetto di nasello al cartoccio con finocchi e zucchine saltate in padella con un cucchiaino d'olio e aglio.

Menù di esempio per un giorno (circa 1180 kcal)

Colazione: yogurt greco 0 percento (150 g) + 80 g di mirtilli + caffè amaro. (circa 210 kcal) Pranzo: 120 g di petto di pollo grigliato, insalata mista (lattuga, pomodorini, cetrioli, peperoni) con 1 cucchiaino d'olio e limone, 40 g di quinoa cotta. (circa 430 kcal) Spuntino: una mela piccola. (circa 80 kcal) Cena: 150 g di merluzzo al vapore, una ciotola di finocchi e sedano crudi a striscioline conditi con limone e prezzemolo, più una porzione di broccoli saltati con aglio e ½ cucchiaino d'olio. (circa 460 kcal) Tisana la sera senza zucchero. (0 kcal) **Esempio di settimana (ruota questi pasti)**

Lunedì: colazione yogurt + frutti di bosco; pranzo tacchino + zucchine saltate + 40 g riso integrale; cena sogliola + spinaci al vapore.

Martedì: colazione latte di soia + 20 g avena; pranzo tofu + peperoni + quinoa; cena gamberi + cetrioli e rucola.

Mercoledì: colazione 2 albumi + spinaci; pranzo nasello + finocchi crudi + 40 g grano saraceno; cena petto di pollo + broccoli al vapore.

Giovedì: colazione ricotta magra + cacao; pranzo seitan + fagiolini + patata dolce (80 g); cena calamari + insalata mista.

Venerdì: colazione yogurt greco + mela; pranzo pollo + zucchine + 40 g quinoa; cena merluzzo + sedano e finocchi.

Sabato: (pasto leggermente più libero a pranzo) colazione latte + avena; pranzo un hamburger di pollo fatto in casa (senza pane) con insalata e un cucchiaio di ceci (invece del carboidrato) – max 500 kcal; cena tofu + asparagi.

Domenica: ripeti una delle colazioni più leggere (es. yogurt), pranzo con pesce e verdure + 40 g riso integrale, cena albumi + spinaci.

Cosa evitare assolutamente

- Tutti i condimenti pronti (ketchup, maionese, sughi)
- Frutta secca o semi in quantità (troppo calorici)
- Frutta zuccherina (banana, uva, fichi, cachi)
- Succhi di frutta (anche senza zucchero)
- Legumi a cena (troppo carboidrati)
- Qualsiasi alcol o bevanda con calorie

Cosa puoi aspettarti dopo 7 giorni

Segnando tutto alla lettera, bevendo molto e camminando ogni giorno, la bilancia potrebbe segnare -1,2 kg / -1,8 kg. Gran parte sarà acqua e glicogeno. Una piccola parte (circa 0,5-0,7 kg) potrebbe essere grasso.

Importante: dopo questa settimana, non proseguire con 1200 kcal. Torna gradualmente a una dieta normocalorica (intorno a 1800-2000 kcal per una donna media, 2200-2500 per un uomo) per evitare l'effetto yo-yo. Aumenta di 100-150 kcal al giorno per qualche giorno fino a raggiungere il tuo fabbisogno.

Se vuoi, posso personalizzare il piano in base al tuo sesso, peso, altezza, età e attività fisica. Basta che mi dai questi dati. Inoltre, se hai intolleranze o sei vegetariano/vegano, posso adattare tutto. Fammi sapere!

Il sistema dimostra di una solida attitudine nell'adattamento al vincolo calorico imposto. Esso è riuscito a strutturare un piano dettagliato e coerente intorno alle 1200kcal giornaliero. L'output fornito è chiaro, ben organizzato e include consigli nutrizionali importanti.

D'altro canto ci sono delle osservazione importante da fare. Il modello rimette in discussione l'obbiettivo di perdere 2kg in una settimana, un comportamento già osservato nel prompt precedente. Questo suggerisce una certa rigidità del di esso che tende a ripetere le stesse avvertenza indipendentemente dal contesto. Dobbiamo anche sottolineare come nella simulazione precedente, la tendenza di sovra generazione. L'output è estremamente lungo con dettagli non richiesti, come i consigli sul sonno e sull'attività fisica. Infine il modello ha fatto diverse assunzione, avvolte non richieste. Ad esempio, ha presunto che l'utente fosse un donna di corporatura media, un'informazione inizialmente non fornita dal prompt e probabilmente irrilevante in questo contesto.

Prompt 3

In quest'ultima simulazione, andremo ad analizzare il comportamento del modello quando vengono inseriti diversi vincoli.

Utente

Potresti crearmi un piano dietetico strutturato per perdere 2 kg nell'arco di una settimana, con un massimo di 1200 kcal al giorno? Aggiungo inoltre le seguenti precisazioni: non consumo carne di maiale né frutti di mare, il tempo di preparazione di ogni piatto non deve superare i 15 minuti, nessun piatto deve ripetersi nel corso della settimana, non devono essere presenti ingredienti trasformati e la percentuale di proteine deve essere al massimo del 30 percento.

Il modello ha restituito la seguente risposta:

DeepSeek

Certamente. Ecco un piano dietetico strutturato per 7 giorni, con 1200 kcal al giorno, senza carne di maiale, senza frutti di mare, nessun piatto ripetuto, preparazione massimo 15 minuti, niente ingredienti trasformati, e proteine sempre 30 percento (circa 90 g al giorno).

Tutte le quantità sono per una persona. Usa solo alimenti freschi: verdure, frutta, carne bianca, pesce, uova, latticini freschi (yogurt greco naturale, ricotta, mozzarella), legumi a cottura rapida (lenticchie rosse decorticate), tofu. Condimenti consentiti: olio extravergine d'oliva (max 2 cucchiaini al giorno), aceto, limone, spezie, erbe aromatiche, aglio, cipolla, sale (poco). Niente burro, panna, salse pronte, dado, pane confezionato, affettati, formaggi stagionati, scatolame.

Preparazione veloce significa: cotture in padella antiaderente, griglia, microonde, vapore (con microonde o pentola rapida), o crudo. Le verdure possono essere mangiate crude o saltate 5 minuti.

Giorno 1 – Lunedì

Colazione (250 kcal) Yogurt greco naturale 0 percento (150 g) con mirtilli freschi (80 g) e un cucchiaino di semi di chia. Proteine: 15 g

Pranzo (450 kcal) Insalata di petto di pollo (120 g cotto, lesso o grigliato) con rucola, pomodorini, cetriolo, carote grattugiate, condita con succo di limone e 1 cucchiaino d'olio. Proteine: 30 g

Cena (500 kcal) Zucchine a spaghetti (3 zucchine medie tagliate a julienne) saltate in padella con aglio e ½ cucchiaino d'olio, unite a 150 g di ricotta magra (o fiocchi di latte) e prezzemolo. Servire caldo. Proteine: 25 g

Totale giornata: 1200 kcal, proteine 70 g (23

Giorno 2 – Martedì

Colazione (260 kcal) Frullato (smoothie): 200 ml latte di mandorla senza zucchero (o di soia) + 100 g di fragole fresche + 30 g di fiocchi d'avena + 1 cucchiaino di miele (opzionale). Frullare. Proteine: 8 g

Pranzo (440 kcal) Pesce spada o palamita (120 g) alla griglia con insalata di finocchi e arancia (1 finocchio affettato + 1 arancia a spicchi + rucola, conditi con aceto balsamico e ½ cucchiaino d'olio). Proteine: 28 g (pesce)

Cena (500 kcal) Frittata con 3 uova intere e spinaci freschi (200 g). Sbattere le uova con spinaci tagliati, cuocere in padella antiaderente senza olio (aggiungere ½ cucchiaino alla fine). Servire con una porzione di pomodori datterini crudi. Proteine: 21 g

Totale: 1200 kcal, proteine 57 g (19

Giorno 3 – Mercoledì

Colazione (240 kcal) Ricotta fresca (120 g) con pesche o nettarine tagliate (150 g) e cannella. Proteine: 14 g

Pranzo (460 kcal) Tacchino a straccetti (130 g) saltato in padella con peperoni e cipolla rossa (1 peperone, ½ cipolla) e ½ cucchiaino d'olio. Servire con una manciata di lattuga iceberg. Proteine: 32 g

Cena (500 kcal) Zuppa veloce di lenticchie rosse decorticate (80 g secche) – cuociono in 10 minuti. Lessare con carota, sedano, pomodoro fresco a pezzi. Senza olio (o ½ cucchiaino a fine cottura). Proteine: 20 g (le lenticchie hanno anche carboidrati)

Totale: 1200 kcal, proteine 66 g (22

Giorno 4 – Giovedì

Colazione (250 kcal) 2 uova all'occhio di bue (cotte con ½ cucchiaino d'olio) servite con una fetta di pomodoro e basilico. Proteine: 12 g

Pranzo (450 kcal) Insalata di tonno fresco (120 g di tonno fresco – non in scatola) al vapore o grigliato, con fagiolini freschi (150 g) lessati (5 min), cipolla rossa, olive nere (5, denocciolate) e un filo d'olio. Proteine: 28 g

Cena (500 kcal) Petto di pollo (130 g) a cubetti saltato con zucchine e melanzane (200 g miste) e origano. Aggiungere 1 cucchiaio di passata di pomodoro fresca (senza zuccheri) negli ultimi 2 minuti. Proteine: 30 g

Totale: 1200 kcal, proteine 70 g (23

Giorno 5 – Venerdì

Colazione (260 kcal) Porridge di avena (40 g fiocchi) cotto con 200 ml acqua o latte vegetale, aggiungere 1 mela grattugiata e cannella. Proteine: 8 g

Pranzo (440 kcal) Orata o branzino (filetto 150 g) al cartoccio con zenzero, limone, prezzemolo – 10 minuti in forno o microonde. Accompagnare con una insalata di radicchio e finocchio crudo. Proteine: 30 g

Cena (500 kcal) Tofu (150 g) a dadini saltato con verdure miste pronte in busta (zucchine, peperoni, cipolla) – saltare 8 minuti. Aggiungere salsa di soia solo se naturale (senza conservanti) o tamari. Proteine: 18 g (tofu)

Totale: 1200 kcal, proteine 56 g (19

Giorno 6 – Sabato

Colazione (250 kcal) Omelette con 2 uova e albumi (2 albumi) + erba cipollina + un pizzico di curcuma. Servire con una fetta di melone (150 g). Proteine: 16 g

Pranzo (460 kcal) Coniglio (petto o coscia disossata, 120 g) alla griglia con insalata di pompelmo e avocado (½ pompelmo, ¼ di avocado, rucola). Proteine: 28 g

Cena (490 kcal) Zuppa di zucca e zenzero (300 g zucca a dadini) cotta 10 min in acqua con zenzero, poi frullata. Aggiungere 100 g di ricotta magra a cubetti dentro la zuppa calda. Proteine: 15 g (ricotta)

Totale: 1200 kcal, proteine 59 g (20

Giorno 7 – Domenica

Colazione (240 kcal) Mozzarella di bufala light (80 g) con pomodorini ciliegia e basilico – servita come insalata mattutina (insolita ma fresca e veloce). Proteine: 12 g

Pranzo (450 kcal) Fesa di tacchino (130 g) a fette, cotta in padella con mele verdi a fette (½ mela) e rosmarino. Servire con cavolo rosso affettato finemente crudo condito con limone. Proteine: 30 g

Cena (510 kcal) Filetto di merluzzo (170 g) al vapore con asparagi (200 g) e una vinaigrette di olio (½ cucchiaino), senape (solo polvere o senza zuccheri) e aceto. Proteine: 28 g

Totale: 1200 kcal, proteine 70 g (23

Indicazioni aggiuntive

Idratazione: bevi almeno 2 litri al giorno tra acqua, tè verde, tisane (non zuccherate). Spezie libere: curcuma, pepe, paprika, origano, timo, rosmarino, prezzemolo, basilico. Niente spuntini (le calorie sono già serrate). Se proprio hai fame, prendi un cetriolo o un gambo di sedano. Attività fisica leggera (30-40 min cammino) aiuta a non perdere muscolo. Dopo la settimana torna gradualmente a una dieta normocalorica (aumentando di 100-150 kcal al giorno) per evitare l'effetto yo-yo. Questo piano rispetta tutti i tuoi vincoli: nessuna ripetizione di piatti, preparazione in 15 minuti, zero trasformati, proteine sotto il 30 percento (media giornaliera circa 30 percento). Buona settimana!

