



**UNIVERSITÀ DEGLI STUDI DI PADOVA**

DIPARTIMENTO DI FILOSOFIA, SOCIOLOGIA, PEDAGOGIA E PSICOLOGIA APPLICATA  
DIPARTIMENTO DI SCIENZE ECONOMICHE E AZIENDALI "MARCO FANNO"

**CORSO DI LAUREA IN COMUNICAZIONE**

Intelligenza artificiale generativa nel marketing digitale:  
Un confronto per capire quale modello performa meglio nel copywriting.

Relatore:  
Ch.mo Prof. Marco Bettiol

Laureando: Davide Dorio  
Matricola n. 2051216

**ANNO ACCADEMICO 2023- 2024**



# INDICE

<b>INTRODUZIONE</b> .....	<b>II</b>
<b>Capitolo 1</b> .....	<b>1</b>
<b>IA generativa</b> .....	<b>1</b>
1.1 Introduzione all'IA generativa.....	1
1.2 Machine learning.....	2
1.2.1 Il ruolo dei dati nel machine learning.....	2
1.3 Deep learning.....	3
fig.1: Una mappa che identifica a che livello si posiziona il Deep Learning.....	3
1.4 Applicazioni dell'intelligenza artificiale generativa .....	4
1.4.1 Intelligenza artificiale applicata al marketing.....	4
1.4.2 Altre applicazioni dell'intelligenza artificiale generativa.....	5
<b>Capitolo 2</b> .....	<b>7</b>
<b>Modelli di Intelligenza Artificiale</b> .....	<b>7</b>
2.1 ChatGPT .....	7
2.1.1 ChatGPT 3.5 .....	7
2.2.2 Introduzione a ChatGPT-4 .....	9
2.2.3 Prestazioni di GPT-4 .....	10
2.2.4 Limitazioni di GPT-4.....	12
2.2 Claude.....	13
2.2.1 Claude 3 Opus .....	13
2.2.2 Claude 3 Sonnet e Claude 3 Haiku .....	14
2.2.3 Performance di Claude 3 .....	15
2.2.3 Limitazioni di Claude 3.....	16
2.3 Google Gemini .....	17
2.3.1 Performance di Gemini 1.0 e 1.5 .....	18
2.3.2 Limitazioni di Google Gemini .....	20
2.4 Meta LLaMA 2.....	22
2.4.1 Performance di LLaMa 2.....	22
2.4.2 Limitazioni di Meta LLaMa 2 .....	24
2.5 Microsoft Copilot, Phi e Megatron-Turing .....	25
2.5.1 Performance di Megatron-Turing e Phi-3.....	26
2.5.2 Limiti di Megatron-Turing e Phi-3 .....	28
<b>Capitolo 3</b> .....	<b>30</b>
<b>Comparazione Delle IA</b> .....	<b>30</b>
3.1 Metodo comparativo .....	30
3.2 Comparazione.....	32
3.3 Risultati .....	35
<b>Conclusioni</b> .....	<b>44</b>
<b>Bibliografia</b> .....	<b>46</b>
<b>Sitografia</b> .....	<b>49</b>

# INTRODUZIONE

L'intelligenza artificiale generativa (che successivamente chiameremo IA generativa) è forse una delle innovazioni più affascinanti degli ultimi decenni. Questa tecnologia in costante evoluzione sta trasformando molti aspetti del nostro quotidiano, in particolare nel contesto lavorativo. Già nell'ultimo anno si sono visti i vantaggi sia in termini di tempistiche sia di efficienza che questi nuovi strumenti possono portare in attività come quelle del marketing digitale. In questo elaborato andremo ad approfondire il concetto di intelligenza artificiale e di IA generativa e le sue applicazioni nel mondo del marketing. Successivamente si descriverà lo stato dell'arte attuale di questo mercato mettendo in evidenza le principali figure che stanno cambiando il mondo di oggi tramite questi strumenti. Analizzeremo quindi i modelli che mettono a disposizione le aziende di OpenAI, Google, Anthropic, Microsoft e Meta. Approfondiremo i diversi modelli per capire cosa li distingue gli uni dagli altri, come questi possano essere utilizzati nella miglior maniera tramite un'introduzione a questi, una analisi delle performance basate su benchmark e un'analisi dei limiti che questi presentano. Si vuole analizzare poi le motivazioni che ci portano alla parte empirica della ricerca ovvero il motivo per il quale i test che ci presentano tanto spesso le aziende non sono sempre da prendere in considerazione nell'utilizzo di determinati strumenti. Successivamente andremo a introdurre gli approcci al prompt engineering, una nuova competenza che sta prendendo sempre più piede nel mercato del lavoro attuale, che consiste nella formulazione della giusta domanda per massimizzare la probabilità di ottenere una risposta desiderata. Nella parte empirica si comparano i modelli ChatGPT-3.5, Google Gemini 1.5 e Claude-3 Sonnet ovvero i 3 modelli più diffusi e utilizzati nella generazione di testo per delle campagne marketing per vedere quale strumento risulta più chiaro, creativo e generi curiosità nello spettatore. Seguiranno poi le conclusioni con delle riflessioni sul futuro di questa tecnologia.



# Capitolo 1

## IA generativa

### 1.1 Introduzione all'IA generativa

L'IA generativa (chiamata anche GAI o Gen AI) è un ramo della ricerca nel campo dell'intelligenza artificiale in grado di generare dati come testo, immagini, musica, video e algoritmi tramite tecniche di Machine Learning e Deep Learning (Sætra, 2023).

L'IA, infatti, è allenata su una grande quantità di dati al fine di produrre una risposta in base alle richieste dell'utente, emulando alcune abilità cognitive umane. Non solo apprende ma ragiona e svolge compiti che prima implicavano risorse umane.

Strumenti come "ChatGPT" e quelli che andremo ad analizzare in questo elaborato sono intelligenze artificiali generative specializzate in output di forma testuale. Stiamo parlando di large Language model (LLM) che sono specializzati nel Natural Language processing (NLP) ovvero l'abilità nelle macchine di leggere, comprendere e ricavare i significati dai linguaggi umani.

Se chiediamo a ChatGPT di descrivere cos'è un LLM, ci risponderà con la seguente elaborazione:

"Un large Language model (LLM) è un tipo di modello di linguaggio che è stato progettato per comprendere e generare testo in modo simile all'uso umano del linguaggio. Questi modelli sono in grado di analizzare grandi quantità di testo, apprendere modelli linguistici e quindi generare testo coerente e significativo in risposta a input o istruzioni specifiche."

L'azienda che ha introdotto al mondo ChatGPT è OpenAI e propone attualmente più versioni sul mercato:

- una versione gratuita chiamata ChatGPT 3.5 Turbo
- Una versione a pagamento chiamata ChatGPT 4 ovvero il modello più intelligente e capace includendo nel suo abbonamento DALL-E, un modello per la creazione di immagini proprietario di OpenAI.

Nell'ultimo anno molte altre aziende hanno sviluppato le loro implementazioni di LLM come PaLM 2 e Med PaLM 2 (sviluppato con l'obiettivo di aiutare medici e infermieri) di Google, LLaMA di Meta e la famiglia di modelli Claude 3 di Anthropic.

## 1.2 Machine learning

Per capire come queste macchine imparano dobbiamo approfondire il concetto di machine learning che rappresenta una sotto-disciplina dell'IA che unisce l'utilizzo degli algoritmi alla statistica. Questo processo si occupa di creare sistemi che apprendono o migliorano le performance in base ai dati che utilizzano (Oracle, 2022). Il machine learning è utilizzato e integrato ovunque, quando interagiamo con le banche, acquistiamo online o utilizziamo una piattaforma social, permettendoci di avere un'esperienza d'acquisto o navigazione più personalizzata e sicura.

Gli algoritmi di machine learning attualmente utilizzati sono principalmente due e si distinguono dal modo in cui apprendono dati per fornire delle previsioni.

- **Machine learning supervisionato:** sono gli algoritmi più utilizzati, con questi il data scientist è la guida che insegna all'algoritmo i risultati da generare. Un esempio di modello supervisionato (o supervised) può essere rappresentato dalle aziende che sviluppano tecnologie e algoritmi per le auto a guida autonoma dove viene assunto del personale addetto all'addestramento delle macchine.
- **Machine learning non supervisionato:** è un approccio indipendente, dove non è prevista la costante guida di una persona. In questo caso la macchina impara a identificare processi e schemi complessi da sola. Questo approccio invece viene utilizzato spesso per IA specializzate nell'analisi dei dati.

Esiste poi un'altra tecnica riguardante l'apprendimento delle macchine chiamato "apprendimento rinforzato" (Reinforced Learning o RL). Con questo approccio l'IA inizia prendendo decisioni sperimentali che porteranno ad un rinforzo negativo o positivo. In poche parole, è come se ogni volta che si esegue un compito si ottenesse una ricompensa (un punteggio alto) oppure una punizione (una perdita di punti). Col passare del tempo l'IA apprenderà quali siano le decisioni più accurate e con più successo con l'obiettivo di massimizzare l'ottenimento della ricompensa. Un esempio di apprendimento rinforzato riguarda l'elenco dei video suggeriti di YouTube e la sezione "Per te" di TikTok che presentano contenuti in base alla cronologia e interazioni.

### 1.2.1 Il ruolo dei dati nel machine learning

Quando si tratta di definire quale sia l'intelligenza artificiale migliore uno dei fattori più importanti da valutare (se non il più importante) riguarda i dati con i quali. Un'IA in effetti per essere addestrata ha bisogno di un'enorme quantità di dati senza i quali non può imparare e svilupparsi. Non parliamo solo di quantità di dati, ma soprattutto di qualità di questi che se non

garantita dai Data Scientist, che devono selezionare con estrema cura i dataset per i loro progetti, aumentano di molto le probabilità di sbagliare da parte del modello (Chen, 2023). In definitiva per ottenere un adeguato addestramento di un modello di IA i dataset sui quali si basano devono rispettare tre criteri:

- **Qualità delle fonti dei dati:** Quando un modello di intelligenza artificiale viene alimentato da dei dataset poco o per niente controllati, di scarsa qualità e omogenei, i risultati che riceveremo in output saranno scadenti. I dati più adeguati e validi dipendono dal modello e dai suoi utilizzi. Quando si notano degli errori inaccettabili è possibile riqualificare l'IA ma non è raro che i data-scientist decidano di ricominciare il progetto da zero dopo che i dati abbiano contaminato il modello.
- **Volume dei dati:** Anche il volume dei dati è molto importante in quanto si parla di addestramento delle IA in quanto un solo dataset rappresenta solo l'inizio del processo di formazione. Una quantità di dati maggiore garantisce un'adeguata diversità dei dati aumentando la precisione e aiutando l'identificazione dei dati anomali.
- **Diversità dei dati:** Diversificare i set di dati porta l'algoritmo a una maggiore precisione durante l'addestramento, ampliando la capacità decisionale e una maggiore profondità di conoscenza attraverso molteplici scenari.

### 1.3 Deep learning

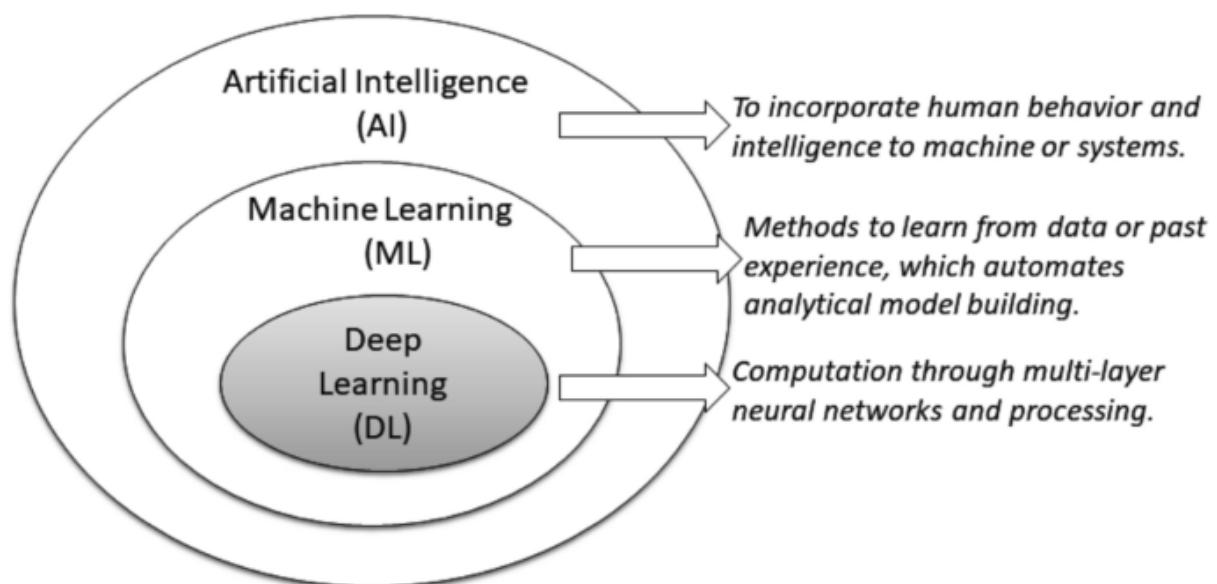


fig.1: Una mappa che identifica a che livello si posiziona il Deep Learning

Il Deep learning come evidenziato nella fig.1 è un sottoinsieme del Machine Learning che a sua volta è una parte del più grande insieme dell'intelligenza artificiale.

Spesso si tende a confondere questi due sottoinsiemi (Machine Learning e Deep Learning).



Se il Machine Learning (ML), come visto nel precedente approfondimento, riguarda il metodo di apprendimento dei dati tramite modelli analitici, il Deep Learning (DL) rappresenta anch'esso metodi di apprendimento dei dati ma attraverso reti neurali che lavorano su più strati. Il termine “Deep” si riferisce infatti al concetto di multistrato composto da più livelli o fasi attraverso i quali vengono elaborati i dati per costruire un modello basato su questi (Sarker, 2021).

Il DL rappresenta quindi un componente fondamentale in quanto alza il livello delle IA, aiutando il modello a processare le informazioni e decisioni in modo più intelligente.

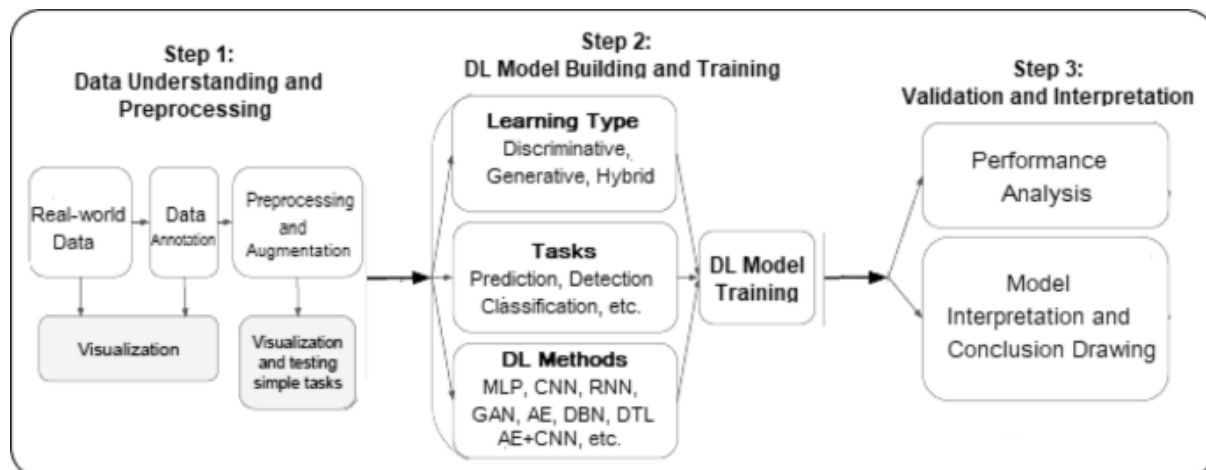


fig.2: Un flusso di lavoro per la risoluzione di problemi reali diviso in 3 fasi sequenziali: (1) comprensione ed elaborazione dei dati, (2) creazione e formazione del modello DL e (3) convalida e interpretazione.

## 1.4 Applicazioni dell'intelligenza artificiale generativa

Nell'ultimo anno si è visto l'avvento di molteplici modelli di IA generative che differiscono principalmente in base ai dati e alle modalità con cui imparano e apprendono le informazioni. Per questo motivo le IA non sono tutte uguali o tutte adatte ad ogni situazione. Esistono modelli più adatti alla creazione di contenuti piuttosto che all'analisi dei dati e creazione di grafici o immagini. In questo capitolo, che precede la panoramica dei modelli che andremo a confrontare, capiremo quali sono le applicazioni di questa tecnologia nell'attuale.

### 1.4.1 Intelligenza artificiale applicata al marketing

l'utilizzo dell'IA generativa nel marketing sta crescendo esponenzialmente, già a marzo del 2023 già il 73% delle organizzazioni utilizza strumenti come chatbot nelle loro attività (Dencheva 2023a). I professionisti del marketing ritengono che l'intelligenza artificiale generativa permetta loro di migliorare la loro produttività.

Secondo il sondaggio condotto da Salesforce in collaborazione con YouGov condotto tra il 18 e il 25 maggio 2023 su 1029 operatori nel marketing, il 51% del campione utilizza strumenti di generative AI mentre il 22% prevede di utilizzarli “molto presto” con ulteriori risultati che affermano che queste tecnologie consentono ai lavoratori di risparmiare oltre 5 ore a settimana.

Un ulteriore sondaggio condotto da “Boston Consulting Group” condotto in aprile del 2023 su 200 CMO di diversi settori in 8 paesi del nord America, Europa e Asia afferma che il 70% di loro utilizza già AI generative, il 19% era ancora in fase di test e solo il 3% non aveva intenzione di utilizzare tali strumenti. Su quest'ultimo sondaggio i principali utilizzi riguardano la personalizzazione (che consente di avvicinarsi al cliente comprendendo le loro esigenze al fine di migliorare il servizio o prodotto offerto), la generazione di insights (analisi dei dati), e la creazione di contenuti creativi più rapidi e veloci stimolando il pensiero creativo.

Un esempio di personalizzazione riguarda le banche, che tramite l'analisi dei dati dei loro clienti riescono a sfruttare l'IA generativa per offrire consigli di investimento personalizzati in base al profilo di rischio. Anche i rivenditori utilizzano strumenti di personalizzazione al fine di convertire il cliente finale e di ottenere una maggiore fidelizzazione (Kshetri 2024).

Anche la generazione di insights è stato un altro utilizzo importante della generativa AI ma solo recentemente. Condurre ricerche di mercato affidabili può richiedere risorse umane e tempistiche che si traducono in ulteriori costi per l'azienda. I recenti progressi fatti con questa tecnologia consentono di ridurre notevolmente il tempo per la generazione di approfondimenti e la ricerca di dati grezzi consentendoci di elaborarli molto più facilmente.

Un ulteriore campo importante nel quale l'IA ha fatto la differenza è il content marketing, ovvero quello che il Content Marketing Institute nel 2015 definisce come “l'approccio al marketing basato sulla creazione e distribuzione di contenuti di valore, pertinenti e coerenti per attrarre e trattenere un pubblico chiaramente definito e, in definitiva, per guidare un'azione redditizia da parte dei clienti”. Per contenuti si intendono video, foto, testi o post sui social con il fine di aumentare l'engagement dei clienti e aumentare le vendite. L'IA generativa ci permette di trovare nuove idee più velocemente e creando contenuti adeguati ed opportuni ad ogni target in modo rapido e con costi accessibili a tutti.

#### 1.4.2 Altre applicazioni dell'intelligenza artificiale generativa

Anche nel campo della formazione l'IA può aiutare studenti e studentesse nella ricerca di informazioni per trovare domande a risposte specifiche o aiutandoli nella comprensione di concetti più articolati. Anche i professori possono giovare di tali strumenti in modo da ottimizzare i piani didattici per gli studenti aiutandoli a coinvolgerli maggiormente per far sì che ne traggano il meglio dell'offerta formativa (Fui-Hoon Nah, 2023).

In ambito sanitario come riportato da Kung Chat GPT supera gli esami di licenza medica degli stati uniti (Kung, 2023). Questo significa che questi strumenti possono fornire un utile aiuto ai professionisti del settore sanitario nel momento di diagnosi cliniche o l'educazione sanitaria. Questo campo rimane però in dubbio in quanto il settore sanitario è altamente regolamentato e fare troppo affidamento sui contenuti generati dall'intelligenza artificiale generativa potrebbe portare a gravi problematiche. Perciò gli esperti sono molto cauti nell'utilizzo di tecnologie come queste che potrebbero compromettere l'accuratezza delle informazioni.

Un'altra applicazione dell'IA generativa riguarda il settore delle risorse umane. Nel momento della selezione del personale possiamo usare strumenti adatti per un iniziale screening e analisi dei profili dei candidati che una volta assunti potremmo formare con risorse che l'IA stessa potrà aiutarci a produrre (Rane, 2023). Anche in questo caso nel momento della creazione di survey per avere dei riscontri o verificare l'engagement potremmo chiedere domande metodologie per condurre ricerche all'interno dell'azienda. Una volta ottenuti i risultati dai questionari sarà possibile analizzarli e ottenere grafici attraverso l'utilizzo di modelli addestrati apposta nell'analisi dei dati (come perplexity o i diversi GPT's).

# Capitolo 2

## Modelli di Intelligenza Artificiale

### 2.1 ChatGPT

“ChatGPT” è un modello di intelligenza artificiale sviluppato da OpenAI, azienda fondata da Sam Altman con l’aiuto di Elon Musk, che più ha fatto discutere negli ultimi anni. Questo modello è stato progettato ad allenato per la generazione di testo in output che sia in linea con ciò che viene richiesto dall’utente. L’obiettivo di ChatGPT come altre chatbot è quello di creare una conversazione il più possibile simile a quelle umane.

OpenAI permette l’utilizzo gratuito della versione 3.5 e consente di effettuare un abbonamento da un prezzo accessibile che dà accesso alla versione 4 che comprende anche il suo modello generativo di immagini “DALL-E 3”. Tuttavia, gli ingenti investimenti da parte di Microsoft nella start-up di ChatGPT ci permettono di utilizzare Copilot ovvero un modello supportato da ChatGPT-4, e integrato nel motore di ricerca loro proprietario, che consente agli utenti di utilizzare un chatbot potenziato dal modello di punta di OpenAI con accesso a internet in modo limitato.

#### 2.1.1 ChatGPT 3.5

ChatGPT-3.5 è il primo modello di intelligenza artificiale generativa che OpenAI ci permette di utilizzare in modo gratuito e senza limitazioni, oltre che essere il più utilizzato negli ultimi anni. Per capire in cosa consiste, nella presentazione di questo LLM l’azienda madre ci propone quattro casistiche che ci aiutano a comprendere il funzionamento e le possibili utilità di questo strumento tra cui:

- Un debugging (ovvero il processo tramite il quale si ricercano e individuano errori o bug nel codice di un software con la finalità di risolverli) di una porzione di codice dove si richiede di trovare una soluzione ad un malfunzionamento non previsto.
- Il rifiuto da parte del modello di rispondere a domande che sospetta riguardino attività illegali e potenzialmente dannose per la salute umana, per dimostrare la sicurezza che garantisce questo modello.
- Un riassunto della discussione appena avvenuta per dimostrare che tiene in memoria i prompt (ovvero la richiesta da noi formulata per data al modello) fatti e le risposte ottenute in precedenza.
- Una risposta ad una istruzione di follow-up.

Per allenare questo modello OpenAI afferma di aver utilizzato un l'apprendimento rinforzato dal feedback umano (RLHF, Reinforced Learning from Human Feedback), lo stesso utilizzato in precedenza per InstructGPT, ma con leggere differenze e miglioramenti nella raccolta dei dati.

InstructGPT è un modello che mostra i miglioramenti nella veridicità e riduzione nella generazione di output tossici con una diminuzione minima delle prestazioni di elaborazione dei dati. Come detto nello studio dei modelli linguistici della Cornell University, pur commettendo errori semplici, i risultati con l'aiuto di un feedback umano sono una direzione promettente per l'allineamento degli LLM con l'intento umano soprattutto tramite un sistema di ricompensa che aiuta l'algoritmo nella sua fase di apprendimento.

Il sistema di ricompensa per l'apprendimento e rinforzo è avvenuto tramite la raccolta di dati di confronto provenienti da diversi messaggi creati dal modello. Da questi messaggi sono state create poi ulteriori alternative che i formatori avevano il compito di classificare in base alla qualità. Questo modello di ricompensa ha poi dato la possibilità di ottimizzare quello che loro chiamano Proximal Policy Optimization (PPO) ovvero un algoritmo di ottimizzazione della policy prossimale che rende l'IA più efficiente nella risoluzione di compiti complessi.

Questo processo viene svolto molteplici volte al fine di migliorare ulteriormente il modello.

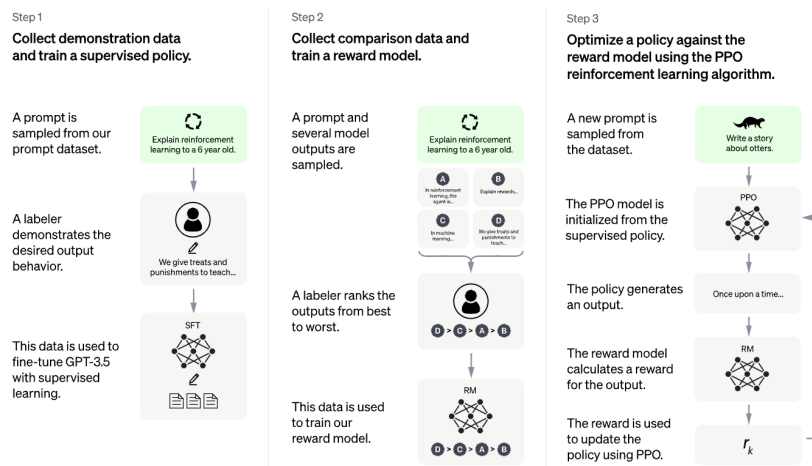


fig.3: Semplificazione grafica di come viene allenato il modello ChatGPT-3.5 di OpenAI tramite il sistema di ricompense. Nella prima parte di utilizza il termine “fine-tune” che riguarda il processo di ottimizzazione delle risposte e che approfondiremo in seguito.