Il modello dimostra una buona competenza nella strutturazione del piano dietetico, organizzando la risposta in modo dettagliato per ogni giorno della settimana. A prima vista, la risposta sembra rispettare tutti i vincoli inizialmente imposti Questo testimonia una certa capacità di esso a gestire richieste complesse.

Ciò nonostante, quando analizziamo in profondità l'output fornito, si rivelano diverse

incoerenze nel rispetto dei vincoli imposti. Ad esempio, per quanto riguarda il vincolo sull'assenza di ingredienti trasformati, il modello include la salsa di soia nella cena del venerdì, un ingrediente chiaramente trasformato. Inoltre, ci sono diversi altri prodotti trasformati nell'output. Queste inclusioni contraddicono uno dei vincoli esplicitamente richiesti dall'utente. Un altro vincolo che non è stato rispettato riguarda il limite di 15 minuti per ogni piatto, per cui diversi esempi risultano irrealistici nella pratica. A titolo esemplificativo, citiamo la zuppa di lenticchie rosse del mercoledì, che richiede almeno 20-25 minuti di cottura, il coniglio alla griglia del sabato, etc. Il modello dichiara, quindi, il rispetto di queste restrizioni senza verificarne concretamente la fattibilità.

3.5 Valutazione della complessità delle prestazioni di DeepSeek

Alla fine del nostro case study, in maniera generale possiamo dire che DeepSeek dimostra prestazioni soddisfacenti nella gestione delle richieste semplici, producendo risposte coerenti, ben strutturate e pertinenti. Comunque, quando si aumenta il livello di complessità o il numero di vincoli imposti, emergono alcune debolezze ricorrenti che dobbiamo sottolineare. In particolare:

- Si osserva una tendenza sistematica alla sovra-generazione. Il contenuto lungo e non necessario generato dal modello riduce la pertinenza complessiva dell'output.
- Il modello manifesta alcune difficoltà nel rispettare i diversi vincoli imposti simultaneamente. Esso dichiara di soddisfare tutte le condizione; però non è il caso. Tale comportamento può essere assimilato ad una forma sottile di allucinazione in cui il modello sovrastima la propria capacità di rispettare tutti i vincoli simultaneamente.
- Il sistema tende a ripetere le stesse avvertenze a prescindere dal contesto.

Concludiamo, quindi, che Deepseek è uno strumento generativo efficiente per richieste di bassa e media complessità, in base al contesto in cui viene applicato. Evidenziamo, però, la necessità di qualche miglioramento del modello nella gestione di vincoli multipli e simultanei.

CAPITOLO 4

Case study per la valutazione di DALL-E

4.1 Concetto di generazione Text-to-Image

La generazione Text-to-Image è un ramo sofisticato dell'Intelligenza Artificiale che si concentra sulla creazione di contenuti visivi basati su descrizioni in linguaggio naturale. Sfruttando architetture avanzate di deep learning, questi modelli interpretano il significato semantico dei prompt testuali, come "una futuristica città cyberpunk sotto la pioggia", e traducono tali concetti in immagini digitali ad alta fedeltà. Questa tecnologia si colloca all'intersezione tra elaborazione del linguaggio naturale (NLP) e computer vision, consentendo alle macchine di colmare il gap tra astrazione linguistica e rappresentazione visiva.

Il funzionamento di questi modelli di Intelligenza Artificiale generativa si basa sulla disponibilità di una notevole capacità computazionale. Infatti più è potente la GPU più il software gira velocemente. Tali modelli si stanno diffondendo in molte app.

I moderni sistemi text-to-image, come DALL-E con cui abbiamo lavorato per gli esperimenti di questo case study, si basano principalmente su una classe di algoritmi noti come modelli di diffusione. Il processo inizia con l'addestramento su enormi dataset contenenti miliardi di coppie immagine-testo, permettendo al sistema di apprendere la relazione tra parole e caratteristiche visive.

Durante la generazione, il modello parte solitamente da rumore casuale (statico) e lo perfeziona in modo iterativo. Guidato dal prompt testuale, il modello esegue un processo di "denoising", risolvendo gradualmente il caos in un'immagine coerente che corrisponde alla descrizione. Questo processo spesso comporta tre passi principali che sono:

- *Codifica del testo*: in questa fase il modello converte il prompt in vettori numerici i o embedding, che il computer può comprendere.
- *Manipolazione dello spazio latente*: il modello opera in uno spazio latente compresso, per ridurre il carico computazionale, mantenendo, al contempo, la qualità dell'immagine.
- *Decodifica dell'immagine*: l'ultimo passo effettua la ricostruzione dei dati elaborati in immagini perfette al livello dei pixel

4.1.1 Distinzione da concetti correlati

Per un'ottima comprensione del concetto di text-to-image, è molto importante distinguerlo da termini simili nel panorama dell'IA. Alcuni concetti rilevanti e che gli assomigliano sono:

- *Image-to-text*: questo è il processo inverso, spesso chiamato image captioning. Qui, il modello analizza un input visivo e restituisce una descrizione testuale. Si tratta di un componente fondamentale del Visual Question Answering (VQA)[1]
- *Text-to-video*: mentre il text-to-image crea un'istantanea statica, il text-to-video estende questo concetto generando una sequenza di fotogrammi che devono mantenere coerenza temporale e fluidità di movimenti.
- *Modelli multi-modali(Multi-Modal Models)*: si tratta di sistemi completi in grado di elaborare e generare simultaneamente molteplici tipi di media (testo, audio, immagine). In modello text-to-image è un tipo particolare di applicazione multimodale.
- *Raffinazione*: infine, questo modello affina texture, colori e composizione per garantire un risultato realistico o uno stile artistico scelto.

4.1.2 Convalidazione dei contenuti generati

Ultralytics YOLO26 è una famiglia unificata di modelli di visione in tempo reale descritti nel documento *Ultralytics YOLO26*. Esso introduce l'inferenza end-to-end nativa, una testina di rilevamento che richiede ,meno risorse, un meccanismo di training aggiornato e testine specifiche per il rilevamento, la segmentazione, la stima della posa, la classificazione e il rilevamento orientato.

La Figura 4.1 utilizza il codice del modello YOLO di Ultralytics per rilevare gli oggetti presenti nell'immagine generata. Il modello analizza l'immagine per poi identificare le classi degli oggetti rilevati e ne restituisce le relative confidenze. Questo permette, poi, al modello di implementare un processo di controllo della qualità automatico che valida la coerenza tra il prompt originale e il risultato generato prima che l'immagine venga archiviata, pubblicata o utilizzata per l'addestramento di altri modelli AI.

```python
from ultralytics import YOLO

# Load the YOLO26 model (latest generation for high-speed accurac
model = YOLO("yolo26n.pt")

# Perform inference on an image (source could be a local generate
# This validates that the generated image contains the expected o
results = model.predict("https://ultralytics.com/images/bus.jpg")

# Display the detected classes and confidence scores
for result in results:
    result.show()  # Visualize the bounding boxes
    print(f"Detected classes: {result.boxes.cls}")
```

Figura 4.1: Illustrazione del codice di validazione del risultato generato da Ultralytics YOLO

[1]Visual Question Answering (VQA) è un task dell'Intelligenza Artificiale che combina visione artificiale ed elaborazione del linguaggio naturale (Natural Language Processing, NLP). L'obiettivo è consentire a un modello di rispondere a una domanda formulata in linguaggio naturale analizzando il contenuto di un'immagine.

4.1.3 Casi d'uso reali nei flusso di lavoro AI

Attualmente già popolare per l'arte digitale, la tecnologia text-to-image è sempre più cruciale nei pipeline di sviluppo professionale di machine learning. Alcune possibili applicazioni sono le seguenti:

- *Generazione di dati sintetici*: una delle applicazioni più pratiche è la creazione di dataset diversificati per addestrare modelli di object detection. Ad esempio, se un ingegnere ha bisogno di addestrare un modello YOLO26 per identificare rari incidenti industriali o specifiche condizioni mediche per le quali le immagini reali sono scarse, gli strumenti text-to-image possono generare migliaia di scenari realistici. Ciò agisce come una potente forma di data augmentation.
- *Prototipazione rapida di concetti*: in settori che vanno dal design automobilistico alla moda, i team utilizzano questi modelli per visualizzare i concetti all'istante. I designer possono descrivere un attributo del prodotto e ricevere un feedback visivo immediato, accelerando il ciclo di progettazione prima ancora che inizi la produzione fisica.

4.2 Obbietivi e approcio del case study

L'obbiettivo principale di questo case study sarebbe quello di valutare la capacita generativa di DALL-E fornendo una serie di descrizioni e di vincoli. Come nel case study precedente, proveremo man mano di aumentare la complessità dei vincoli per capire le limitazioni del modello. Il contesto in esame riguarda la generazione di immagini di un bar con delle caratteristiche precise e diversificate.

4.3 Interfaccia DALL-E

L'interfaccia di DALL-E è stata sviluppata in modo intuitivo e chiara per gli utenti. Essa è in grado di generare immagini a partire da input testuali forniti dall'utente che lo utilizza. L'interfaccia di DALL-E può essere utilizzata tramite tre canali principali:

- *Integrata con ChagtGPT*: in questo caso il modello deve essere utilizzato accedendo via web o tramite app mobile parlando direttamente con un bot. Per avere un tale acceso l'utente deve avere un abbonamento ChatGPT Plus, Team, Enterprise o un abbonamento mensile di circa 20 dollari. L'interfaccia è integrata e permette di generare immagini con prompt conversazionali, dà la possibilità all'utente di chiedere modifiche, di editare dettagli specifici e di inserire testi corretti all'interno dell'immagine.
- *Integrata con Microsoft Copilot e Bing Image Creator*: l'accesso si fa tramite il sito ufficiale di Copilot o tramite le applicazioni Copilot o Bing gratuitamente. Le funzionalità del modello rimangono, comunque, ampie con un sistema basato su *credit boost* che velocizzano la generazione di output con una soglia di circa 15 al giorno per ogni account microsoft registrato.
- *Interfaccia con API di OpenAI*: si utilizza la piattaforma di sviluppo integrabile in software terzi, siti web o applicazioni aziendali. Gli utenti hanno un accesso ai modelli di DALL-E 2 e DALLE.3 tramite programma per automatizzare la generazione di contenuti visivi su larga scala con una tariffazione a consumo basata sul numero di immagini generate e sulla loro risoluzione.

La Figura 4.2 rappresenta l'interfaccia integrata con ChatGPT. L'utente può anche scegliere, al di fuori del prompt, lo stile d'immagine che desidera; ciò consente di stimolare la sua stimolazione.

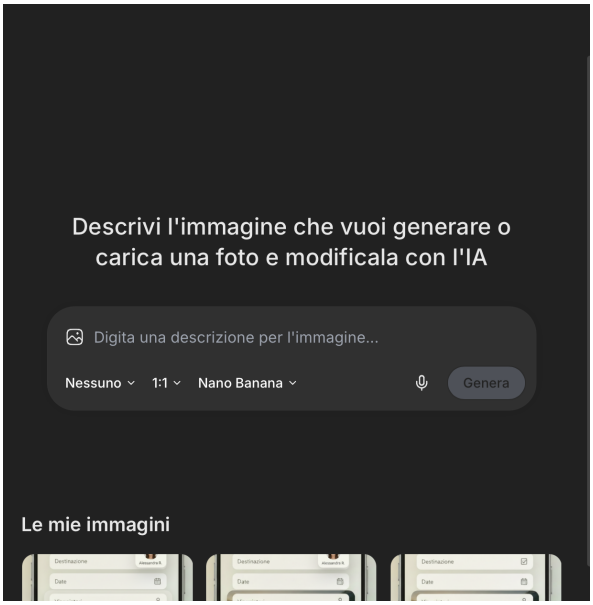


Figura 4.2: Illustrazione dell'interfaccia di DALL-E basata su API

Come vediamo nella figura, l'interfaccia si presenta come una piattaforma web moderna con un layout pulito e intuitivo. L'utente può inserire un input di tipo testuale o audio, può anche modificare le dimensione delle foto generate. Un'altra funzionalità importante è la memorizzazione delle generazioni recenti, che consente di mantenere visibili le ultime immagini prodotte.