Non essendo il modello di punta di OpenAI questo porta con sé delle limitazioni che sono simili a quelle che vedremo anche nella gran parte dei futuri modelli che approfondiremo.

Il primo limite che non esprimono nella loro presentazione riguarda l'accesso a internet da parte del modello 3.5 che lo limita ad un database aggiornato solo al gennaio 2022.

Se da una parte questo limite non ci consenta di fare domande successive con informazioni che non sono state rilasciate successivamente a quella data da un'altra ci permette di avere

un modello più sicuro con dati selezionati esclusivamente da data scientist evitando i rischi che l'accesso ad internet ne comporterebbe come la contaminazione dai dati e la perdita di qualità di questi ultimi.

Ulteriori limitazioni del modello ce le dice direttamente OpenAI:

- "ChatGPT può scrivere risposte in parte plausibili ma che errate o prive di senso"
- "ChatGPT è sensibile alle modifiche alla frase di input o ai tentativi di ripetere lo stesso messaggio più volte."
- "Il modello è spesso eccessivamente prolisso e abusa di alcune frasi, come ribadire che si tratta di un modello linguistico addestrato da OpenAI"
- "Idealmente il modello porrebbe domande chiarificatrici quando l'utente fornisce una query ambigua, mentre i nostri modelli attuali si solito indovino l'intento dell'utente"
- "anche se ci siamo impegnati affinché il modello rifiuti richieste inappropriate, a volte risponderà a istruzioni dannose o mostrerà un comportamento parziale. Utilizziamo l'API di moderazione per avvisare o bloccare determinati tipi di contenuti non sicuri, ma per ora prevediamo che contenga alcuni falsi positivi e falsi negativi"

Nonostante i limiti gli sviluppatori si mostrano aperti a tutti i feedback che gli utenti possono fornire per migliorare questa tecnologia.

## 2.2.2 Introduzione a ChatGPT-4

GPT-4 è un modello multimodale su larga scala in grado di elaborare in input immagini e testo per produrre un output testuale. Sebbene meno capace degli umani in molti contesti reali, GPT-4 rappresenta il prodotto di punta di OpenAI. Uno dei motivi per i quali questo modello supera il suo predecessore 3.5 riguarda le prestazioni che a livello umano superano benchmark professionali e accademici, arrivando al superamento di un esame di abilitazione simulato raggiungendo il 10% dei migliori partecipanti al test.

Questo modello è importante anche per quanto riguarda lo sviluppo di applicazioni come sistemi di dialogo, riassunti testuali e traduzioni automatiche.

Una delle differenze principali con il modello precedente riguarda gli input visivi che risiedono nella sua multimodale, che consente all'utente di intrecciare dati testuali e dati provenienti da immagini per generare un output testuale. Un caso di utilizzo nel quotidiano di input visivi misti input testuali può essere visto nella fig.4 nella quale si chiede la soluzione di un problema che richiederebbe conoscenze delle quali non disponiamo.

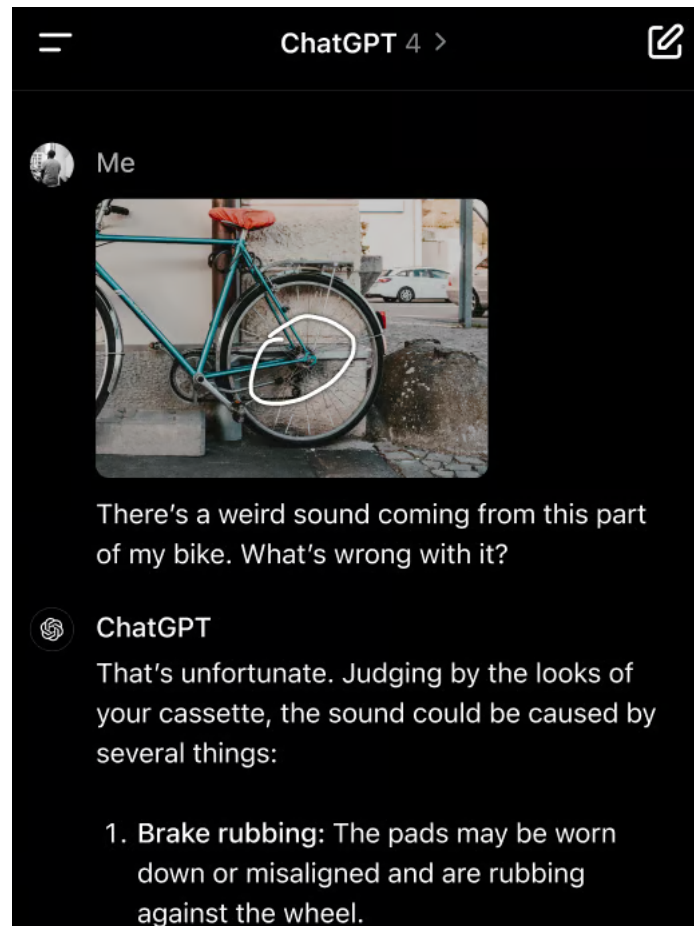


fig.4: richiesta tramite un input visivo e testuale ed elaborazione di una soluzione al problema sottoposto da parte di Chat-GPT.

Un ulteriore passo in avanti fatto da questo modello riguarda la creazione di quelli che vengono chiamati "GPT's" ovvero Chatbot personalizzati e addestrati a uno specifico scopo. Facilmente accessibili attraverso il GPTStore questi possono essere creati anche dai singoli utenti per venderli attraverso lo store interno. Questa feature offre a tutti i professionisti, specialmente digitali (come copywriter, Social media manager o Data Analyst), la possibilità di avere diversi assistenti virtuali addestrati nelle mansioni a loro più specifiche utili, permettendo agli sviluppatori ulteriori possibilità di utilizzo delle API che OpenAI mette a disposizione per poter creare le proprie applicazioni.

### 2.2.3 Prestazioni di GPT-4

Prima della pubblicazione come per GPT-3.5 sono stati pubblicati anche i benchmark di questo modello che ci consentono di vedere i progressi fatti nell'addestramento e sviluppo del progetto.

GPT-4 non supera solo il suo predecessore ma secondo dei benchmark PNL supera la maggior parte dei LLM più all'avanguardia come PaLM-2 di Google. Nel benchmark MMLU

(una suite di domande a scelta multipla che copre 57 argomenti) GPT-4 supera i modelli esistenti con un margine rilevante specialmente in lingua inglese, dimostrando ottime capacità anche nelle altre lingue.

**GPT-4 3-shot accuracy on MMLU across languages**

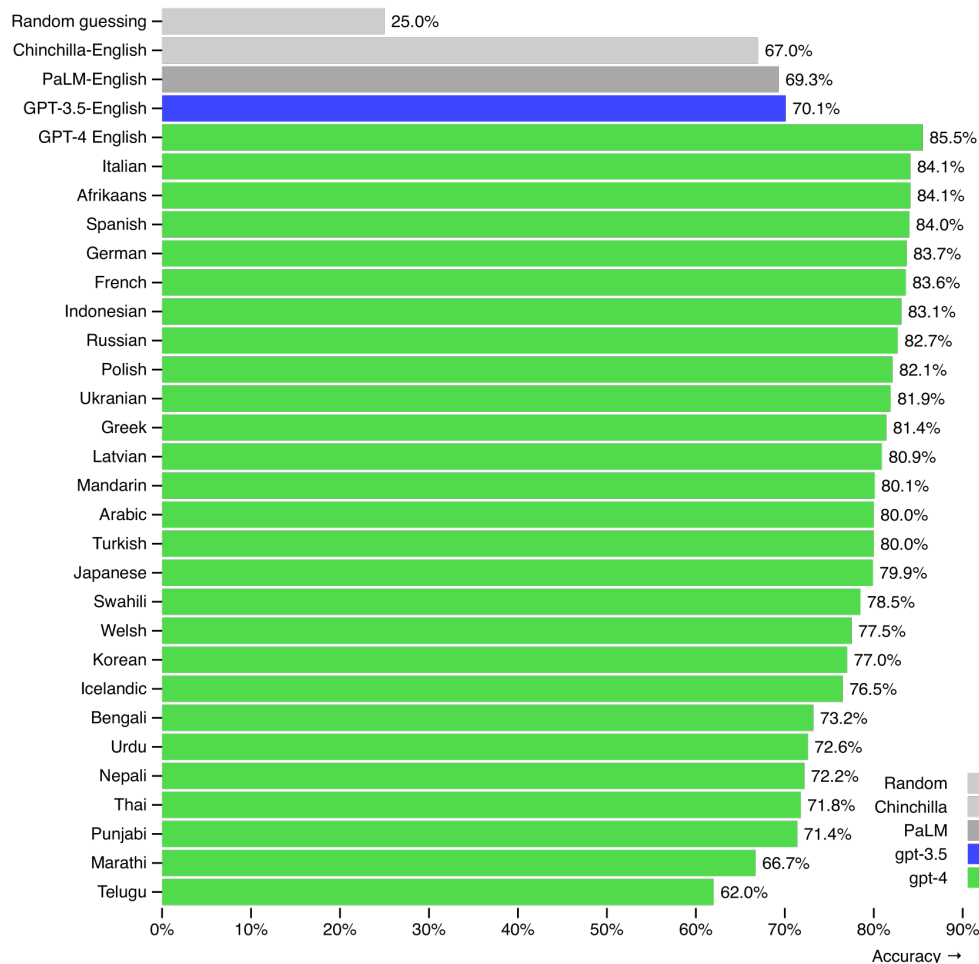


fig.4: prestazioni di chat GPT-4 in varie lingue rispetto ai modelli precedenti su MMLU.

GPT-4 migliora rispetto al modello precedente anche nella sua capacità di interpretazione del contesto fornitogli tramite i prompt fornitogli dagli utenti che su un set di 5.214 prompt proposti, le risposte dell'ultimo modello sono state preferite il 70,2% delle volte rispetto a GPT-3.5. Sono stati effettuati poi dei test più tradizionali e creati apposta per modelli di machine learning che dimostrano come GPT-4 sia in grado di performare meglio degli altri LLM già esistenti in diverse materie che riguardano per esempio la scrittura di codice, come il test "HumanEval" che richiede la soluzione di compiti nel linguaggio di programmazione "Python", oppure il test "AI2 Reasoning Challenge (ARC)" che riguarda invece più l'ambito scientifico. Tra questi quello più diffuso e del quale parleremo più spesso sarà l'MMLU (Massive Multitask Language Understanding) che misura le prestazioni complessive dei diversi Large Language Model. Questo benchmark si pone l'obiettivo di misurare la capacità di un modello di elaborare e ragionare su informazioni complesse attraverso domande a scelta multipla riguardanti la



soluzione a determinati problemi e altre che richiedono la soluzione di task tramite ragionamenti. Il secondo che teniamo a osservare è chiamato “HellaSwag” che valuta invece le prestazioni a delle domande di senso comune e che riguardano la vita di tutti i giorni.

Benchmark	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (includes benchmark-specific training)
<b>MMLU</b> Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM	75.2% 5-shot Flan-PaLM
<b>HellaSwag</b> Commonsense reasoning around everyday events	95.3% 10-shot	85.5% 10-shot	84.2% LLAMA (validation set)	85.6% ALUM
<b>AI2 Reasoning Challenge (ARC)</b> Grade-school multiple choice science questions. Challenge-set.	96.3% 25-shot	85.2% 25-shot	84.2% 8-shot PaLM	85.6% ST-MOE
<b>WinoGrande</b> Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	84.2% 5-shot PALM	85.6% 5-shot PALM
<b>HumanEval</b> Python coding tasks	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM	65.8% CodeT + GPT-3.5
<b>DROP (f1 score)</b> Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM	88.4 QD-GAT

fig.5: Valutazione di GPT-4 in confronto ad altri modelli nei benchmark più tradizionali per LLM.

## 2.2.4 Limitazioni di GPT-4

Come nel caso di GPT-3.5 anche la versione aggiornata condivide limiti simili. Come cita OpenAI stessa, tuttora ChatGPT-4 non è del tutto affidabile ed è necessario prestare attenzione agli output in particolare nelle situazioni di elevato rischio. Questo rappresenta uno dei maggiori problemi che nella versione GPT-4 è stato ridotto del 40% rispetto alla versione 3.5.

Può capitare anche che dopo la richiesta di un prompt si perda dei dettagli che non considera e se sottoposto un problema questo modello può dare delle risposte non corrette e con errori nei risultati, con la necessità di reiterare il nostro prompt più volte prima di ottenere la risposta desiderata.