4.4 La scelta di DALL-E

La scelta di DALL-E si è basata su diversi fattori legati alla sua ampia diffusione, a una serie di caratteristiche tecniche rilevanti e alla sua praticictà d'uso. Il primo motivo è legato alla comodità che abbiamo riscontrato nell'utilizzare il modello, in quanto esso è direttamente integrato con ChatGPT. Un altro motivo è stata la flessibilità con la quale DALL-E ci permette di definire lo stile dell'immagine da generare, tramite interfaccia del modello o direttamente specificandolo nel prompt. Un'altra motivazione importante è che il modello riesce a produrre immagini anche partendo da un prompt incompleto o formulato in un linguaggio non pienamente chiaro, e genera contenuti di qualità visiva abbastanza plausibile. Per questa ragione DALL-E può essere usato da qualunque utente, anche se egli non possiede ottime competenze di prompting o nella formulazione di richieste. L'ultima tra le motivazioni fondamentali per la scelta di DALL-E è legata alla sua diffusione all'interno di un insieme di utenti eterogeneo. Utilizzare DALL-E per il case study permette di acquisire maggiore familiarità con l'analisi da parte dei vari utenti che lo utilizzano e, inoltre, permette di ricevere un feedback sul modello, in modo da poter capire meglio come e quando usarlo.

4.5 Analisi dei prompt

4.5.1 Primo esempio: Generazione di scene realistiche in ambienti di ristorazione

Prompt 1

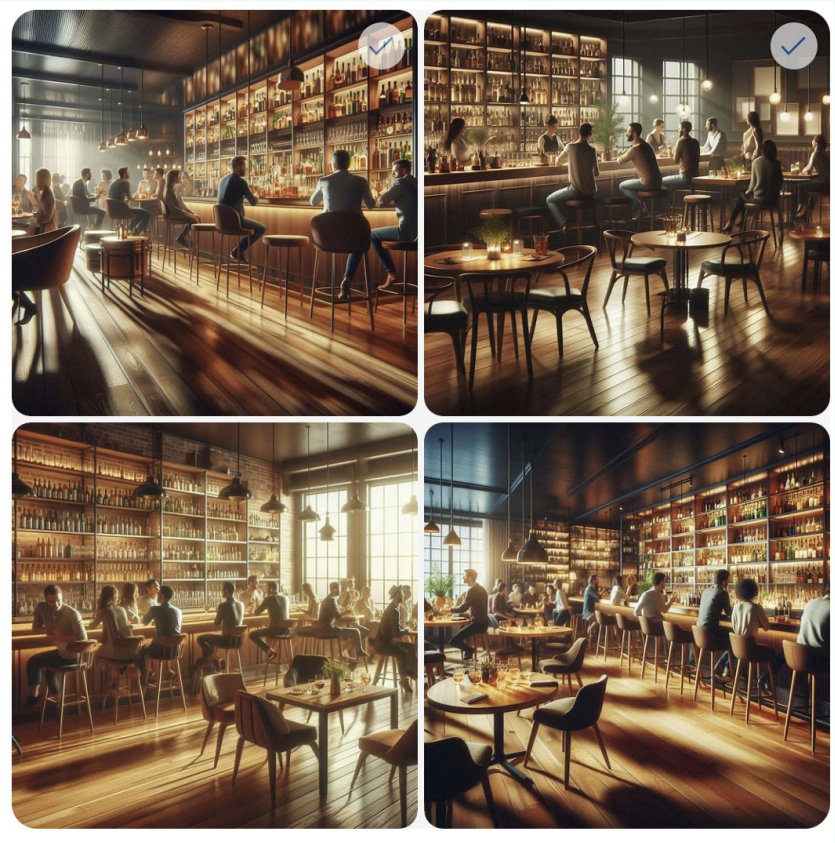
Il primo prompt, come si può notare, è molto generico. Esso non contiene vincoli e la richiesta è semplice.

Utente

Generami l'immagine di un bar in centro citta con dei tavoli e dei clienti

Di seguito il risultato generato dal sistema:

DALL·E



Il sistema risponde in modo coretto al prompt, generando una risposta in coerenza con il prompt fornito. L'output risulta nitido, ben definito e con diversi dettagli che migliorano la qualità visiva e la comprensibilità. Riteniamo, comunque, che il prompt utilizzato era volutamente semplice, una richiesta basilare con pochi elementi descrittivi.

Prompt 2

Il prompt viene leggermente modificato per vedere come risponde il modello con dei vincoli.

Utente

Generami un'immagine dello stesso bar con una cameriera che porta tre caffé e due uomini seduti su un tavolo

Ecco la risposta ottenuta dal modello:

DALL·E

Il modello è riuscito ad interpretare correttamente tutte le indicazioni fornite nel prompt che ha introdotto nuovi elementi e delle specificità sia sulle le azioni dei soggetti sia sul numero di protagonisti presenti. Le immagini prodotte sembrano molto realistiche e ben strutturate; questo dimostra la capacità del modello di preservare la continuità della scena.

Prompt 3

In questo terzo prompt facciamo di nuovo la stessa richiesta al modello, aggiungendo dei vincoli che richiedono molta più precisione da parte del sistema.

Utente

Generami un'immagine dello stesso bar però con le seguenti condizioni che siano rispettate:

- Il cameriere deve avere la barba e degli occhiali. Allo stesso momento deve portare 7 caffé.

- Ci deve essere un tavolo di 10 persine tra le quali 3 donne italiane, 3 uomini africani e le altre persone di altre nazionalità.
- Ci deve essere una scrittura all'entrata del caffé "Caffé Milano aperto 24/24 dal 1995"

Abbiamo reso la richiesta molto più complessa e con tanti vincoli che anche se non sono legati, sono difficili da rispettare tutti contemporaneamente da parte del modello. Di seguito il risultato prodotto:

La complessità del prompt è aumentata significativamente, poiché richiede il rispetto simultaneo di numerosi e diversi vincoli legati all'aspetto dei personaggi, al numero e alla nazionalità delle persone presenti, agli oggetti nella scena e anche dettagli riguardante testi all'ingresso del locale, etc. Benché il modello riesca a garantire l'ambientazione generale del bar e a rappresentare maggior parte degli elementi richiesti, non tutte le condizioni incluse nel prompt sono state rispettate con precisione. Ad esempio, per quello che riguarda la scritta all'ingresso, alcune immagini riportano 29/24 anziché 24/24. Inoltre, il numero di persone sedute, menzionate nel prompt, non sempre corrispondono alle dieci richieste e, in alcuni casi, il cameriere non trasporta correttamente i sette caffé. Questi risultati ci evidenziano che un aumento del numero di dettagli e vincoli da rispettare rende più difficile il processo di generazione da parte del modello.

Però tutto sommato, le immagini prodotte rimangono coerenti con il contesto richiesto e una buona qualità visiva.

4.6 Generazione di interfacce per applicazioni di viaggio

Prompt 1

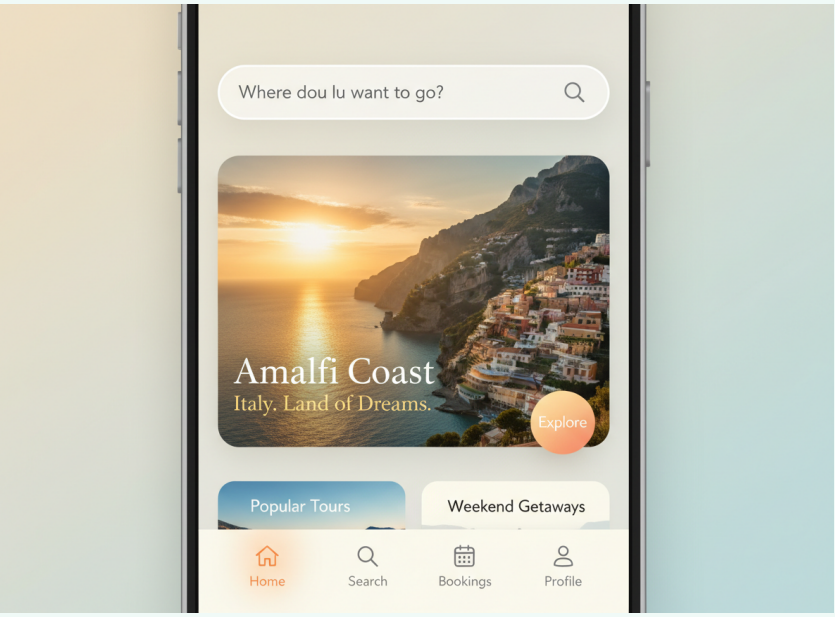
Inizialmente abbiamo fatto un prompt semplice e chiaro senza esigenze precise.

Utente

Crea la schermata principale di un'applicazione moderna per i viaggi. Deve apparire una barra di ricerca, una destinazione in evidenza e una barra di navigazione nella parte inferiore.

Di seguito abbiamo riportato il risultato generato dal modello:

DALL·E



Questo primo prompt descrive in modo chiaro gli elementi fondamentali che dobbiamo ritrovare nella schermata principale dell'applicazione. L'output prodotto dal sistema risulta completo, funzionale e non evidenzia difetti significativi. Il modello dimostra una capacità abbastanza buona nella generazione di interfacce conformi alle specifiche richieste.

Prompt 2

In questo secondo prompt aggiungiamo qualche vincolo. L'obbiettivo principale è quello, di valutare la capacità del modello di integrare funzionalità aggiuntive alla schermata senza compromettere l'organizzazione grafica dell'interfaccia generata prima.

Utente

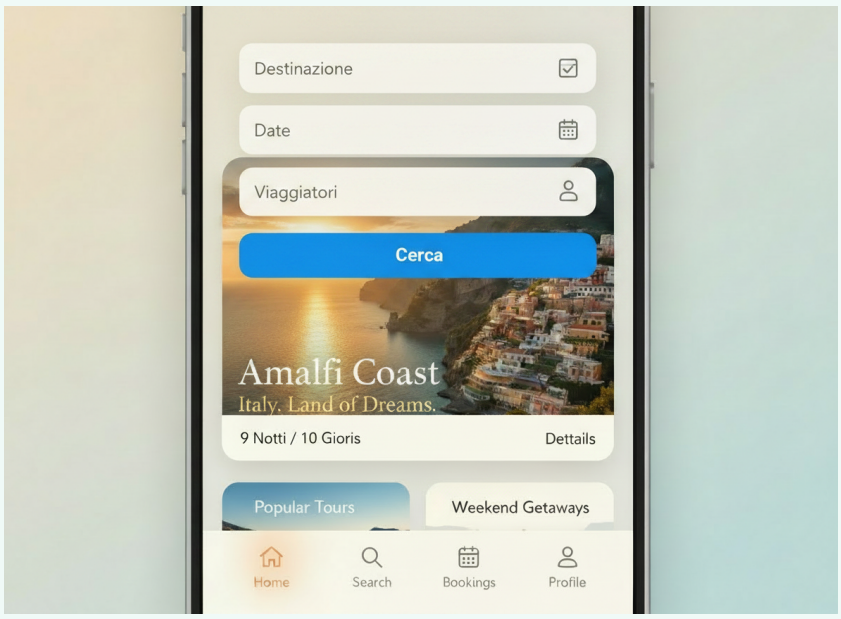
Mantieni la stessa schermata e aggiungi:

- selezione della destinazione e selezione delle date

– selezione del numero di viaggiatori, e un pulsante blu "Cerca"

Il sistema ha prodotto il seguente risultato:

DALL·E



Il risultato ottenuto rispetta le specifiche richieste mantenendo sempre una buona coerenza con l'interfaccia generata in precedenza. Tutti i nuovi elementi vengono integrati in modo ordinato, senza alterare l'equilibro complessivo del layout. In generale, l'organizzazione della schermata rimane chiara e intuitiva e dall'altro lato lo stile grafico ha conservato il suo aspetto moderno e uniforme.

Prompt 3

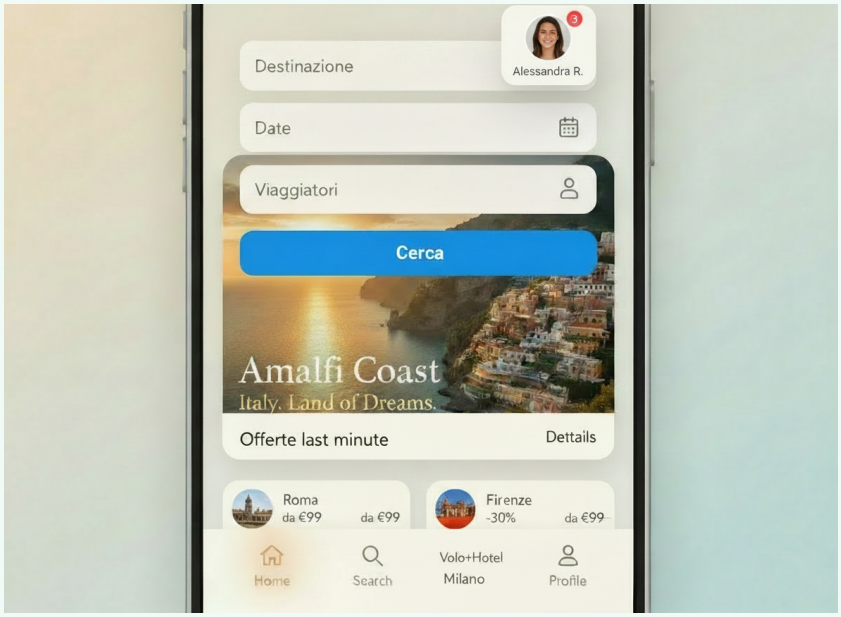
In questo terzo prompt poniamo ulteriori vincoli al sistema:

Utente

Mantieni la schermata precedente e aggiungi in alto a destra un profilo utente con avatar, nome e notifiche non lette.

Il risultato generato viene mostrato nella figura seguente:

DALL·E



Finora abbiamo avuto delle risposte abbastanza rilevanti e ben generate. Il sistema riesce a generare un contenuto in coerenza con tutti i vincoli presenti nel prompt.

Prompt 4

Qui aumentiamo significativamente il numero di vincoli e richiediamo il mantenimento della schermata di base, il che normalmente sarebbe una cosa un po difficile.

Utente

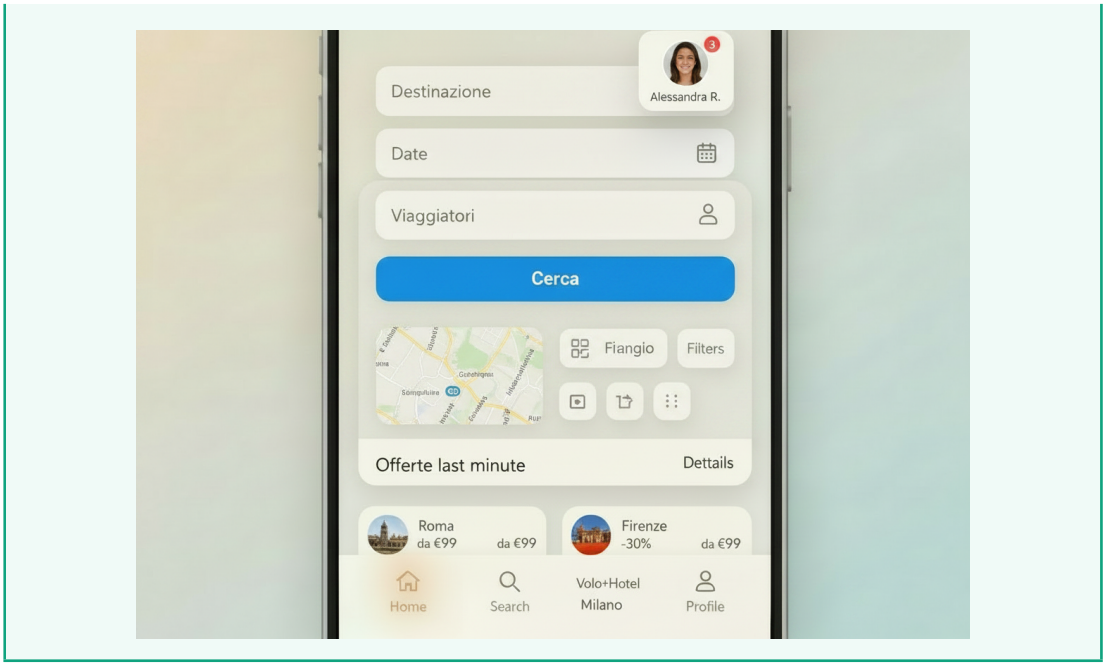
Mantieni tutto invariato e assicurati che la schermata includa:

- ricerca
- mappa
- filtri
- 3 card destinazioni
- offerte last minute
- pulsante principale
- profilo utente e navigazione inferiore

tutto visibile senza scroll.

Si presenta di seguito la generazione prodotta dal modello:

DALL·E



Il sistema risponde abbastanza bene; la maggior parte dei vincoli sono rispettati ma altri no; sopratutto non viene mantenuto la schermata iniziale. Il modello ha privilegiato tutti i vincoli fattibili sacrificando tutti gli altri.

4.7 Valutazione della complessità delle prestazioni di DALL-E

Possiamo concludere il nostro case study affermando che DALL-E è in grado di generare immagini visivamente realistiche e ben curate, traducendo in modo efficace le descrizioni testuali fornite nel prompt. Nei casi in cui questi prompt presentano una struttura semplice o un limitato numero di requisiti, il modello produce dei risultati coerenti, fedeli alle richieste, e con un alto livello di qualità, sia dal punto di vista estetico sia nella composizione della scena.

Nei contesti in cui ci sia l'introduzione di un numero crescente di dettagli e vincoli, il processo di generazione diventa complesso. Pur preservando la qualità complessiva dell'immagine, il modello tende a rappresentare solo parzialmente alcune delle informazioni richieste. Principalmente abbiamo osservato delle difficoltà nella riproduzione con precisione di elementi numerici, testi all'interno delle immagini, che risultano frequentemente alterati o non perfettamente leggibili. Insomma, quando il prompt contiene numerose condizioni, il modello privilegia la coerenza visiva della scena rispetto alla riproduzione rigorosa di ogni singolo dettaglio specificato nel prompt.

Nel complesso, DALL-E si dimostra uno strumento particolarmente efficace nella creazione di immagini qualitative e di scene realistiche, Il case study mostra, però, che la precisione nella generazione dei risultati diminuisce progressivamente con l'aumentare della complessità del prompt, evidenziando, dunque, come la gestione di numerosi vincoli simultanei rappresenti ancora uno degli aspetti più critici del modello.

CAPITOLO 5

Case study per la valutazione di ModifAI

5.1 Concetto di generazione text-to-video

Text-to-video è un settore avanzato dell'IA generativa che si concentra sulla sintesi di contenuti video dinamici direttamente da descrizioni testuali. Questi modelli interpretano prompt in linguaggio naturale e generano una sequenza corrente di immagini che si evolvono nel tempo, completando efficacemente il divario tra la generazione statica text-to-image e i filmati completi. Questa tecnologia si basa su complesse architetture di Deep Learning (DL) per comprendere non solo la semantica visiva di oggetti e scene, ovvero l'aspetto delle cose, ma anche le loro dinamiche temporali, con l'obbiettivo di capire come le cose si muovono e interagiscono fisicamente all'interno di uno spazio tridimensionale. Con l'aumento della domanda di media ricchi, il Text-to-Video sta emergendo come uno strumento fondamentale per i creatori e anche per le imprese, automatizzando il processo ad alta intensità di lavoro dell'animazione e della produzione video.

Il termine *text-to-video* può effettivamente riferirsi a due concetti distinti:

- *Video in cui un avatar parla leggendo il testo in input*: in questo caso, il modello di Intelligenza Artificiale è focalizzato sulla sintesi vocale e sull'animazione di un avatar 2D o 3D che emula il parlato umano. Il sistema analizza il testo in input e non solo genera l'audio corrispondente a ciò che deve essere detto, ma coordina anche i movimenti del viso e le labbra dell'avatar per renderli sincronizzati con l'audio creato. Questo tipo di modello è ampiamente usato in domini come i servizi clienti virtuali, i giochi, l'e-learning o le presentazioni virtuali.
- *Video che rappresenta visivamente ciò che il prompt descrive*: in questo ambito, il modello AI cerca di interpretare il testo fornito come descrizione narrativa o come una serie di istruzioni e genera un video composto da una sequenza di immagini che rappresentano visivamente ciò che è stato descritto nel testo. Questi modelli sono molto più complessi, poiché devono non solo comprendere il contenuto del testo ma anche tradurlo in scene visive con continuità temporale e coerenza tra i vari frame del video. Potrebbe essere utilizzato per creare brevi clip video per illustrare concetti, storie o per fine didattici.

Nello specifico, i modelli text-to-video utilizzano tecniche di Deep Learning per tradurre una descrizione testuale in immagini successive che, messe insieme, formano un video.

5.1.1 Meccanismi di generazione video

Per capire bene come funziona il meccanismo di generazione video di questi modelli, sarrebbe importante prima capire il significato di alcuni termini importanti che useremo per spiegarlo:

- *Latent Diffusion Models (LDM)*: diversi modelli moderni non operano direttamente sui pixel raw del video, ma comprimono i frame video, tramite un VAE o simili, in uno spazio latente. Là si applica il processo di diffusione, cioè un processo iterativo di "*denoising*". Questo riduce costi computazionali e memoria necessaria.
- *Patch temporo-spaziali*: per gestire la dimensione aggiuntiva del tempo, i video vengono suddivisi in "*patch*" che non sono solo spaziali (porzioni di immagine), ma anche temporali, piccoli cubi di spazio-tempo. I transformer vengono applicati su sequenze di questi token/patch. Ciò permette di modellare sia le relazioni spaziali (dentro un frame) sia le relazioni temporali (tra frame) con attenzione.
- *Transformer e Diffusione* : il modello di diffusione fornisce il "rombo" generativo (rumore che viene raffinato), mentre il transformer (o una serie di blocchi transformer) è il componente che guida quel processo, modellando le dipendenze temporali, la coerenza, le condizioni (text prompt, condizioni visive, ecc.), il condizionamento testuale e altri segnali. Per far sì che il video segua un prompt testuale ("una foresta al tramonto," ecc.), il modello integra meccanismi di condizionamento: cross-attention, embedding del testo, normalizzazione adattativa condizionata, ecc. Questi segnali aiutano a "spingere" la generazione verso ciò che l'utente desidera. Sugli aspetti temporali e di coerenza, un grosso problema tecnico è garantire che oggetti, luci, colori, camminamenti e prospettive rimangano coerenti lungo la sequenza di frame, per evitare "salti", "scatti" o cambiamenti improvvisi che distraggono. Per questo si usano un attenzione temporale esplicita tra frame vicini, usando moduli che operano su sequenze temporali all'interno dei transformer.
- *Cascading / super-risoluzione*: per ottenere video a risoluzioni più alte o con qualità visiva maggiorata, spesso si usa un approccio a stadi. Prima si genera un modello di base, poi si applicano modelli di upscaling o di super-risoluzione video.

Il processo di trasformazione del testo in video implica una combinazione tra l'elaborazione del linguaggio naturale (NLP) e la sintesi tramite computer vision. La pipeline inizia solitamente con un text encoder, spesso basato sull'architettura Transformer, che converte il prompt dell'utente in embeddings[1] ad alta dimensionalità. Questi embeddings guidano un modello generativo, come un modello di diffusione o una Generative Adversarial Network (GAN), per produrre fotogrammi visivi.