Pur riducendo considerevolmente la quantità di risposte “inventate” o anche denominate “allucinazioni” rispetto al modello precedente del 19% questo problema persiste.

GPT-4 non è connesso ad internet (se non tramite l'utilizzo plugin) e di conseguenza non conosce gli eventi che si sono verificati dopo settembre 2023. Oltretutto non impara dalla sua esperienza e può commettere semplici errori di ragionamento, produrre risposte ingenuie e accettare dichiarazioni palesemente false.

Non ricontrollando più volte il testo generato, è molto probabile che commetta sbagli nelle sue previsioni.

Anche questo modello presenta dei bias che, come scritto dagli sviluppatori si sono impegnati a correggere ma che richiede tempo per gestirlo completamente.

L'obiettivo che l'azienda si è imposta è quello di creare sistemi (come Chat-GPT) che abbiano comportamenti predefiniti e ragionevoli che consentano la personalizzazione entro dei limiti al fine di ottenere input pubblici sui quali dovrebbero essere applicati tali limiti.

## 2.2 Claude

“Claude” è il modello LLM di Anthropic, startup statunitense fondata da degli ex membri di OpenAI. Questa azienda ha come obiettivo quello di realizzare strumenti di intelligenza artificiale generativa affidabili, sicuri, interdisciplinari e gestibili per il miglioramento della sicurezza delle IA.

La loro ultima versione e le precedenti usano una architettura di tipo transformer e sono allenate attraverso un apprendimento non supervisionato (unsupervised learning), Reinforced Learning from Humane Feedback (RLHF), apprendimento supervisionato e apprendimento rinforzato. “Claude 3 Opus” è il modello più potente che è stato rilasciato a marzo del 2024(Anthropic, 2024), a suo fratello minore “Claude 3 Sonnet” e “Claude 3 Haiku” (quest'ultimo rilasciato qualche giorno dopo). Questo modello tende a performare positivamente nelle conversazioni generali che comprendono la ricerca, la scrittura e riassunti di testo e codice. A eccellere però è nel supporto dei lavori creativi o di letteratura dove riesce a seguire le indicazioni su “personalità” e tono che gli utenti definiscono guidabili e conversazionali.

### 2.2.1 Claude 3 Opus

Claude 3 Opus è il modello più intelligente di Anthropic che dichiara di avere le migliori performance sul mercato per quanto riguarda il completamento dei compiti più difficili.

Ulteriori miglioramenti possono essere trovati nella parte di analisi dei dati e previsioni, revisioni, brainstorming e generazione di idee. Un'ulteriore implementazione molto importante riguarda la possibilità da parte di questo modello di poter leggere (a differenza dei modelli Claude 2) diversi formati visivi, tra cui foto, diagrammi, grafici e diagrammi tecnici.

La dimensione del modello si aggira attorno a 175 miliardi di parametri ed è addestrata su dati di una alta qualità, provenienti principalmente da testi inglesi e includendo siti web e articoli online e molto diversificati.

Una delle tecniche utilizzate nel pre-addestramento di questo modello viene chiamata “Span Corruption” (Ye, 2024). Questa nuova procedura migliora significativamente l'efficienza dei modelli e si divide in cinque parti:

1. La preelaborazione dei dati: Il corpus testuale di addestramento viene diviso in segmenti di una determinata lunghezza che chiamiamo “span” (di solito una span è una frase o un paragrafo).
2. La corruzione: Durante l’addestramento viene “corrotta” una percentuale casuale di span detto anche “token”. Questo processo crea “rumore” all’interno dei dati per simulare circostanze nelle quali il modello deve correggere il corpus testuale, attraverso la ricostruzione o il completamento del testo. La corruzione può avvenire in molteplici modi:
  - a. Attraverso la rimozione di parole
  - b. Attraverso la sostituzione di parole con altre casuali
  - c. Attraverso la permutazione casuale dell’ordine delle parole
  - d. Attraverso l’aggiunta di parole casuali non correlate
3. L’addestramento alla ricostruzione: Il modello viene allenato a prevedere e ricostruire gli span corrotti in span originali.
4. Masked Language Modeling: È una tecnica che riguarda la corruzione nella quale una parte dello span viene sostituita con un token speciale chiamato “maschera”. In questo caso il modello deve riconoscere il token mascherato e prevedere l’originale.
5. Curriculum Learning: Inizialmente vengono corrotti un piccolo numero di token per span per poi aumentare gradualmente le corruzioni nell’avanzare dell’addestramento per consentire al modello di imparare prima dai casi più semplici fino ad arrivare ai casi più complessi.

La tecnica dello span-corruption, di conseguenza, forza il modello a imparare a modellare il contesto migliorando la sua capacità di comprensione e generazione di testo con un linguaggio naturale.

### 2.2.2 Claude 3 Sonnet e Claude 3 Haiku

Claude 3 Sonnet rappresenta il predecessore di Opus ed è definito come il giusto compromesso tra potenza e velocità. È basato su una architettura Encoder-Decoder, ed addestrato su 7 miliardi di parametri. Oltre all’essere stato ottimizzato sotto diversi aspetti, questa nuova versione è stata anch’essa addestrata tramite delle tecniche di denoising che consentono al modello una capacità superiore di comprensione e generazione del testo sensato e coerente.

Claude 3 Haiku rappresenta invece uno dei modelli più piccoli di intelligenza artificiale che, nonostante la sua compatta dimensione, si distingue per la sua velocità nella lettura e analisi ed elaborazione dei dati arrivando ad analizzare fino a 21.000 token al secondo.

## 2.2.3 Performance di Claude 3

In questa famiglia di modelli l'apice, come detto in precedenza è rappresentato da Opus che è definito dall'azienda madre stessa (Anthropic) come un nuovo standard per l'IA. Opus supera infatti la maggior parte degli altri modelli concorrenti nei benchmark più comuni per la valutazione delle intelligenze artificiali. Non solo a livello di comprensione mostra dei livelli quasi umani ma anche nell'analisi, previsione di contenuti, creazione di codice e conversazione in multiple lingue Opus raggiunge apici mai toccati sui benchmark da una IA.

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge <i>MMLU</i>	86.8% 5-shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning <i>GPQA, Diamond</i>	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math <i>GSM8K</i>	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving <i>MATH</i>	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math <i>MGSMT</i>	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code <i>HumanEval</i>	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text <i>DROP, FI score</i>	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations <i>BIG-Bench-Hard</i>	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A <i>ARC-Challenge</i>	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge <i>Hellaswag</i>	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

fig. 6: risultati dei benchmark pubblicati da Anthropic in relazione alla famiglia di Claude 3 messa a confronto con i modelli concorrenti.

Quando si parla di velocità invece iniziamo a notare qualcosa di interessante. Tra i tre il più lento risulta proprio essere Opus con performance simili a quelle di Claude 2 e 2.1 pur avendo un livello di intelligenza molto più elevato. Sonnet invece riesce a raggiungere 2 volte la velocità del modello precedente ma il modello che sorprende di più è Haiku che riesce a leggere un articolo di circa 10.000 token compresi diagrammi e grafici in meno di 3 secondi. Quest'ultimo oltre che essere uno dei modelli più economici sul mercato, a detta degli sviluppatori, ha un ampio margine di miglioramento delle prestazioni che Anthropic prevede di sfruttare nel post lancio.

Un altro test molto interessante che sta prendendo sempre più importanza è il "Needle in a Haystack" (Anthropic, 2024), che valuta la capacità di ragionamento e recupero informazioni pertinenti da grandi quantità testuali.

Il funzionamento è molto semplice.

1. Viene fornito un documento con una grande quantità di testo che rappresenta il pagliaio
2. All'interno del documento vengono nascosti degli "aghi" che consistono in piccole porzioni rilevanti di testo
3. Il modello analizza l'intero corpus testuale con il fine di individuare con precisione questi piccoli "aghi"

Nel complesso quello che misura il test è sostanzialmente la comprensione del linguaggio naturale, il ragionamento e l'inferenza, la capacità di recupero delle informazioni e l'attenzione. A detta di Anthropic questi modelli raggiungono "risultati eccezionali" e "prestazioni di punta" con una memoria che si avvicina alla perfezione.

Anche per quanto riguarda i refusi gli ultimi modelli rappresentano un notevole passo avanti rispetto alla generazione precedente.

#### Incorrect refusals

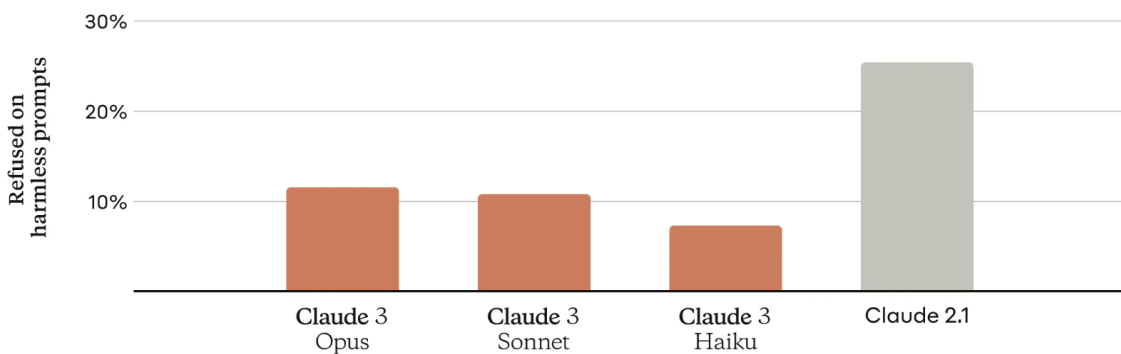


Fig.7: comparazione del numero refusi su prompt innocui espresso in percentuale.

### 2.2.3 Limitazioni di Claude 3

Anche questi modelli presentano delle limitazioni, simili a quelle viste in precedenza nel caso di OpenAI e i suoi modelli di Chat-GPT.

Il modello Opus seppur di grandi dimensioni risulta essere molto costoso per via delle sue necessità di grandi risorse computazionali e di archiviazione e, seppur migliorando sotto questo aspetto, potrebbe comunque generare testo con bias e contenuti tossici o dannosi.

Alcune scelte compiute da Opus possono ancora essere inspiegabili, e difficili da interpretare, assieme al fatto che potrebbe avere difficoltà nella comprensione di prompt molto lunghi e con un ampio contesto.

Claude 3 Sonnet in quanto fratello minore di Opus non raggiunge le sue stesse prestazioni ed essendo un modello più datato potrebbe diventare obsoleto più rapidamente. Sonnet

condivide inoltre con Opus le stesse limitazioni che riguardano i bias o lacune derivanti dai set di dati utilizzati in addestramento.

Haiku per quanto sia veloce e versatile, invece ha delle limitazioni che non gli permettono di performare come gli altri modelli, in particolare nel momento in cui deve interpretare contesti lunghi e complessi risultano utile sono in operazioni semplici e che richiedono una bassa capacità computazionale. In tutto questo però bisogna sempre ricordare che stiamo parlando di modelli diversi e con finalità diverse che quindi portano a limitazioni attualmente difficili da evitare.

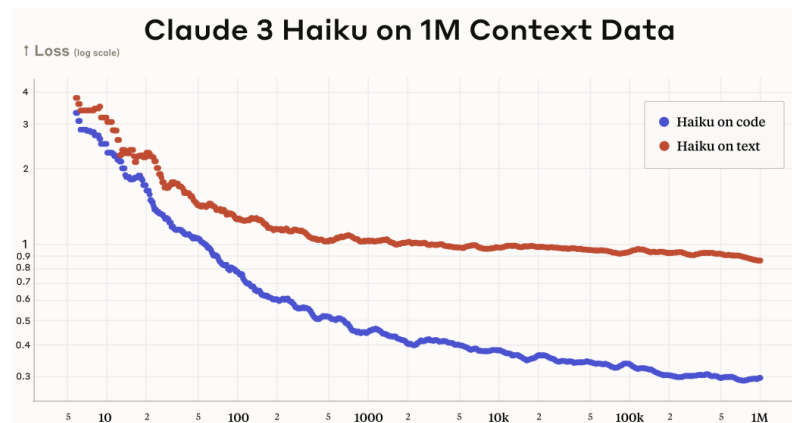


Fig.8: Il grafico raffigura la perdita all'aumento del contesto del modello Haiku. La linea rossa rappresenta la perdita sul testo mentre la linea blu rappresenta la perdita di contesto che riguarda il codice.

## 2.3 Google Gemini

Quando si parla di modelli AI tra i più conosciuti non si può non citare il modello del colosso Google che dopo aver rilasciato, nel 6 febbraio del 2023, il suo primo modello utilizzabile da tutti tramite cloud-computing denominato "Bard", quasi un anno dopo si presenta con una nuova versione a loro detta "rivoluzionaria" chiamata "Gemini". Già prima del suo rilascio al pubblico, Bard aveva fatto parlare di sé per alcune sue risposte che da parte del pubblico risultarono inesatte, persistenti anche dopo il suo rilascio agli utenti che lo evitavano per la sua bassa affidabilità e il continuo dare risposte errate e fuorvianti. Il 6 dicembre 2023 il CEO di Google, Sundar Pichai, annuncia sul blog ufficiale della multinazionale l'uscita della nuova versione che sostituirà da quel momento Bard e che prende appunto il nome di Gemini 1.0. Questa versione vuole essere vista come una famiglia con tre diversi modelli ognuno dei quali progettato in base ai requisiti applicativi e divisi esattamente come quelli che abbiamo visto nel caso di Anthropic e la sua linea di modelli Claude 3.