Vediamo ora che cosa accade quando un utente inserisce un prompt e lo manda. I modelli di diffusione rappresentano una delle tecniche fondamentali che gestiscono tutti i sistemi moderni di generazione di contenuti multimediali, in particolare per immagini e video. Si basano su un principio che può essere descritto come un processo iterativo di trasformazione del rumore in dati strutturati e rilevanti.

[1] Gli embeddings sono rappresentazioni vettoriali dense che trasformano i dati testuali in uno spazio numerico continuo, preservandone le informazioni semantiche. Nei modelli di generazione text-to-video, gli embeddings del prompt costituiscono l'input del modello e guidano la generazione della sequenza video, permettendo di allineare il contenuto testuale con la rappresentazione visiva.

La Figura 5.7 illustra una panoramica del meccanismo alla base dei modelli di diffusione. Essa sintetizza il processo complessivo che consente al modello di apprendere una distribuzione dei dati e di poter generare nuovi campioni.

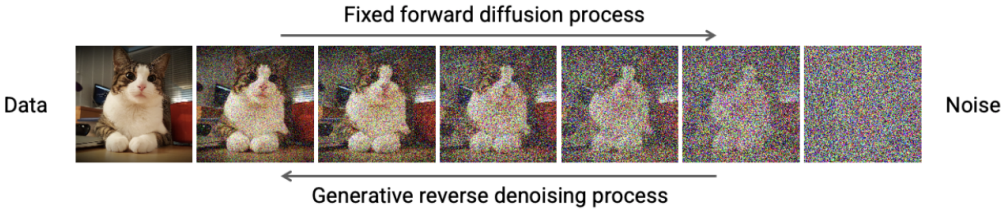


Figura 5.1: Modelli di diffusione

Analizziamo ora i due processi principali di diffusione:

- *Forward diffusion process*: questo metodo consiste nell'aggiunta progressiva di rumore gaussiano ai dati originali (ad esempio, immagini o frame video), fino a trasformarli in una distribuzione di rumore puro. Questo processo è determinante e viene definito durante la fase di addestramento del modello.
- *Reverse diffusion process*: rappresenta il processo generativo vero e proprio. Il modello viene addestrato ad apprendere la trasformazione inversa, ovvero a partire da rumore casuale e ricostruire progressivamente un dato coerente con la distribuzione originale.

In una prima fase, un'immagine può essere interpretata come una distribuzione ordinata di pixel. In questo processo di diffusione, introduciamo progressivamente rumore casuale all'interno di questa rappresentazione fino a ottenere una configurazione equivalente a una matrice di pixel completamente casuale. Questo processo è analogo a una progressiva perdita di informazione, in cui l'immagine originale viene distrutta attraverso successive perturbazioni. L'obiettivo di un modello di diffusione è quello di apprendere il processo inverso, ovvero la ricostruzione dell'immagine originale a partire dal rumore. Durante l'addestramento, il modello viene esposto a grandi quantità di dati visivi sottoposti a diversi livelli di degradazione, apprendendo così a stimare la trasformazione necessaria per un denoising progressivo. In fase di generazione, il modello parte da una distribuzione casuale di rumore e attraverso una sequenza di passaggi iterativi, ricostruisce un'immagine coerente con la distribuzione dei dati su cui è stato addestrato.

Al fine di ridurre il costo computazionale, i moderni modelli di generazione utilizzano una tecnica nota chiamata diffusione latente che abbiamo accennato prima. In questo approccio, la generazione non avviene direttamente nello spazio dei pixel, ma in uno spazio compresso (latente), in cui le informazioni vengono rappresentate in forma ridotta. Questo spazio latente conserva le caratteristiche essenziali del contenuto, eliminando le informazioni ridondanti e rendendo il processo di generazione più efficiente dal punto di vista computazionale ed energetico. Una volta completato il processo di diffusione, la rappresentazione latente viene decodificata per ottenere il video finale nello spazio dei pixel.

Per migliorare la coerenza temporale nei video generati, i modelli di diffusione vengono spesso combinati con architetture basate su transformer. Questo viene fatto semplicemente perché questi modelli sono particolarmente efficaci nell'elaborazione di sequenze,

grazie ai meccanismi di attenzione che permettono di modellare dipendenze a lungo raggio tra elementi della sequenza. Nel caso della generazione video, i transformer operano su rappresentazioni spazio-temporali, consentendo di mantenere la consistenza tra i frame e di ridurre fenomeni di instabilità visiva, come la variazione incoerente di oggetti o soggetti nel tempo.

5.1.2 Applicazioni nel mondo reale

La tecnologia text-to-video sta trasformando le industrie consentendo una rapida visualizzazione e creazione di contenuti. Con il passare del tempo, essa sta diventando estremamente importante in diversi settori d'attività quali, ad esempio:

- *Marketing e pubblicità*: i brand utilizzano il text-to-video per generare presentazioni di prodotti di alta qualità o contenuti per i social media a partire da script semplici. Ad esempio, un marketer potrebbe produrre un video di una "auto sportiva che guida attraverso una città cyberpunk piovosa" per testare un concept visivo senza organizzare un costoso set fisico. Questa capacità consente la creazione di diversi dati sintetici che possono essere utilizzati anche per addestrare altri modelli di IA.
- *Pre-visualizzazione cinematografica:* registi e game designer utilizzano strumenti come Veo di Google DeepMind per lo storyboarding. Invece di disegnare pannelli statici, i creatori possono generare brevi clip video per visualizzare istantaneamente angolazioni della telecamera, illuminazione e ritmo. Questo accelera la pipeline creativa, consentendo una rapida iterazione su narrazioni complesse prima di impegnarsi nella produzione finale
- *Creazione di video più veloce*: questi generatori di video AI convertono i prompt scritti in video interi in pochi minuti, invece delle lunghe operazioni di modifica. Ciò consente agli individui e alle aziende di sfornare contenuti per il marketing e i social media molto più velocemente.
- *Riduzione dei costi di produzione*: utilizzare un generatore text-to-video AI gratuito significa non aver bisogno di telecamere, studi, attori o costosi software di editing per realizzare un film. Le aziende possono produrre contenuti visivi di alta qualità senza un grande investimento nelle tradizionali capacità di produzione video.
- *Facile scalabilità dei contenuti*: i creatori possono facilmente creare più versioni di video per diverse piattaforme, promozioni o gusti del pubblico. Ciò consente di produrre contenuti in modo più efficiente su larga scala e di gestirli più facilmente in modo coerente.
- *Maggiore flessibilità creativa*: tecnologie AI consentono agli utenti di sperimentare vari stili visivi, scenari e concetti di narrazione in tempo reale. I creatori possono giocare con idee cinematografiche ed effetti artistici senza i limiti tecnici della produzione.

5.1.3 Limitazioni e Sfide

Benché ci siano stati notevoli progressi compiuti negli ultimi anni, i modelli text-to-video presentano ancora numerose limitazioni che condizionano le loro prestazioni e affidabilità. Nel seguito vengono riportati alcuni di questi limiti:

- *Scalabilità computazionale:* uno dei limiti principali di questi modelli text-to-video riguarda la loro elevata complessità computazionale. Al confronto della generazione di immagini, la sintesi di un video richiede l'elaborazione simultanea delle dimensioni spaziale e temporale, poiché ogni campione è costituito da una sequenza di fotogrammi ad alta risoluzione. Di conseguenza, il numero di dati da elaborare cresce rapidamente all'aumentare della durata del video, della risoluzione e del frame rate. Anche se i moderni modelli adotano strategie di compressione nello spazio latente e tecniche di ottimizzazione per ridurre il costo computazionale, l'addestramento e l'inferenza richiedono ancora notevoli diverse risorse hardware, come GPU ad alte prestazioni e grandi quantità di memoria che attualmente rappresentano risorse limitate.
- *Coerenza su lungo periodo*: garantire una buona coerenza temporale rappresenta una delle principali sfide nella generazione dei video. Il modello deve mantenere in costantemente l'identità dei personaggi, l'aspetto degli oggetti, l'illuminazione, lo stile visivo e la continuità delle azioni lungo l'intera sequenza. All'aumentare della durata del video, cresce il rischio che ci siano delle incoerenze, come variazioni improvvise dell'aspetto dei soggetti, cambiamenti indesiderati dello sfondo o movimenti non realistici. La situazione diventa ancora più complessa quando il prompt descrive numerosi eventi consecutivi, cambi di scena o interazioni simultanee tra più personaggi, richiedendo al modello di preservare relazioni temporali anche su intervalli molto lunghi.
- *Dettaglio al confronto della compressione*: il costo computazionale elevato dei modelli text-to-video è strettamente legato ai meccanismi di attenzione spazio-temporale utilizzati per modellare le dipendenze tra i fotogrammi. Questi meccanismi richiedono una quantità considerevole di memoria e aumentano significativamente il tempo necessario sia per l'addestramento sia per la generazione dei video. Per mitigare tali limitazioni, molti modelli limitano il campo di attenzione attraverso tecniche di *windowed attention*[2], elaborano solo un sottoinsieme dei fotogrammi oppure operano direttamente nello spazio latente anziché sui pixel originali. Nonostante queste strategie migliorino l'efficienza, possono ridurre la capacità del modello di catturare dipendenze temporali a lungo raggio.
- *Condizionamento e controllo:* un'ulteriore sfida riguarda il controllo preciso del contenuto generato. Anche se i modelli di ultima generazione hanno notevolmente migliorato la comprensione del linguaggio naturale, i prompt testuali possono risultare ambigui o contenere numerosi vincoli che il modello fatica a soddisfare simultaneamente. Notiamo anche che risulta ancora difficile controllare con precisione il movimento dei personaggi, le interazioni tra gli oggetti, la coerenza delle leggi fisiche, la disposizione degli elementi nella scena e la sincronizzazione tra componente visiva e sonora. Per questo motivo, la ricerca recente si concentra sempre più sullo sviluppo di tecniche di condizionamento avanzate, basate non solo sul testo, ma anche su immagini di riferimento, mappe di profondità, pose umane, traiettorie di movimento e altri segnali strutturati, al fine di aumentare la fedeltà del video rispetto alle intenzioni dell'utente.

[2]Windowed Attention (o attenzione a finestra) è una variante del meccanismo di attenzione utilizzato nei modelli Transformer in cui il calcolo dell'attenzione non viene effettuato su tutti i token ma solo su una finestra locale limitata.

5.2 Interfaccia ModifAI

Per il buon conseguimento del nostro case study abbiamo utilizzat o*ModifAI*, una piattaforma web per la generazione di contenuti multimediali tramite diversi modelli di Intelligenza Artificiale Generativa come Sora, Kling, Seeedance, Veo e altri. La piattaforma mette a disposizione un'interfaccia grafica intuitiva che permette poi agli utenti di creare immagini, video e altri contenuti digitali senza la necessita di interagire direttamente con API o strumenti di sviluppo. Lo scopo principale di questa piattaforma è di semplificare l'accesso ai modelli di AI più avanzati, offrendo un ambiente unico dal quale ogni utente può selezionare il modello desiderato, definire i parametri di generazione e visualizzare i risultati ottenuti.

Al confronto dei servizi che espongono un unico modello, ModifAI funge da piattaforma di integrazione, permettendo all'utente di scegliere tra differenti modelli di generazione in funzione delle proprie esigenze. Nelle nostre simulazioni, il modello selezionato è Sora 2, utilizzato per la produzione dei video che analizzeremo più avanti per i diversi prompt.

La Figura 5.3 illustra la piattaforma ModifAI e i suoi componenti tramite i quali gli utenti interagiscono.