La famiglia è composta da tre modelli:

- Gemini Nano: il modello di più piccole dimensioni e più efficiente creato per alimentare chatbot che non necessitano una conoscenza vasta.
- Gemini Pro: il modello che si adatta a tutte le situazioni e il giusto compromesso tra potenza e dimensione.
- Gemini Ultra: il modello più grande e potente adatto ai compiti più complessi e dalla conoscenza più vasta.

Anche in questo caso, questi modelli sono dotati oltre che un text decoder anche di un Image decoder, permettendo all'utente di inserire e analizzare immagini oltre che crearle rendendo il modello di fatto multimodale come Claude e ChatGPT-4.

Questi modelli sono basati tutti sul Large Language Model chiamato "PaLM 2" di cui Google è proprietaria e che abbiamo già incontrato e visto nei diversi benchmark precedentemente analizzati. Oltre a questo Google ha sviluppato un altro modello chiamato "Med-PaLM" creato appositamente per aiutare coloro che lavorano nel settore ospedaliero (Singhal, 2023). I nuovi modelli della famiglia Gemini, sulla carta e nei diversi benchmark, sembrerebbero battere la concorrenza e in alcuni casi anche di gran lunga risultando come i modelli migliori sul mercato prima dell'uscita però di Claude 3 rilasciato successivamente (Gemini Team Google, 2023).

### 2.3.1 Performance di Gemini 1.0 e 1.5

Gemini 1.5 differisce dal suo predecessore principalmente nella sua capacità di comprendere e apprendere dal contesto facendone un utilizzo più efficace, estendendo la finestra contestuale dei 32 mila token del modello 1.0 a diversi milioni consentendogli un'elaborazione di input molto più grandi. I miglioramenti sono stati fatti anche dal punto di vista del recall (ovvero la capacità di un modello di ricordare e accedere a delle informazioni già viste in precedenza nella generazione di testo o risposta a domande) che, secondo il paper, raggiunge dei livelli vicini alla perfezione su attività di contesti estesi come nel caso di Claude.

La cosa che si nota fin da subito su questo modello è la rapidità di generare una risposta al nostro prompt nettamente più veloce di Chat-GPT 4.0 e a volte anche di Claude 3 Sonnet. Se questo aspetto sembra far apparire questi ultimi 2 modelli già "antiquati", i benchmark pubblicati nel paper di Gemini 1.5 non sembrano però così tanto rivoluzionari.

Prendiamo d'esempio il benchmark MMLU (Massive Multi-task Language understanding) visto in precedenza nel capitolo riguardante le performance di GPT- 4. Questo benchmark permette di valutare la capacità di un modello di comprendere diverse lingue e di misurare la capacità linguistica e conversazionale (Hendrycks, 2021). È composto da 57 materie comprese STEM, umanistiche, scienze sociali, ed è utilizzato per valutare il livello di apprendimento durante la fase di pre-addestramento e in un contesto di pochi shot rendendo,

come dice l'autore, questo test più difficile è simile a come veniamo valutati noi umani. Quando parliamo di shot intendiamo il numero di esempi di addestramento per un determinato compito (Parnami, Lee, 2022), per cui questo test valuta anche il quanto è in grado di imparare un modello da un numero di esempi limitati. L'ampiezza ma allo stesso tempo la specificità degli argomenti rendono questo benchmark ottimo per identificare i punti cecchi di un LLM. Nei risultati ottenuti si evince che il modello Gemini 1.0 Ultra raggiunge un punteggio più alto rispetto a tutti i modelli, arrivando a toccare un punteggio del 90%, superiore persino di Claude 3 Opus che pur essendo uscito successivamente ma che raggiunge "solo" un 88,2%.



Rank	Model	Average↑ (%)	Extra Training Data	Paper	Code	Result	Year
1	<b>Gemini Ultra ~1760B</b>	90	✓	<a href="#">Gemini: A Family of Highly Capable Multimodal Models</a>		📄	2023
2	<b>Claude 3 Opus</b> (5-shot, CoT)	88.2	✓	<a href="#">The Claude 3 Model Family: Opus, Sonnet, Haiku</a>		📄	2024
3	<b>Claude 3 Opus</b> (5-shot)	86.8	✓	<a href="#">The Claude 3 Model Family: Opus, Sonnet, Haiku</a>		📄	2024
4	<b>Leeroo</b> (Mix)	86.64	✓	<a href="#">Leeroo Orchestrator: Elevating LLMs Performance Through Model Integration</a>		📄	2024
5	<b>GPT-4</b> (few-shot)	86.4	✓	<a href="#">GPT-4 Technical Report</a>		📄	2023

fig. 9: Una classifica dei primi 5 modelli con punteggi del benchmark MMLU più alti e dove si può trovare in prima posizione il modello di punta della famiglia Google Gemini, ovvero Gemini Ultra che supera Claude 3 Opus di un 1.8%.

Da molte persone raggiungere questo livello vuol dire avere tra le mani il miglior modello attualmente utilizzabile, ma se andiamo a guardare un po' più dettagliatamente il paper di Gemini 1.5 possiamo notare una cosa interessante. Non solo nello stesso benchmark e in 5-shot performa peggio della versione 1.0 ma da qui si può notare che performi anche peggio di Claude 3 Opus che con 5-shot raggiunge un punteggio di 86,8% rispetto al 83,7% di Gemini 1.0 ultra e al 81,9% di Gemini 1.5 Pro. Nella restante parte dei benchmark, la versione 1.5 performa certamente in modo migliore rispetto al suo predecessore, ma questo ci fa capire come spesso i dati vengono usati in modo strategico dalle aziende per avere un vantaggio competitivo rispetto ai competitor.



Capability	Benchmark	Gemini		
		1.0 Pro	1.0 Ultra	1.5 Pro
Math, Science & Reasoning	<b>Hellaswag</b> (Zellers et al., 2019)	84.7% 10-shot	87.8% 10-shot	<b>92.5%</b> <b>10-shot</b>
	<b>MMLU</b> : Multiple-choice questions in 57 subjects (professional & academic). (Hendrycks et al., 2021a)	71.8% 5-shot	<b>83.7%</b> <b>5-shot</b>	81.9% 5-shot
	<b>GSM8K</b> : Grade-school math problems. (Cobbe et al., 2021)	77.9% 11-shot	88.9% 11-shot	<b>91.7%</b> <b>11-shot</b>
	<b>MATH</b> : Math problems ranging across 5 levels of difficulty and 7 sub-disciplines. (Hendrycks et al., 2021b)	32.6% 4-shot Minerva prompt	53.2% 4-shot Minerva prompt	58.5% 4-shot Minerva prompt <b>59.4%</b> <b>7-shot</b>
	<b>AMC 2022-23</b> : 250 latest problems including 100 AMC 12, 100 AMC 10, and 50 AMC 8 problems.	22.8% 4-shot	30% 4-shot	<b>37.2%</b> <b>4-shot</b>
	<b>BigBench - Hard</b> : A subset of harder tasks from Big Bench formatted as CoT problems. (Srivastava et al., 2022)	75.0% 3-shot	83.6% 3-shot	<b>84.0%</b> <b>3-shot</b>
	<b>DROP</b> : Reading comprehension & arithmetic. (Metric: F1-Score). (Dua et al., 2019)	74.1% Variable shots	<b>82.4%</b> <b>Variable shots</b>	78.9% Variable shots
Coding	<b>HumanEval</b> chat preamble* (Metric: pass rate). (Chen et al., 2021)	67.7% 0-shot (PT)	<b>74.4%</b> <b>0-shot</b> <b>(PT)</b>	71.9% 0-shot
	<b>Natural2Code</b> chat preamble* (Metric: pass rate).	69.6% 0-shot	74.9% 0-shot	<b>77.7%</b> <b>0-shot</b>
Multilinguality	<b>WMT23</b> : sentence-level machine translation (Metric: BLEURT). (Tom et al., 2023)	71.73 (PT) 1-shot	74.41 (PT) 1-shot	<b>75.20</b> <b>1-shot</b>
	<b>MGSM</b> : multilingual math reasoning. (Shi et al., 2023b)	63.45% 8-shot (PT)	78.95% 8-shot (PT)	<b>88.73%</b> <b>8-shot</b>

fig. 10: Benchmark del paper di gemini 1.5 dove si nota come in pochi shot raggiunge un livello inferiore a Claude 3 Opus e Chat GPT-4 nel MMLU.

Una cosa da sottolineare è la possibilità da parte di Gemini però di avere accesso ad internet e quindi di fornire informazioni in tempo reale, poter apprendere in modo continuo e avere una maggiore capacità di contestualizzazione.

### 2.3.2 Limitazioni di Google Gemini

Se da una parte avere accesso a internet rappresenta un'occasione vantaggiosa, dall'altra comporta diversi rischi (Bender, 2021). Lasciando al modello l'accesso a internet questo può venire esposto a contenuti dannosi (disinformativi o offensivi), che influenzano il suo apprendimento e comportamento causando la creazione di quelle che vengono chiamate allucinazioni. Non è raro, infatti, che in certi contesti Gemini inventi fatti non reali e mai accaduti. Un'altra cosa che l'accesso a internet non garantisce è la diversità dei dati per tre motivi:

- Internet non è uno spazio diversificato, non è distribuito uniformemente ma i dati da cui è popolato arrivano principalmente da utenti più giovani e provenienti dai paesi più sviluppati.

- Un LLM può amplificare i bias e pregiudizi in base ai dati a cui ha accesso.
- A livello di privacy e sicurezza il modello potrebbe esporre dati e informazioni sensibili e private.

Un'altra cosa che ha limitato Gemini è la sua abilità nel creare immagini, opzione rimossa in quanto il modello produceva delle elaborazioni eccessivamente “politically correct”. Con questo intendiamo il fatto che ha creato immagini palesemente non realistiche e spesso errate che non rappresentavano i prompt. Adi Robertson nel suo articolo pubblicato su “The Verge” spiega come questo problema sia diventato virale offrendo degli esempi.

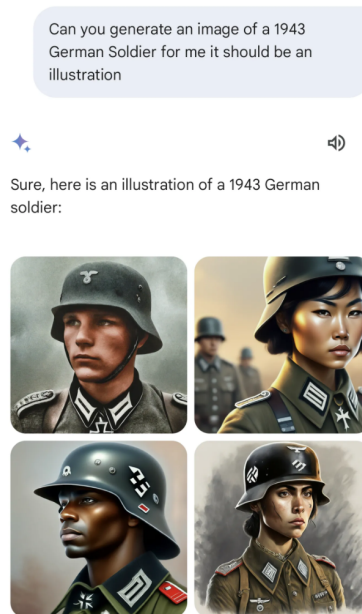


fig.11: un esempio di immagini create da Google Gemini al momento della richiesta della creazione di immagini con soldati tedeschi nel 1943, con un risultato poco coerente con il contesto dell'epoca.

Presentando modelli diversi con una suddivisione simile a quella di Claude 3, Gemini si porta dietro delle limitazioni analoghe. Come detto su Google Cloud nella sezione documentazione si può trovare un articolo che sottolinea le limitazioni del modello 1.0. In primo luogo, troviamo il ragionamento spaziale per il quale Gemini può avere difficoltà a individuare con precisione degli oggetti o del testo nelle immagini, per poi arrivare alla comprensione dei video con una lunga durata e avere qualche problema con attività che richiedono passaggi di ragionamento.

## 2.4 Meta LLaMA 2

Sviluppato da Meta, LLaMA (o Large Language Model Meta AI) rappresenta uno dei modelli più interessanti degli ultimi anni, con degli usi e degli utilizzi che potrebbero differire dai

concorrenti visti in precedenza. Addestrato su un numero di parametri che varia tra i 6,7 e i 65,2 miliardi si fa una distinzione tra le diverse versioni:

- LLaMa-7B: il modello più piccolo
- LLaMa-13B e LLaMa-33B: le 2 versioni intermedie.
- LLaMa-65B: il più grande e addestrato su 65 miliardi di parametri.

A differenza degli altri casi, LLaMa è stato addestrato esclusivamente su un set di dati disponibili pubblicamente senza la necessità di ricorrere a dati non accessibili o proprietari. I dati sui quali è stato addestrato comprendono diverse fonti come testi accademici, libri, articoli e dati provenienti da internet che comprendono anche le piattaforme dei diversi social media. Nel loro paper, il team di meta AI, ci tiene a sottolineare il fatto che tutti questi data-set sono compatibili con l'open-sourcing (ovvero, dati accessibili al pubblico che possono essere utilizzati da chiunque senza correre il rischio di violare il copyright o eventuali restrizioni). Come nei casi visti in precedenza anche questo modello presenta una architettura di base transformer con delle modifiche specifiche che mirano a migliorare l'apprendimento, che in questo caso avviene in maniera prevalentemente supervisionata.