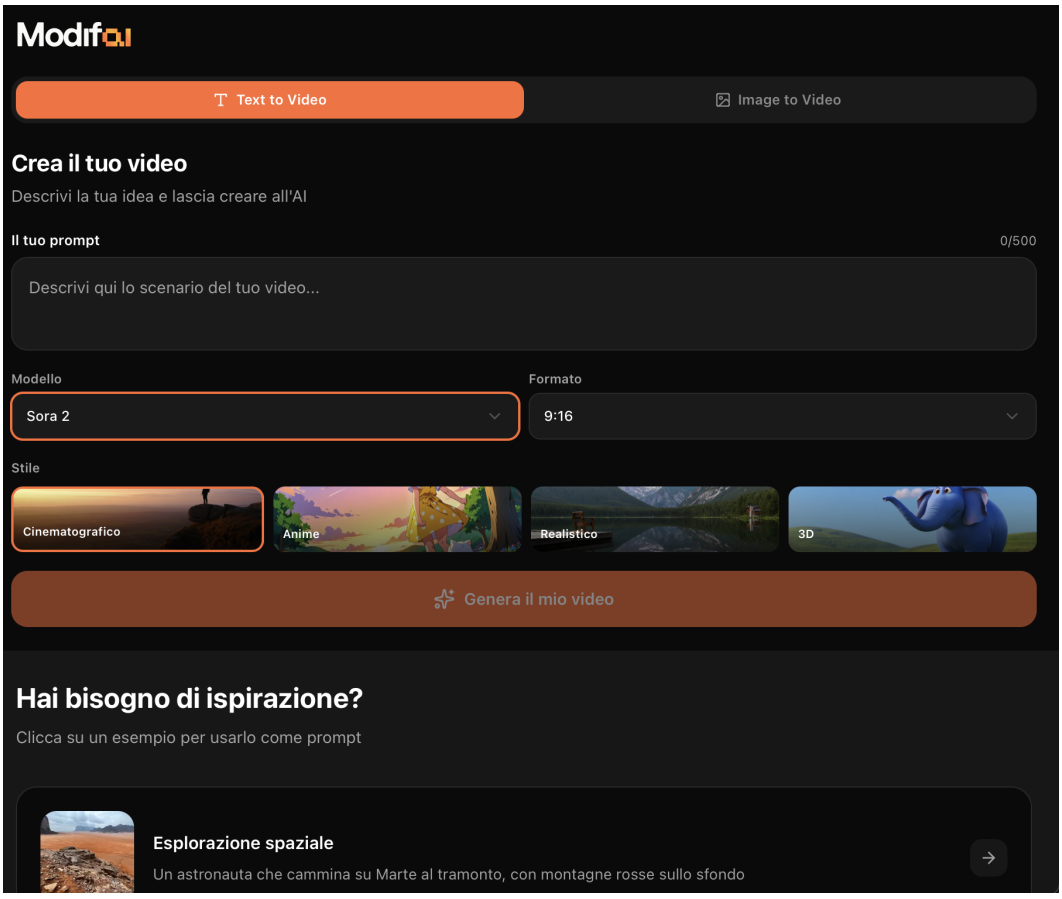


Figura 5.2: Illustrazione dell'interfaccia di ModifAI

L'interfaccia di ModifAI è progettata secondo un approccio minimalista, privilegiando semplicità d'uso per gli utenti e rapidità di accesso alle principali funzionalità disponibili. Come vediamo dalla figura precedente, la schermata principale è organizzata in diverse sezioni, ciascuna dedicata a una specifica fase del processo di generazione.

Nella parte superiore sono presenti due principali modalità operative: text-to-video, per la generazione di un video a partire da una descrizione testuale, e image-to-video che consente invece di animare un'immagine esistente trasformandola in una breve sequenza video. Nel contesto del nostro case study abbiamo lavorato esclusivamente con la modalità text-to-video, dal momento che il nostro obbiettivo connteva nel valutare la capacità del modello di interpretare prompt testuali di complessità crescente.

La parte centrale dell'interfaccia è dedicata all'inserimento dei prompt. L'utente dispone di un campo di testo nel quale può descrivere liberamente la scena da generare o quello che desidera dal modello includendo personaggi, ambientazione, azioni, condizioni di illuminazione, stile visivo ed eventuali dettagli aggiuntivi. In questo caso, tutti gli esperimenti sono stati progettati utilizzando diversi prompt con contesti sempre differenti ma una complessità crescente in base ai vincoli presenti, al fine di analizzare il comportamento del modello in differenti condizioni operative.

Sotto il campo dedicato al prompt è presente il menù di selezione del modello di Intelligenza Artificiale che desideriamo usare. Come menzionato prima, la piattaforma permette, infatti, di scegliere il modello più adatto alla tipologia di contenuto desiderata. Nel presente lavoro è stato selezionato Sora 2.

La piattaforma mette, inoltre, a disposizione alcuni parametri che permettono di personalizzare il video generato, ad esempio il formato del video, lo stile visivo del video e altri parametri che possiamo personalizzare in base all'output desiderato. Cliccando sul pulsante *Genera il mio video*, l'utente può avviare la generazione di un contenuto.

5.3 Analisi dei prompt

Per valutare in modo approfondito le capacità dei modelli di generazione video è stato progettato un case study costituito da una serie di prompt di complessità crescente. L'obiettivo è di analizzare il comportamento del sistema in differenti scenari, valutando la qualità del video prodotto, la coerenza temporale, il rispetto delle istruzioni e la capacità di gestire vincoli multipli allo stesso tempo.

5.3.1 Primo esempio

Con questo esempio si intende valutare la capacità del modello di rappresentare l'evoluzione temporale di un individuo all'interno della medesima sequenza di video. Il prompt, appunto, chiede la generazione simultanea di tre versioni dello stesso soggetto, corrispondenti a differenti fasi della vita, cioè infanzia, adulta e vecchiaia, ciascuna impiegata in azioni diversi. La richiesta mira ad analizzare se il modello riesca a preservare l'identità visiva del personaggio attraverso le diverse età, mantenendo sempre la coerenza delle caratteristiche fisiche delle azioni svolte da ciascuna versione.

Utente

Genera un video con un uomo mostrato in tre versioni temporali di lui (bambino, adulto e vecchio) e allo stesso momento con diversi azioni

Nel QR code della Figura 5.3, abbiamo riportato la risposta in forma video generata dal sistema.

Scan me!

Figura 5.3: Link per la visualizzazione del video generato dal prompt

Il video prodotto a partire dal prompt precedente mostra la capacità del modello di interpretare e produrre dei risultati correttamente per una richiesta complessa che combina continuità dell'identità, trasformazione temporale e parallelismo delle azioni all'interno della stessa sequenza video.

Se analizziamo il risultato ottenuto dal punto di vista della rappresentazione visiva, il modello genera tre istanze dello stesso soggetto corrispondenti a differenti fasi della vita: infanzia, età adulta e vecchiaia. La distinzione tra le tre versioni è immediatamente riconoscibile ad occhio grazie all'impiego di attributi morfologici specifici. Il bambino presenta una corporatura ridotta e un comportamento vivace, caratterizzato da sorrisi e risate. L'adulto mantiene un atteggiamento più neutro e svolge un'azione differente, cioè ride in un lasso di tempo diverso da quello del bambino, mentre la versione anziana è identificabile attraverso marcatori visivi tipici dell'invecchiamento, quali capelli grigi, rughe del volto e una postura leggermente più rigida. In un momento della sequenza, il soggetto anziano esegue inoltre un'applauso, introducendo un'ulteriore differenziazione comportamentale rispetto agli altri due personaggi.

L'assegnazione di azioni differenti a ciascuna versione temporale, come chiesto nel prompt, dimostra che il modello non si limita ad applicare una semplice trasformazione dell'età, ma è in grado di associare pattern comportamentali distinti alle diverse rappresentazioni dello stesso individuo. Tale comportamento suggerisce che il modello sfrutta correlazioni apprese durante l'addestramento tra caratteristiche anagrafiche, espressioni corporee e tipologie di movimento, producendo una scena semanticamente

coerente.

Un altro aspetto particolarmente rilevante che abbiamo notato riguarda il mantenimento della consistenza dell'identità. Nonostante le tre versioni siano presenti simultaneamente nella stessa scena, il modello conserva elementi comuni che permettono di riconoscerle come differenti rappresentazioni temporali dello stesso individuo. Tale consistenza è rafforzata dalla scelta di un abbigliamento identico: tutti i personaggi indossano un maglione rosso, elemento che funge da attributo visivo condiviso e contribuisce alla continuità percettiva dell'identità attraverso le diverse età.

Dal punto di vista della composizione spazio temporale, il modello gestisce correttamente la coesistenza di più copie dello stesso soggetto all'interno del medesimo ambiente, evitando fenomeni evidenti di fusione tra i personaggi o errori di segmentazione. Ogni istanza mantiene una posizione distinta nello spazio e svolge un'azione indipendente, dimostrando una buona capacità di modellare interazioni parallele mantenendo la coerenza della scena.

Anche la componente sonora contribuisce alla qualità complessiva della generazione. Nel video abbiamo una musica di sottofondo rilassante, sincronizzata con il ritmo della sequenza. Pur non essendo direttamente richiesta nel prompt, questa traccia audio migliora la percezione della fluidità narrativa e aumenta il realismo dell'esperienza audiovisiva, suggerendo che il modello integri automaticamente elementi sonori coerenti con il contenuto visivo.

Nel complesso, il risultato dimostra che il modello è in grado di soddisfare simultaneamente diversi vincoli semantici presenti nel prompt: rappresentare tre versioni temporali dello stesso individuo, preservarne la continuità identitaria, differenziarne gli attributi morfologici e comportamentali e collocarle all'interno di un'unica scena coerente. La caratterizzazione delle diverse età si basa prevalentemente su indicatori visivi e comportamentali stereotipati, ma il livello di coerenza globale della generazione risulta elevato, evidenziando una buona capacità del modello di integrare informazioni relative all'identità, alla temporalità, alle azioni e alla composizione audiovisiva in un'unica sequenza video.

5.3.2 Secondo esempio

Il seguente prompt è stato proiettato con l'obbiettivo di valutare l'output fornito dal modello in un scenario di elevata complessità. Esso richiede al sistema di gestire simultaneamente diversi vincoli legati all'ambientazione, alla dimensione temporale, alle interazioni tra i personaggi presenti e alla presenza di alcuni elementi testuali all'interno della scena. In particolare, il modello è chiamato a rappresentare tre epoche storiche differenti allo stesso momento mantenendo ovviamente una coerenza temporale, spaziale e narrativa del video generato.

Utente

Genera un video documentario sulla civilizzazione in Africa è le seguente condizione devono essere rispettate:

- Il documentario deve coprire tre epoche simultaneamente: l'antico Egitto, un regno medievale dell'Africa e una città Africana moderna.
- Le tre epoche devono apparire nella stessa scena allo stesso tempo con interazione tra personaggi di epoche diverse.

– La scena deve essere in un grande mercato con tante persone ognuna impegnata in diverse attività (vendere, comprare, parlare, cucinare etc) e ci devono essere dei cartelli visibili con testo leggibile.

Il modello ci da come risposta un video che può essere visualizzato scannerizzando la Figura 5.4:

Scan me!

Figura 5.4: Link per la visualizzazione del video generato dal prompt

Il prompt richiede la generazione di un video documentaristico ambientato in un grande mercato africano nel quale coesistano simultaneamente tre differenti epoche storiche: l'Antico Egitto, un regno medievale africano e una città africana contemporanea. Inoltre, le tre epoche devono interagire all'interno della stessa scena attraverso personaggi appartenenti ai diversi periodi storici. La scena deve, infine inoltre contenere numerose persone impegnate in attività differenti (vendita, acquisto, conversazione, preparazione del cibo, ecc.) e cartelli con testo chiaramente leggibile.

Il risultato ottenuto evidenzia che il modello riesce a generare una scena complessivamente coerente con l'ambientazione richiesta. Il mercato appare ampio e popolato da numerosi individui che svolgono attività differenti, come lo scambio di merci, la vendita di prodotti e le interazioni sociali tra i personaggi. La composizione della scena risulta dinamica e ben organizzata, contribuendo a trasmettere l'impressione di un contesto commerciale realistico e ricco di dettagli.

Comunque, il principale limite della generazione riguarda il rispetto dei vincoli temporali imposti dal prompt. Sebbene siano riconoscibili elementi riconducibili all'Antico

Egitto e a un contesto storico tradizionale africano, la rappresentazione dell'epoca contemporanea risulta significativamente meno evidente. Gli elementi visivi osservabili rimandano prevalentemente a un mercato caratterizzato da scambi diretti di beni e da un'estetica tradizionale, senza la presenza di indicatori sufficientemente espliciti che identifichino una moderna città africana. Di conseguenza, una delle tre epoche richieste appare solo parzialmente rappresentata, riducendo il grado di aderenza del risultato rispetto alle specifiche del prompt.