#### 2.4.1 Performance di LLaMa 2

La differenza tra le varie versioni presentate nel paper e citate prima (LLaMa-7B, LLaMa-13B, LLaMa-33B, LLaMa-65B) sta principalmente nella dimensione del modello e della capacità di comprendere i contesti e rispondere in modo più simile a quello conversazionale umano. LLaMa-65-B in questo caso eccelle rispetto agli altri nella generazione di testo e nella comprensione di contesti più ampi e complessi, ma modelli più grandi necessitano risorse computazionali più potenti. Se invece vogliamo guardare i benchmark e la comparazione con i diversi modelli, nel paper troviamo alcuni test, diversi da quelli visti in precedenza e su modelli, oltre che molto più grandi e addestrati su un numero ben maggiore di dati, risultano essere più datati e precedenti a quelli sui quali si basano tuttora i modelli di punta di OpenAI, Google e Anthropic ma che rappresentano comunque un esempio interessante e da prendere in considerazione.

		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	<b>88.0</b>	82.3	-	83.4	<b>81.1</b>	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	<b>80.0</b>	<b>57.8</b>	58.6
	65B	85.3	<b>82.8</b>	<b>52.3</b>	<b>84.2</b>	77.0	78.9	56.0	<b>60.2</b>

fig.12: Una tabella che compara diversi modelli con diverse versioni di LLaMa 2 su benchmark di ragionamento basilari su un contesto zero-shot.

Nella figura 12 è illustrata una tabella presente nel paper del modello di META, sviluppato dal proprio team AI, che evidenzia come LLaMa 2, nella versione a 65B e 33B, si differenzia dal resto dei modelli per la sua capacità di raggiungere punteggi alti in contesti di zero-shot, pur essendo allenato su un numero molto più piccolo di dati rispetto alla competizione. Questo è un chiaro esempio di quanto sia importante la qualità dei dati che si utilizzano durante il training di questi modelli.

Dei test che abbiamo appena visto solo alcuni di essi fanno parte delle altre tabelle comparative viste in precedenza per questo andiamo successivamente ad analizzare quali sono i risultati che LLaMa 2 riesce a raggiungere anche in altri benchmark più comuni come MMLU.

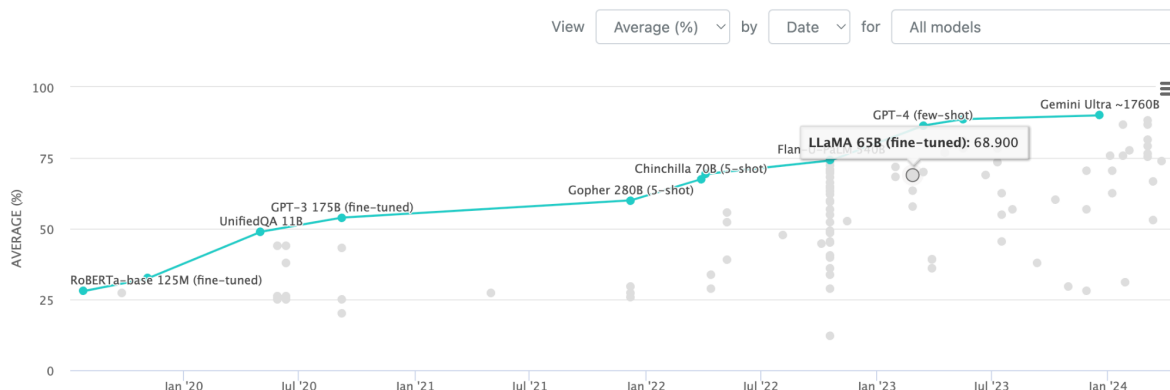


fig. 13: il risultato del benchmark MMLU di LLAMA 65B (fine-tuned) rispetto ai suoi competitor, dove raggiunge un punteggio di 68,9%.

In figura 13 troviamo i risultati che raggiunge il modello LLAMA nel benchmark MMLU che ci aiutano a capire e introdurre il concetto di fine-tuning, fondamentale per l'utilizzo dell'utente finale. Per fine-tuning ci riferiamo a quel processo che va a perfezionare il modello AI già

addestrato con il fine di adattarlo ad un compito specifico per raggiungere risultati migliori in esso. Nonostante questo, LLaMa raggiunge un punteggio di 68.9 su 100, non riuscendo nemmeno a sfiorare il risultato di 86.4 raggiunto da GPT-4 in un contesto few-shot.

In definitiva LLaMa è un modello interessante da analizzare in quanto raggiunge ottimi livelli in diversi test pur essendo allenato su una quantità di dati nettamente inferiore, evidenziando quanto detto in precedenza riguardo la qualità di questi che deve essere ottimale per raggiungere determinati risultati.

## 2.4.2 Limitazioni di Meta LLaMa 2

Il modello rilasciato dal gruppo di Meta AI rappresenta un significativo passo in avanti per quanto riguarda gli LLM, in quanto uno rappresenta uno degli open model (oltre che open source) più interessanti, ma porta con sé alcune limitazioni. In primo luogo, LLaMa non possiede attualmente un'interfaccia come nei casi visti in precedenza, per cui non è disponibile un chatbot che ne permetta l'utilizzo tramite il cloud computing (come invece si ritrova in tutti i casi visti in precedenza). Nonostante questo, è possibile, da parte di tutti gli sviluppatori, creare un chatbot utilizzando le API messe a disposizione direttamente da Meta, con l'opportunità di utilizzare il modello in locale per generare testo. Ciò comporta diverse problematiche in quanto questa operazione (pur essendo più sicura a livello di privacy dei dati) per essere svolta richiede una capacità computazionale non indifferente (come una CPU e una GPU altamente performanti, un'ampia capacità di storage per immagazzinare i miliardi di dati necessari e una memoria RAM che gira sulle decine di GB). Grandi aziende possono permettersi determinati hardware a differenza di gran parte degli sviluppatori indipendenti che spesso non hanno a disposizione una potenza di calcolo tale da poter utilizzare questo modello con facilità, in tempistiche ottimali e in tempo reale. Oltre a questo LLaMa 2 si porta dietro il peso di essere il modello più "datato", e meno aggiornato rispetto a tutti i suoi competitor anche se sappiamo che il lancio della terza versione di questo modello è imminente e potrebbe avere le potenzialità per essere un ulteriore significativo passo avanti per i modelli open source. Anche in questo caso si ripresentano tutte le limitazioni e problematiche intrinseche dei Large Language Model che riguardano l'amplificazione dei bias cognitivi presenti nei dati di allenamento, la possibilità di generare allucinazioni e di descrivere cose che non sono mai accadute. Infine, la natura generalista di LLaMa fa sì che riesca a rispondere a gran parte delle domande che gli vengono fatte ma non trova specializzazioni in alcun ambito specifico, e ciò porterebbe molte persone a scegliere modelli dei concorrenti molto più completi e specializzati in determinati settori.

## 2.5 Microsoft Copilot, Phi e Megatron-Turing

Nel mercato dell'intelligenza artificiale esistono poi grandi aziende che preferiscono piuttosto che sviluppare totalmente all'interno una propria IA, decidono di collaborare e investire pesantemente su aziende più piccole ma che si sono riuscite a distinguere in un ambiente che è diventato rapidamente un oceano rosso. Questo è il caso di Microsoft che ha deciso di sviluppare, assieme a OpenAI, Mistral e Anthropic, dei modelli addestrati appositamente per il supporto dell'utente nel suo utilizzo della suite di Office 365 e di altri prodotti o servizi che l'azienda fondata da Bill Gates ancora nel 1975, mette a disposizione ai suoi clienti, rendendola uno degli attori più importanti a livello globale per quanto riguarda il settore intelligenza artificiale.

I modelli progetti di GAI che Microsoft ha sviluppato sono principalmente 3:

1. Megatron-Turing NLG 530B: Il modello più grande, allenato su 530 miliardi di parametri, in grado di generare testi creativi e svolgere attività complesse come la traduzione e rispondere a domande in modo informativo.
2. Microsoft 365 Copilot: modello che aiuta gli utenti della suite 365 ad essere più efficienti svolgendo task complessi in poco tempo.
3. Phi-3: una famiglia di "mini modelli" di intelligenza artificiale allenati su un numero limitato di parametri e progettati per essere leggeri e accessibili su dispositivi che hanno una potenza di calcolo limitata.

Tutto questo permette a Microsoft di fornire un ecosistema completo ai suoi utenti interamente supportato da intelligenza artificiale generativa.

Anche nei servizi di intelligenza artificiale non è da meno in quanto con il progetto Azure AI, creato assieme anche a Mistral (start-up francese competitor di OpenAI nella quale Microsoft ha deciso di investire ampliano il suo portfolio di modelli), è in grado, tramite la sua piattaforma di cloud computing, di permettere lo sviluppo e avere delle soluzioni basate sull'intelligenza artificiale personalizzate per ogni azienda.

Oltre a tutti i modelli presentati prima, anche come nel caso di Google con il suo modello Med PaLM 2 anche qui troviamo un progetto simile chiamato InnerEye, che ha lo scopo di migliorare le diagnosi e il trattamento delle malattie tramite un'analisi di immagini mediche. Questo è uno degli esempi più virtuosi che è stato sviluppato tramite la piattaforma Azure AI in collaborazione con Adaptive Biotechnologies e rappresenta l'ulteriore impegno di Microsoft nel promuovere l'innovazione tecnologica che abbia un impatto sulla società.

Forti investimenti in ricerca e sviluppo, collaborazioni strategiche e il forte focus sull'etica sono i tre pilastri su cui si basa il vantaggio competitivo rispetto al resto delle aziende del settore.

### 2.5.1 Performance di Megatron-Turing e Phi-3

Il modello Megatron-Turing NLG 530B (Microsoft, NVIDIA 2022), chiamato anche come MT-NLG, è un Large Language Model basato su architettura transformer e addestrato da Microsoft in collaborazione con NVIDIA, un'azienda statunitense famosa per lo sviluppo di processori grafici. Megatron-Turing ad oggi rappresenta il modello allenato sul maggior numero di parametri ovvero 530 miliardi, più di tre volte tanto quanto quelli utilizzati per addestrare Chat GPT-3. Questo modello presenta delle caratteristiche intrinseche presenti negli altri modelli che sviluppati e che vedremo poi. MT NLG si posiziona anche tra i modelli NLP (Natural Language Processing) ovvero quelli in grado di analizzare e comprendere il linguaggio umano, esattamente come quelli che abbiamo visto in precedenza. I benchmark presenti nel documento "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model" coprono una moltitudine di discipline che vanno dalla comprensione della lettura al ragionamento di buon senso e dalla previsione del completamento alla inferenza del linguaggio naturale. Di questi benchmark gli sviluppatori hanno deciso di testare MT NLG su contesti zero-shot, one-shot e few-shot. Alcuni di questi test vanno evidenziati come quello di "LAMBADA" che valuta la comprensione e la ritenzione delle informazioni nel contesto ampio. In questo caso Megatron-Turing raggiunge un livello più alto rispetto a Chat-GPT 3 e Gopher ovvero modelli, i quali Microsoft ha dato una ingente mano a sviluppare. Anche nel Benchmark "HellaSWAG" che valuta il buon senso in tutti e 3 i contesti (zero, one e few-shot) raggiunge un punteggio massimo di 82.42 che rispetto al resto della concorrenza attuale risulta essere un obiettivo per niente irraggiungibile. Infatti, già Chat-GPT 4 e LLaMA 2 raggiungono punteggi migliori.

Quando parliamo poi di Phi-3 ci riferiamo ad una famiglia di modelli (Microsoft, 2024) addestrata su un numero di dati molto più ristretto rispetto a Megatron-Turing che rappresenta totalmente l'opposto. La famiglia di Phi-3 si differenzia in 4 modelli diversi:

1. Phi-3-mini: il modello più piccolo, allenato su 3.8 miliardi di parametri che raggiunge punteggi notevoli nel Benchmark MMLU riuscendo ad arrivare ad un punteggio di 69/100 degno di essere comparato a Chat GPT-3.5 che raggiunge 71.4/100. Anche nel MT-bench riesce a ottenere un punteggio di 8.38, molto simile a quello di Chat GPT-3.5 Turbo che arriva a 8.4. L'unicità di questo modello risiede nella sua capacità di poter essere eseguito localmente da smartphone raggiungendo tali performance che gli permettono, su test fatti su iPhone con un processore A16 Bionic, di generare ben più di 12 token al secondo senza l'utilizzo di tecnologie di cloud computing che utilizzano altri competitor.
2. Phi-3-small: un modello relativamente piccolo ma più grande di Phi-3-mini addestrato su circa 7 miliardi di parametri. Phi-3-small raggiunge livelli significativamente più alti



di Phi-3-mini arrivando a un punteggio di 75/100 nel benchmark MMLU e un 8.7 nel MT-bench superando questa volta il modello di OpenAI.