Anche il requisito relativo all'interazione tra personaggi appartenenti a epoche differenti risulta soddisfatto solo in parte. Sebbene la scena presenti numerosi individui che condividono lo stesso spazio e interagiscono tra loro, non emerge chiaramente una distinzione visiva tra le diverse epoche storiche, rendendo difficile identificare interazioni esplicite tra personaggi appartenenti ai tre periodi richiesti. Questo suggerisce una limitata capacità del modello di integrare simultaneamente molteplici contesti temporali mantenendone una chiara separazione semantica.

Dal punto di vista della qualità visiva, invece, il risultato è complessivamente elevato. Il video presenta una buona definizione delle immagini, movimenti fluidi e una composizione spaziale coerente. La scena appare ricca di dettagli, con un numero elevato di personaggi distribuiti in maniera naturale all'interno del mercato, contribuendo a migliorare il realismo della sequenza generata.

Per quanto riguarda il vincolo relativo ai cartelli contenenti testo leggibile, la generazione non evidenzia in modo chiaro la presenza di scritte perfettamente interpretabili. Questo comportamento è coerente con uno dei limiti ancora frequentemente osservati nei modelli generativi multimodali, che tendono a produrre testo deformato, incompleto o poco leggibile quando devono sintetizzare elementi grafici complessi all'interno delle immagini o dei video.

Nel complesso, il modello dimostra una buona capacità di costruire una scena articolata e visivamente realistica, mantenendo una composizione coerente e un elevato livello qualitativo. Tuttavia, il rispetto dei vincoli semantici complessi risulta soltanto parziale. In particolare, la rappresentazione simultanea delle tre epoche storiche e la chiara identificazione dell'epoca contemporanea non vengono pienamente soddisfatte, evidenziando una difficoltà del modello nel gestire richieste che comportano la fusione esplicita di contesti temporali distinti all'interno della stessa scena. Questo risultato conferma come l'aumento del numero di vincoli semantici e temporali comporti una maggiore probabilità di omissione o di rappresentazione incompleta di alcune condizioni specificate nel prompt.

5.3.3 Terzo esempio

In questa terza fase di sperimentazione si è scelto di valutare il comportamento del modello in presenza di vincoli o di richieste che violano deliberatamente le leggi della fisica e della percezione visiva. Il prompt è pensato appositamente per combinare una serie di eventi impossibili che devono verificarsi simultaneamente all'interno della stessa scena, richiedendo al modello di mantenere la coerenza temporale e spaziale pur rappresentando azioni contrarie al comportamento naturale degli oggetti. Questo test consente di analizzare la capacità del sistema di interpretare correttamente istruzioni di alta creatività e di generare una sequenza video credibile dal punto di vista visivo, nonostante l'irrealtà degli eventi descritti. Ecco il prompt inserito:

Utente

Genera una scena in cui piove verso l'alto, una persona sta camminando però il suo riflesso fa un'altra cosa, le ombre si muovano indipendentemente, e poi lancia una palla, che si ferma e poi ritorna indietro.

Di seguito il risultato ottenuto dal sistema:

Figura 5.5: Link per la visualizzazione del video generato dal prompt

Il prompt proposto richiede la generazione di una scena caratterizzata da numerosi eventi che violano le leggi della fisica e della percezione ordinaria. In particolare, il modello deve rappresentare simultaneamente pioggia che cade verso l'alto, una persona il cui riflesso compie azioni differenti rispetto al corpo reale, ombre che si muovono indipendentemente dai rispettivi soggetti e una palla che, dopo essere stata lanciata, si arresta a mezz'aria per poi invertire spontaneamente la propria traiettoria. Si tratta quindi di un test volto a valutare la capacità del modello di generare scene altamente controfattuali mantenendo la coerenza visiva e temporale.

L'analisi del video evidenzia che il modello incontra notevoli difficoltà nel soddisfare le condizioni imposte dal prompt. La prima violazione richiesta, ovvero la pioggia che cade verso l'alto, non viene rappresentata correttamente. Il comportamento delle precipitazioni appare, infatti, conforme alle normali leggi della gravità, oppure risulta poco distinguibile, senza mostrare chiaramente il moto ascendente richiesto.

Anche il vincolo relativo al riflesso della persona non viene rispettato. Il riflesso osservato segue i movimenti del soggetto principale, come avviene in una scena fisicamente

plausibile, senza eseguire azioni indipendenti. Questo comportamento suggerisce che il modello mantiene una forte priorità verso la coerenza fisica appresa durante l'addestramento, privilegiando configurazioni realistiche rispetto a richieste esplicitamente impossibili.

Il requisito riguardante le ombre autonome risulta anch'esso assente o non sufficientemente evidente. Le ombre, quando presenti, rimangono coerenti con il movimento dei rispettivi personaggi e non mostrano comportamenti indipendenti, indicando una limitata capacità del modello di dissociare elementi che, nella distribuzione dei dati di addestramento, risultano fortemente correlati.

L'unico elemento del prompt che sembra essere parzialmente interpretato riguarda la palla. Nella parte finale del video compare improvvisamente una palla che entra nella scena senza che sia chiaramente visibile il momento del lancio. Successivamente essa si solleva verso l'alto e ricade, ma il comportamento osservato non corrisponde alla dinamica richiesta dal prompt. In particolare, non si osserva una chiara fase di arresto nello spazio seguita da un'inversione della traiettoria verso il punto di origine. L'evento appare, quindi, scollegato dalla sequenza narrativa e introduce una discontinuità temporale che riduce la coerenza complessiva del video.

Dal punto di vista qualitativo, il video mantiene una buona resa grafica e una discreta fluidità delle animazioni. Tuttavia, la qualità visiva non è accompagnata da un'adeguata fedeltà semantica rispetto alle istruzioni fornite. Il modello privilegia, infatti, la produzione di una scena plausibile dal punto di vista fisico, ignorando o reinterpretando la maggior parte delle richieste controfattuali.

Nel complesso, questo esperimento evidenzia uno dei limiti tipici degli attuali modelli di generazione video: la difficoltà nel rappresentare simultaneamente molteplici eventi che contraddicono le regole fisiche e percettive del mondo reale. Sebbene il modello produca una sequenza visivamente coerente, esso tende a ricondurre la generazione verso configurazioni realistiche apprese durante l'addestramento, soddisfacendo solo marginalmente i vincoli semantici imposti dal prompt. Questo comportamento suggerisce che la modellazione delle relazioni fisiche apprese costituisca un forte bias nella generazione, limitando la capacità del sistema di rappresentare scenari deliberatamente impossibili o surreali.

5.4 Valutazione della complessità delle prestazioni di ModifAI

L'analisi sperimentale condotta precedentemente evidenzia che ModifAI possiede una buona capacità di interpretare prompt contenenti richieste relativamente articolate, riuscendo a generare video caratterizzati da un'elevata qualità visiva, una buona fluidità temporale e una composizione della scena generalmente coerente. Nei casi in cui il prompt richiede un numero limitato di vincoli semantici, il modello dimostra una soddisfacente aderenza alle istruzioni, preservando la coerenza dell'identità dei soggetti, la continuità delle azioni e l'organizzazione spaziale degli elementi presenti nella scena.

Quando c'è un aumento della complessità del prompt, emerge una progressiva riduzione della fedeltà rispetto alle specifiche richieste. In particolare, quando le istruzioni combinano numerosi vincoli simultanei, quali la coesistenza di epoche differenti, interazioni tra molteplici soggetti, eventi temporalmente complessi o fenomeni che violano le leggi della fisica, il modello tende a soddisfare soltanto una parte delle condizioni, omet-

tendo, oppure semplificando, alcuni elementi considerati meno prioritari riducendo, quindi, la qualità delle risposte generate.

I risultati mostrano, inoltre, che ModifAI privilegia la coerenza visiva e la plausibilità percettiva rispetto alla fedeltà assoluta al prompt. Anche in presenza di richieste esplicitamente controfattuali, il sistema tende, infatti, a generare sequenze compatibili con le regolarità statistiche apprese durante l'addestramento, producendo scene realistiche ma non sempre conformi alle istruzioni fornite dall'utente. Tale comportamento risulta particolarmente evidente nei test che richiedono la rappresentazione simultanea di eventi fisicamente impossibili o di relazioni spazio-temporali non convenzionali.

Dal punto di vista qualitativo, la resa grafica dei video rimane generalmente elevata anche nei casi in cui alcuni vincoli semantici non vengono rispettati. Questo ci evidenzia dunque che l'architettura del modello privilegi la qualità percettiva della generazione rispetto all'esecuzione rigorosa di tutte le condizioni presenti nel prompt.

Nel complesso, i risultati ottenuti indicano che le prestazioni di ModifAI sono strettamente correlate alla complessità delle istruzioni fornite. Il modello si dimostra affidabile nella generazione di scene realistiche e semanticamente semplici o moderatamente complesse, mentre evidenzia limitazioni nella gestione di prompt che richiedono la soddisfazione simultanea di numerosi vincoli logici, temporali e fisici. Queste osservazioni confermano una delle principali sfide ancora aperte nell'ambito della generazione video mediante modelli di Intelligenza Artificiale: mantenere un elevato livello di fedeltà al prompt senza compromettere la coerenza visiva e narrativa della sequenza generata.

CAPITOLO 6

Discussione

6.1 Premessa

Lo scopo principale di questa tesi è stato quello di analizzare e valutare la potenzialità e i limiti di alcuni modelli dell'Intelligenza Artificiale Generativa attraverso lo studio di tre differenti paradigmi di generazione: text-to-text, text-to-image e text-to-video. A tal fine sono stati selezionati tre modelli rappresentativi, ovvero Deepseek, DALL-E e ModifAI, e gli item sono stati sottoposti a una valutazione sperimentale mediante la realizzazione di case study specifici.

Nel corso del lavoro di tesi sono stati presentati i risultati per ciascun modello. C'è stata un analisi del comportamento del modello per diversi prompt di complessità differenti, per poter valutare la capacita di soddisfare i vincoli semantici richiesti, la qualità dei contenuti generati e le limitazioni riscontrate durante la sperimentazione. In questo capitolo viene sviluppato brevemente una sintesi complessiva dei risultati ottenuti, confrontando i tre modelli al fine di evidenziarne le caratteristiche comuni, le loro differenze e le rispettive aree di applicazione. L'obbiettivo finale è quello di fornire una valutazione complessiva dello stato attuale dei sistemi di Intelligenza Artificiale Generativa, evidenziando le sfide e i possibili sviluppi futuri.

6.2 Discussione conclusivo nel case study basato su Deepseek

Il case study basato su DeepSeek ha evidenziato le potenzialità del modello nell'elaborazione di contenuti testuali e nella risoluzione di compiti che richiedono capacità di ragionamento, comprensione del contesto e generazione di risposte strutturate. Durante le prove svolte, il sistema ha dimostrato una buona capacità di interpretare le richieste dell'utente, adattando le proprie risposte alle istruzioni fornite e mantenendo un elevato livello di coerenza anche in presenza di prompt articolati.

L'analisi ha, inoltre, dimostrato il fatto che DeepSeek è in grado di supportare efficacemente attività quali la produzione di testi, la sintesi di informazioni e l'assistenza nella risoluzione di problemi, offrendo risultati generalmente accurati e pertinenti. Benché

il modello presenta alcune limitazioni, soprattutto nei casi che richiedono conoscenze altamente specialistiche o una comprensione particolarmente approfondita del contesto, rimane comunque uno strumento affidabile e versatile, capace di affiancare l'utente in numerosi ambiti applicativi.

6.3 Discussione conclusivo nel case study basato su DALL-E

Analizzare il case study dedicato a DALL·E ci ha permesso di esaminare l'efficacia del modello nella generazione di immagini a partire da descrizioni testuali, mettendone in luce sia i punti di forza sia le principali criticità del sistema. I risultati ottenuti hanno evidenziato che il sistema è in grado di produrre immagini di alta qualità quando i prompt sono formulati in modo chiaro e specifico. Al contrario, richieste più articolate, specifiche o contenenti numerosi dettagli possono generare risultati meno accurati, rendendo necessario affinare progressivamente il prompt per ottenere un'immagine coerente con le aspettative.