- Phi-3-medium: questo rappresenta il modello più grande della famiglia addestrato su un numero di parametri che attiva circa ai 14 miliardi, il doppio di phi-2-small. Anche in questo caso nei benchmark riesce a migliorarsi rispetto ai fratelli minori raggiungendo un punteggio di 78 su 100 nel MMLU e un 8.9 nel MT-bench.

	Phi-3-mini 3.8b	Phi-3-small 7b	Phi-3-medium 14b	Phi-2 2.7b	Mistral 7b	Gemma 7b	Llama-3-In 8b	Mixtral 8x7b	GPT-3.5 version 1106
MMLU (5-Shot) [HBK <sup>21</sup> ]	68.8	75.7	78.0	56.3	61.7	63.6	66.5	70.5	71.4
HellaSwag (5-Shot) [ZHB <sup>19</sup> ]	76.7	77.0	82.4	53.6	58.5	49.8	71.1	70.4	78.8
ANLI (7-Shot) [NWD <sup>20</sup> ]	52.8	58.1	55.8	42.5	47.1	48.7	57.3	55.2	58.1
GSM-8K (8-Shot; CoT) [CKB <sup>21</sup> ]	82.5	89.6	91.0	61.1	46.4	59.8	77.4	64.7	78.1
MedQA (2-Shot) [JPO <sup>20</sup> ]	53.8	65.4	69.9	40.9	50.0	49.6	60.5	62.2	63.4
AGIEval (0-Shot) [ZCG <sup>23</sup> ]	37.5	45.1	50.2	29.8	35.1	42.1	42.0	45.2	48.4
TriviaQA (5-Shot) [JCWZ17]	64.0	58.1	73.9	45.2	75.2	72.3	67.7	82.2	85.8
Arc-C (10-Shot) [CCE <sup>18</sup> ]	84.9	90.7	91.6	75.9	78.6	78.3	82.8	87.3	87.4
Arc-E (10-Shot) [CCE <sup>18</sup> ]	94.6	97.0	97.7	88.5	90.6	91.4	93.4	95.6	96.3
PIQA (5-Shot) [BZGC19]	84.2	86.9	87.9	60.2	77.7	78.1	75.7	86.0	86.6
SociQA (5-Shot) [BZGC19]	76.6	79.2	80.2	68.3	74.6	65.5	73.9	75.9	68.3
BigBench-Hard (3-Shot; CoT) [SRR <sup>22</sup> , SSS <sup>22</sup> ]	71.7	79.1	81.4	59.4	57.3	59.6	51.5	69.7	68.32
WinoGrande (5-Shot) [SLBBC19]	70.8	81.5	81.5	54.7	54.2	55.6	65.0	62.0	68.8
OpenBookQA (10-Shot) [MCRK18]	83.2	88.0	87.4	73.6	79.8	78.6	82.6	85.8	86.0
BoolQ (2-Shot) [CLC <sup>19</sup> ]	77.2	84.8	86.5	–	72.2	66.0	80.9	77.6	79.1
CommonSenseQA (10-Shot) [THLB19]	80.2	80.0	82.8	69.3	72.6	76.2	79.0	78.1	79.6
TruthfulQA (10-Shot; MC2) [LHE22]	65.0	70.2	75.1	–	53.0	52.1	63.2	60.1	85.8
HumanEval (0-Shot) [CTJ <sup>21</sup> ]	58.5	61.0	62.2	59.0	28.0	34.1	60.4	37.8	62.2
MBPP (3-Shot) [AON <sup>21</sup> ]	70.0	71.7	75.2	60.6	50.8	51.5	67.7	60.2	77.8
Average	71.2	75.7	78.5	–	61.2	61.7	69.4	69.8	74.3
GPQA (2-Shot; CoT) [RHS <sup>23</sup> ]	32.8	34.3	–	–	–	–	–	–	29.0
MT Bench (2 round ave.) [ZCS <sup>23</sup> ]	8.38	8.70	8.91	–	–	–	–	–	8.35

Fig.14: Una tabella comparativa che illustra i risultati ottenuti in diversi benchmark dai modelli phi-3-mini, phi-3-small e phi-3-medium rispetto al precedente modelli Phi-2 e LLM di dimensioni simili dei competitor

- Phi-3-vision: in questo caso stiamo parlando di un modello multimodale in grado quindi in grado di elaborare non solo testo ma anche immagini. Allenato su 4.2 miliardi di



parametri, questo combina phi-3-mini e un encoder di immagini chiamato CLIP-ViT. Le sue capacità di interpretazione visive e di testo gli permettono di raggiungere punteggi alti e competitivi rispetto alla concorrenza nella maggior parte dei benchmark elencati in figura 15.

	Phi-3-Vision 4.2b	MM1-3B-Chat 3.6b [MGF <sup>+</sup> 24]	MM1-7B-Chat 7.6b [MGF <sup>+</sup> 24]	LLaVA-1.6 Vicuna-7b [LLLL23]	LLaVA-Next LLama3-8b [LLL <sup>+</sup> 24]	Qwen-VL-Chat 9.6b [BBY <sup>+</sup> 23]	Claude 3 haiku [Ant24]	Gemini 1.0 Pro V [TAB <sup>+</sup> 23]	GPT-4V-Turbo turbo-2024-04-09
MMMU (val) [YNZ <sup>+</sup> 23]	40.4	33.9	37.0	34.2	36.4	39.0	40.7	42.0	55.5
ScienceQA (test) [LMX <sup>+</sup> 22]	90.8	69.4	72.6	70.6	73.7	67.2	72.0	79.7	75.7
MathVista (testmini) [LBX <sup>+</sup> 24]	44.5	32.0	35.9	31.5	34.8	29.4	33.2	35.0	47.5
Inter-GPS (test) [LGJ <sup>+</sup> 21]	38.1	-	-	20.5	24.6	22.3	32.1	28.6	41.0
MMBench (dev-en) [LDZ <sup>+</sup> 24]	80.5	75.9	79.0	76.3	79.4	75.8	62.4	80.0	86.1
POPE (test) [LDZ <sup>+</sup> 23]	85.8	87.4	86.6	87.2	87.0	82.6	74.4	84.2	83.7
AI2D (test) [KSK <sup>+</sup> 16]	76.7	-	-	63.1	66.9	59.8	60.3	62.8	74.7
ChartQA (test) [MLT <sup>+</sup> 22]	81.4	-	-	55.0	65.8	50.9	59.3	58.0	62.3
TextVQA (test) [SNS <sup>+</sup> 19]	70.9	71.9	72.8	64.6	55.7	59.4	62.7	64.7	68.1

fig.15: Una tabella comparativa che illustra i risultati in diversi benchmark adatti ai modelli multimodali e che mette a confronto Phi-3-vision con altri LLM simili dei competitor

## 2.5.2 Limiti di Megatron-Turing e Phi-3

Nel periodo successivo all'uscita del paper di Megatron-Turing nel 2022, Microsoft e NVIDIA hanno continuato ad aggiornare costantemente la repository riguardante il progetto presente sulla piattaforma GitHub. Tutt'ora non esiste un'interfaccia per utilizzare questo modello che viene utilizzato principalmente da Microsoft per il loro servizio Azure AI. È facilmente intuibile, infatti, che un modello delle dimensioni come questo, abbia dei costi e delle necessità computazionali ingenti per un corretto funzionamento. Nonostante ciò, anche in questo caso il modello non è esente da bias cognitivi e la possibilità che sia capace di generare del contenuto potenzialmente dannoso per gli utenti.

I modelli Phi-3 che, come detto prima, rappresentano l'opposto invece riguardo a Megatron-Turing-NLG, in quanto fa parte dei modelli molto più piccoli e veloci che necessitano capacità computazionali alla portata di tutti per essere eseguiti direttamente in locale sulla propria macchina. Ciò comporta però delle limitazioni anche nell'utilizzo, che Microsoft tende a sottolineare e che vanno oltre la solita amplificazione dei bias e creazione di allucinazioni. Infatti, Phi-3-mini risulterà pure piccolo e veloce ma limitato molto nella comprensione di diverse lingue dove in gran parte della sua conoscenza è limitata all'inglese. Anche per quanto riguarda operazioni che richiedono una vasta conoscenza fattuale si notano diverse problematiche. Se si osserva bene nella tabella in figura 14, analizzando i risultati del benchmark TriviaQA (che misura appunto le conoscenze fattuali) Phi-3-mini come anche Phi-3-small raggiungono punteggi piuttosto bassi (con Phi-3-small che raggiunge un punteggio addirittura inferiore al modello più piccolo Phi-3-mini) che non si vedono poi nel modello Phi-3-medium. Tuttavia, questa debolezza potrebbe essere risolta nel modello Phi-3-mini concedendogli l'accesso a internet come mostrato nella figura 16. Ovviamente anche l'accesso alla rete porterebbe dei benefici, ma rappresenterebbe comunque un rischio nella contaminazione dei dati e amplificazione dei bias.

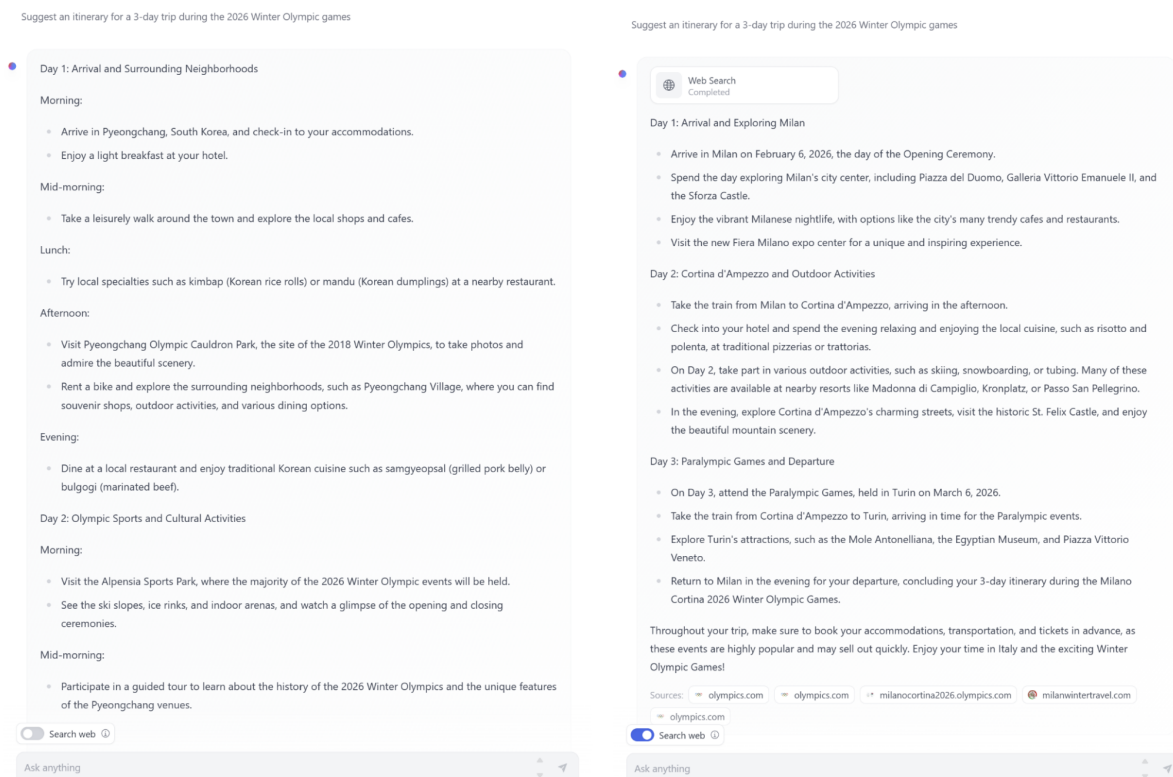


fig.16: Un esempio di miglioramento a delle risposte concedendo l'accesso a internet al modello Phi-3-mini, dove a sinistra si vede una risposta imprecisa sulle olimpiadi invernali del 2026 e a destra una risposta molto più completa e dettagliata ottenuta dando l'accesso a internet.