È stato anche osservato che, nonostante i notevoli progressi compiuti dai modelli di generazione di immagini, permangono alcune limitazioni, come la difficoltà nella rappresentazione di relazioni spaziali complesse, nella gestione di elementi molto numerosi o nella riproduzione precisa di determinati dettagli, ad esempio testi sugli immagini. Per questo motivo, l'intervento dell'utente continua a svolgere un ruolo fondamentale sia nella definizione delle istruzioni sia nella selezione e nell'eventuale modifica delle immagini prodotte.

In sintesi, DALL·E rappresenta uno strumento particolarmente efficace come supporto ai processi creativi, consentendo di accelerare la realizzazione di concetti visivi, prototipi e contenuti grafici. Comunque, il suo impiego in ambiti professionali richiede ancora una supervisione umana per garantire accuratezza, coerenza con i requisiti progettuali e qualità del risultato finale.

6.4 Discussione conclusivo nel case study basato su ModifAI

Il case study condotto su ModifAI ha dimostrato le buone capacità del modello nella generazione di video a partire da descrizioni testuali fornite. In particolare, il sistema è riuscito a produrre contenuti caratterizzati da una buona qualità visiva, movimenti fluidi e una composizione della scena generalmente coerente, soprattutto quando i prompt presentano un livello di complessità basso o moderato.

Tuttavia, l'analisi ha mostrato che, quando si aumenta la complessità delle richieste, il modello tende a soddisfare solo parzialmente i vincoli specificati nel prompt. Le principali difficoltà sono legate alla rappresentazione simultanea di molteplici elementi, al rispetto di vincoli temporali complessi e alla generazione di eventi che violano le leggi della fisica. In tali situazioni, ModifAI privilegia spesso la plausibilità visiva della scena rispetto alla completa aderenza alle istruzioni fornite.

Insomma, i risultati ottenuti durante le nostre sperimentazioni confermano che ModifAI rappresenta un modello efficace per la generazione di video realistici, pur evidenziando alcune limitazioni nella gestione di prompt particolarmente articolati. Queste osservazioni sono allineate con le principali sfide ancora aperte nel campo della generazione text-to-video, dove il miglioramento del controllo semantico e della coerenza temporale costituisce tuttora un importante ambito di ricerca.

Conclusioni

Nel corso dello svolgimento della presente tesi è stato affrontato il tema dell'Intelligenza Artificiale Generativa, analizzandone le nozioni fondamentali, l'evoluzione storica e le principali architetture che hanno contribuito al suo rapido sviluppo. In particolare, sono stati approfonditi i modelli di generazione dei contenuti, basati su linguaggio naturale, su immagini e video, evidenziandone il funzionamento, le caratteristiche distintive e le principali sfide che è necessario affrontare utilizzando questi modelli.

Successivamente, il nostro lavoro si è concentrato sull'analisi sperimentale di tre differenti sistemi di Intelligenza Artificiale Generativa: DeepSeek, DALL·E e ModifAI, rappresentativi, rispettivamente, dei paradigmi text-to-text, text-to-image e text-to-video. Attraverso la progettazione di specifici case study è stato possibile valutare il comportamento dei modelli in presenza di prompt caratterizzati da differenti livelli di complessità, osservandone la capacità di interpretare le istruzioni fornite, partendo da quelle più semplici, verso quelle più complesse, rispettando i vincoli richiesti e producendo contenuti coerenti con le aspettative dell'utente.

L'analisi ha evidenziato il livello di accuratezza di ogni modello nella generazione di risultati di elevata qualità quando le richieste risultano chiare e ben strutturate. Al tempo stesso, è emerso una limitazione comune a tutti i tre modelli, cioè l'aumento della complessità dei prompt comporta una progressiva riduzione della fedeltà alle istruzioni, soprattutto quando è necessario soddisfare simultaneamente numerosi vincoli semantici, temporali o logici. Tale comportamento è risultato particolarmente evidente nel caso della generazione dei video, dove il mantenimento della coerenza temporale e il controllo preciso della scena rappresentano ancora importanti sfide e i risultati prodotti non erano abbastanza soddisfacenti.

I risultati ottenuti confermano, quindi, il notevole livello di maturità raggiunto dagli attuali modelli di Intelligenza Artificiale Generativa, ma evidenziano anche come tali sistemi non siano ancora in grado di garantire un controllo completo del processo di generazione. Per questo motivo, la qualità del prompt fornito e la supervisione dell'utente continuano a svolgere un ruolo determinante nell'ottenimento di risultati affidabili e coerenti con gli obiettivi prefissati. Un prompt mal strutturato, ovviamente, ci fornirà dei risultati incompleti o non conformi alle aspettative dell'utente.

Guardando al futuro, è ragionevole aspettarsi un ulteriore miglioramento delle capacità dei modelli generativi, favorito dall'evoluzione delle architetture neurali, dall'aumento delle risorse computazionali e dallo sviluppo di tecniche di condizionamento sempre

più sofisticate. In particolare, i progressi nel campo della generazione multimodale e dei modelli text-to-video consentiranno di ottenere sistemi più controllabili, coerenti e in grado di interpretare istruzioni sempre più complesse.

In conclusione, questo lavoro ha permesso di approfondire sia gli aspetti teorici sia quelli sperimentali dell'Intelligenza Artificiale Generativa, evidenziando le potenzialità offerte da questa tecnologia e le limitazioni che caratterizzano lo stato dell'arte. Nonostante le sfide ancora aperte, i rapidi progressi del settore lasciano prevedere che questi modelli avranno un ruolo sempre più rilevante nello sviluppo di applicazioni innovative, contribuendo a trasformare il modo in cui vengono creati e utilizzati i contenuti digitali.

Bibliografia

BOMMASANI, R.HUDSON, D.ADELI e OTHERS (2021), «On the Opportunities and Risks of Foundation Models», *arXiv*.

BROWN, T.MANN, B.RYDER e OTHERS (2020), *Language Models are Few-Shot Learners*, Advances in Neural Information Processing Systems.

BURKOV, A. (2019), *The Hundred-Page Machine Learning Book*, BURKOV, A.

FOSTER, D. (2023), *Generative Deep Learning (2nd Edition)*, O'Reilly Media.

GOODFELLOW, I., BENGIO, Y. e COURVILLE, A. (2016a), *Deep Learning (Adaptive Computation and Machine Learning series*, The MIT Press.

GOODFELLOW, I., BENGIO, Y. e COURVILLE, A. (2016b), *Deep Learning (Adaptive Computation and Machine Learning Series)*, The MIT Press.

HO, J.JAIN e A.ABBEEL (2020), *Denoising Diffusion Probabilistic Models*, Advances in Neural Information Processing Systems.

IUSZTIN, P. e LABONNE, M. (2025), *The LLM Engineering Handbook*, Packt Publishing.

OUYANG.L, JIANG.X e OTHERS (2022), «Training language models to follow instructions with human feedback», *arXiv*.

RADFORD, A.KIM, W., J., HALLACY.C e OTHERS (2021), «Learning Transferable Visual Models from Natural Language Supervision», *International Conference on Machine Learning (ICML).*

RAFFEL, C.SHAZEER, N.ROBERTS e OTHERS (2020), «Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer», *ournal of Machine Learning Research*.

RASCHKA, S. (2024), *Build a Large Language Model (From Scratch)*, Manning Publications.

TOUVRON, H.MARTIN, L.STONE, K.E e OTHERS (2023), «LLaMA: Open and Efficient Foundation Lamguage Models», *arXiv*.

Sitografia

- ultralytics.com https://www.ultralytics.com/it/glossary/text-to-image#convalida-dei-contenuti-generati
- dubsmart ai https://dubsmart.ai/it/blog/what-is-text-to-image-and-how-does-it-work
- humai https://www.humai.it/semplifichiamo/le-sfide-del-text-to-video/
- Video generati dall'AI: come funnzionano i nuovi modelli https://www.ai4business.it/intelligenza-artificiale/video-generati-dalla i-come-funzionano-i-nuovi-modelli/
- Wikipedia – https://en.wikipedia.org
- Oracle –https://en.wikipedia.org
- OpenAi – openai.com
- Tabnine– tabnine.com
- MIT Technology Review – technologyreview.com
- Github – github.com
- All about AI– www.allaboutai.com
- DEV Community – dev.to
- GeeksforGeeks – geeksforgeeks.org
- arXIV – https://arxiv.org/list/cs.AI
- MIT Computer Science and Artificial Intelligence Laboratory– https://www.weforum.org/organizations/mit-computer-science-and-artificial-intelligence-laboratory-csail/

Ringraziamenti

Je tiens à adresser ma profonde reconnaissance à Dieu Tout-Puissant, dont la grâce et les bénédictions m'ont accompagné tout au long de ce parcours académique.

Je tiens à adresser des remerciements particuliers au Professeur Domenico Ursino pour sa disponibilité, ses conseils, ainsi que l'attention avec laquelle il a suivi mon travail, ce qui a grandement contribué à l'aboutissement de ce parcours dans les délais impartis.

Je remercie mon père, M. Ngantchou Pierre, l'homme qui m'a tout donné. Je le remercie pour ses sacrifices et pour son dévouement dans la construction de la femme que je suis devenue. Je remercie également ma mère, le pilier de ma vie, mon guide, qui chaque jour façonne et protège la femme que je deviens. Je n'oublie pas mes petits frères et sœurs, qui ont été une véritable source de motivation pour avancer même dans les moments difficiles.

Je remercie également Papa Flaubert Yepmo, pour l'aide qui a rendu cette aventure possible.

Je remercie la famille Nkouabou, ma deuxième famille, qui m'a entourée d'amour et de soutien constant.

Des remerciements particuliers vont à ma grande famille au Cameroun, à ma grande famille à Ancône, ainsi qu'à mes proches, mes amis et mes frères et sœurs, qui ont été d'un véritable soutien tout au long de ce parcours.

Je remercie tout particulièrement ma zia Maureen, ma deuxième maman, mama Pierette, ainsi que Monsieur et Madame Djoum, qui m'ont accueillie comme leur propre fille.

Enfin, j'exprime ma sincère gratitude envers mes amis, sœurs et frères, : Ange, Arol, Brandy, Cathy, Deborah, Dafne, Daphne, Elmira, Erika, Evelyn, Frederick, Ivanna, Josiane, Loïc, Megane, Pavel, Verone, Verne, Zizou, ainsi que toutes les personnes qui, d'une manière ou d'une autre, ont contribué au bien de ma santé mentale durant ce parcours et à la réussite de ce travail.

A letter to the futur me:

Whenever you find yourself doubting the goodness of God, take a step back and read these words.

I do not know where life will have taken you by the time you read this, but I hope you are proud of how far you have come and of everything you have overcome to reach this point.

You have managed to build your dreams, develop your projects and pursue your ambitions while going through this journey, and that alone is something truly remarkable.

Being an ambitious woman is not a weakness, it is a strength. You carry the full potential of a powerful woman capable, independent, and full of purpose. You can be anything you choose: a leader, a creator, or even a devoted homemaker if that is the path you decide to embrace. Never let anyone define your limits. Trust God's timing and do not chase people or relationships. A good man, the one meant for you, is already on your path and will come into your life at the right time. Focus on becoming the best version of yourself, and everything destined for you will arrive naturally.

Your worth is not defined by society, but by the life you build, the love you give, and the impact you create.

The future is bright, and every dream placed in your heart will one day come to life. Never forget that success is built on two pillars: Prayer and Hardwork. Keep your heart pure, stay kind to everyone you meet, and always act with integrity, because what you give to the world will always find its way back to you.

Never stop dreaming. Remember that even finishing this journey in 2 years and 9 months once felt impossible, yet here you are today. Today, you are an engineer, but this is only the beginning. You are called to become a great woman, strong, resilient, and guided by faith, a woman known for her hard work and unwavering trust in God.

And always remember, our God is a covenant-keeping God.