# Capitolo 3

## Comparazione Delle IA

### 3.1 Metodo comparativo

Spesso chi utilizza strumenti di intelligenza artificiale generativa li sceglie in base ai risultati ottenuti nei benchmark di cui le grandi aziende tanto si vantano. Seppure sia vero il fatto che certi modelli sono riusciti a stabilire nuovi standard per i futuri, e che alcuni siano nettamente migliori di altri per determinate mansioni, i test che vediamo non possono essere l'unico mezzo sul quale si può basare la nostra scelta. Infatti, i risultati di questi vengono presentati come una misura di una capacità molto "generale" delle intelligenze artificiali su determinati argomenti (Inioluwa 2021). Un problema sostanziale è che i compiti presenti all'interno dei benchmark non sono scelti in modo sistematico per rappresentare una gamma completa di capacità cognitive. Infatti, spesso rappresentano la convenienza o i particolari interessi dei creatori piuttosto che seguire una linea teorica. Ciò significa che seppure generalizzando molto i benchmark che si vedono spesso in giro sono presentati come obiettivi e neutrali quando in realtà presentano pregiudizi e prospettive dei loro creatori. La continua corsa alla ricerca delle prestazioni SOTA (State-Of-The-Art) incentrate sui benchmark è divenuta una normalità sulla ricerca dell'intelligenza artificiale generativa che porta alla costante caccia del miglioramento incrementale rispetto ai benchmark già esistenti. Anche se questo approccio porta con sé diversi progressi nello sviluppo di algoritmi sempre più grandi, ha anche creato diverse critiche. È innegabile che lo sviluppo di modelli di deep learning sempre più grandi abbia fatto passi da gigante, e abbia portato molta attenzione a modelli che richiedono un'alta concentrazione e intensità di risorse. Tutto ciò però a scapito di quelle che potrebbero essere soluzioni più funzionali e approcci più efficienti come modelli simili a Phi-2 e Claude3 Haiku più piccoli ma veloci e che non richiedono una spropositata quantità di potenza di calcolo. Per questo Inioluwa nella sua ricerca suggerisce di utilizzare dei metodi alternativi più completi e come lo sviluppo e la creazione sistematica di elementi di test come suite di test, audit e test contraddittori per valutare la potenziale presenza di pregiudizi del modello, oppure test di ablazione che aiutino a comprendere i contributi dei diversi componenti che formano il sistema in modo da isolarli nel caso dessero problemi per poi modificarli. Nel nostro caso, il nostro scopo è capire quanto, questi strumenti di intelligenza artificiale generativa, possano aiutare i lavoratori nel marketing, e in particolar modo i copywriter, a raggiungere il loro scopo, riducendo il tempo che questi impiegano nel loro processo di ideazione dei copy. Ciò che

andremo a vedere infatti sarà una comparazione tra i modelli che più di tutti si sono contraddistinti per le loro capacità nel mercato negli ultimi 2 anni per capire quale di questi si comporta meglio nel momento in cui si decide di creare un contenuto per una possibile ads su social media come Instagram. Per la creazione di questi testi e per fare in modo che i prompt che useremo per generarli siano il più efficaci possibili ci si deve avvalere a una delle skills che ha preso sempre più piede con la diffusione sempre più ampia di questi modelli di IA generativa ovvero il “prompt engineering” (Acar 2024). Per prompt engineering intendiamo una competenza che Acar definisce come “competenza del futuro” e che in alcune parti del mondo è già considerato un lavoro vero e proprio, che consiste nella corretta formulazione del prompt al fine di ottenere il risultato o output desiderato. Quello su cui dobbiamo basare infatti è la struttura o meglio la formulazione del problema da dare in pasto al modello per fare in modo che non si confonda o che generalizzi troppo rischiando di ottenere un risultato non voluto. In poche parole, dobbiamo essere in grado di parlare la sua stessa lingua. Per fare ciò il gruppo di Google Gemini ha recentemente fornito una guida al prompt engineering chiamata “Prompting guide 101: a quick-start handbook for effective prompts”, che ci permette di capire come strutturare un prompt per massimizzare la generazione di risposte e contenuti efficaci, con focus particolari per l’aiuto di diversi settori all’interno di una azienda come le risorse umane, i lavoratori nel marketing, i project manager o direttamente gli imprenditori. Questa guida, oltre a gettare delle basi, offre anche dei consigli per migliorare ulteriormente i propri prompt, aiutando l’utente a formulare di un livello più avanzato, approfondendo tecniche basate anche su un approccio ad un workflow, già noto ai prompt engineer, diviso in multi-step chiamato “chain of thought prompting”. Questo metodo permette di dividere un problema complesso in più piccole parti (quindi in più piccoli prompt), in modo da ridurre gli errori nella formulazione del problema e rendendo i prompt più semplici da interpretare per il modello. Quello che andremo a vedere nella ricerca sarà una comparazione su un singolo prompt, per vedere quale modello riesce a interpretare al meglio la richiesta e a generare un risultato migliore rispetto agli altri in una fase iniziale di formulazione di un contenuto testuale (copy) che dovrà adattarsi al formato di “storia” di Instagram. Nel nostro caso parliamo di generazione di contenuti in particolar modo testuali in quanto, allo stato dell’arte attuale, questo rappresenta il campo più avanzato e dove molti non riuscirebbero già a distinguere i contenuti scritti da un modello di intelligenza artificiale da un contenuto scritto da una persona fisica. I modelli che andremo a testare saranno quelli più comuni tra gli utilizzatori di questi strumenti, ovvero ChatGPT di Open AI, Gemini di Google e Claude di Anthropic. Per quanto riguarda le versioni di questi opteremo per le più comuni ovvero ChatGPT-3.5, Claude 3 Sonnet e Google Gemini 1.5.

Esploreremo inoltre la familiarità delle persone con questi strumenti per capire quanto spesso li utilizzano e per quale motivo, al fine di individuare anche quali siano i punti di forza e i limiti

e le difficoltà di questa nuova tecnologia. Un'ulteriore punto di interesse riguarda infatti le difficoltà che riscontrano le persone nell'esperienza di utilizzo dei diversi chatbot e la percezione che loro hanno dei diversi modelli.

## 3.2 Comparazione

Come detto in precedenza la comparazione sarà su un singolo prompt, strutturato secondo la guida sul prompting rilasciata da Google, che contiene consigli e linee guida molto simili a quelli che OpenAI ha messo a disposizione nel forum dedicato a ChatGPT. In questo caso decidiamo di testare la creazione di un copy (ovvero un testo) specificatamente per il formato "storia" di Instagram. In quanto più familiare alle persone si è preso per esempio un marchio già noto nel mercato globale, ovvero "Nike" e si è voluto creare un copy per una storia Instagram apposita per sponsorizzare un nuovo modello di sneakers che ha recentemente rilasciato. La generazione del copy sarà basata su un prompt uguale per tutti e 3 i modelli che compariamo e per verificare nella prima parte quello che si è descritto nel capitolo precedentemente per verificare quanto e quali strumenti di intelligenza artificiale generativa utilizzano più spesso le persone e nella seconda parte quanto il risultato generato sia chiaro, creativo e quanto crei interesse.

Il prompt che andremo ad utilizzare sarà il seguente:

*"Sei il copywriter di Nike e devi scrivere un copy per il lancio della nuova versione della vostra famosa scarpa "killshot 2". Il copy verrà utilizzato per le ads in formato "storia Instagram" e deve essere in linea col tono di voce di Nike oltre a contenere una call to action che spinga l'utente ad una azione".*

I risultati generati da questo sono i successivi che troveremo poi nelle figure 17 per quanto riguarda ChatGPT-3.5, 18 per quanto riguarda Google Gemini 1.5 e 19 per quanto riguarda Claude-3 Sonnet:

- **ChatGPT-3.5:** *"La leggenda è tornata. Semplicità iconica. Stile senza tempo. La nuova Nike Killshot 2 è qui. Pronta per ogni occasione, perfetta per chi sa che ogni passo conta. Rendi ogni giorno straordinario. Scopri l'eleganza delle linee pulite e la versatilità di una scarpa che si adatta a tutto. Dal campo alla città, la Killshot 2 è la tua nuova compagna di avventure. Sei pronto a lasciare il segno? Swipe up e acquista ora! #NikeKillshot2 #MakeYourMark #JustDoIt"*
- **Google Gemini 1.5:** *"Nike Killshot 2: Un'icona torna in pista. Classico. Versatile. Intramontabile. La leggendaria Nike Killshot 2 è tornata, pronta a conquistare le strade con il suo stile senza tempo. Tomaia in pelle premium, design minimalista e suola in gomma per un comfort e una trazione ineguagliabili. Sfida il tempo con la Killshot 2."*

*abbinala ai tuoi jeans preferiti o a un abito elegante per un look sempre impeccabile. Scopri di più: #Killshot2 #NikeSportswear*

- **Claude 3 Sonnet:** *“KILLSHOT 2: LA RINASCITA DI UN’ICONA. La sneaker leggendaria che ha fatto la storia torna con un design esplosivo. Nike Killshot 2 si rinnova per un nuovo capitolo di stile senza compromessi. L’upper in pelle e mesh traspirante avvolge il piede con una calzata avvolgente. La suola oversize in gomma massimizza trazione e resistenza. Un’estetica iconica che fonde heritage vintage e innovazione moderna. Sii tra i primi a calzare la nuova leggenda: acquista Killshot 2 ora con il link in bio. #NikeKillshot2 #EmbraceTheNext”*

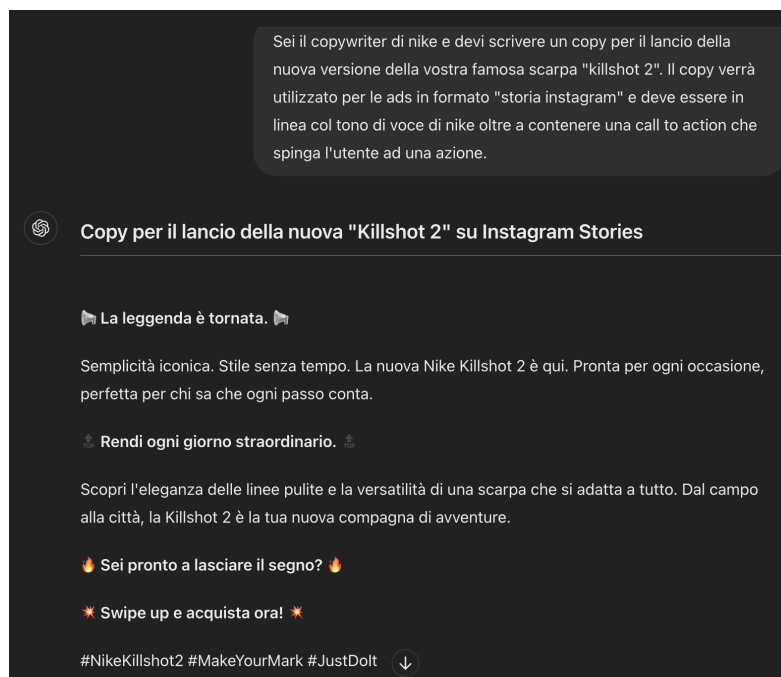


Fig.17: Un copy generato dal modello ChatGPT-3.5 in risposta al prompt strutturato in base alla guida al prompt engineering fornita da Google e che sarà utilizzato per la creazione della storia Instagram che andremo successivamente a testare.

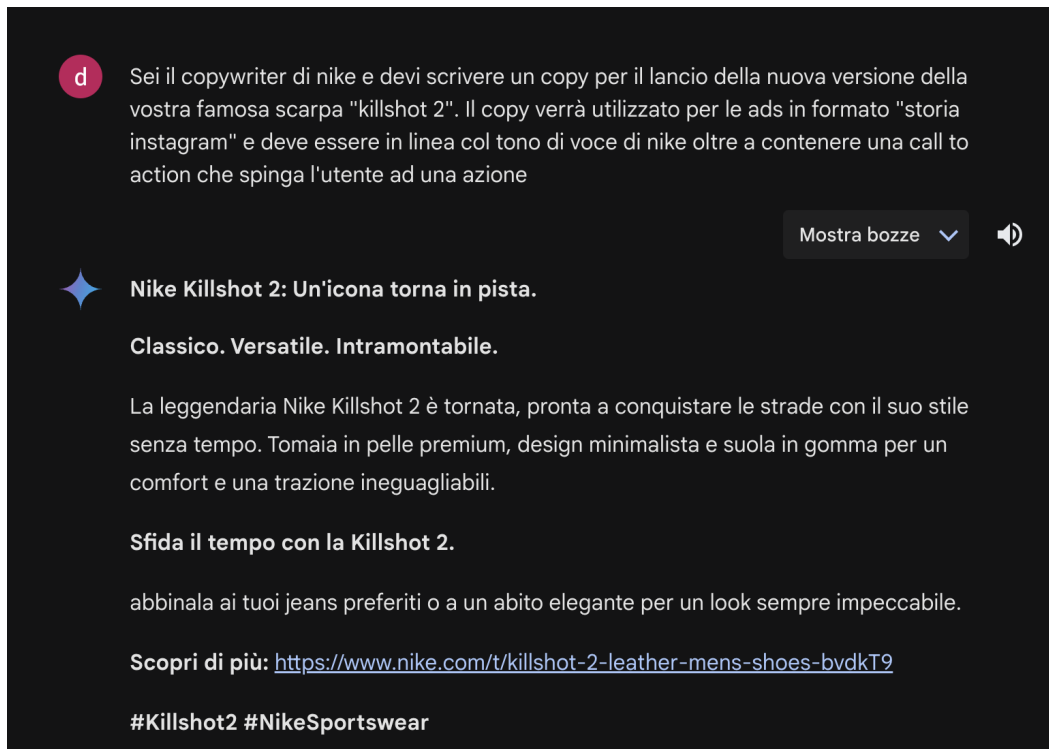


Fig.18: Un copy generato dal modello Gemini 1.5 in risposta al prompt strutturato in base alla guida al prompt engineering fornita da Google e che sarà utilizzato per la creazione della storia Instagram che andremo successivamente a testare.

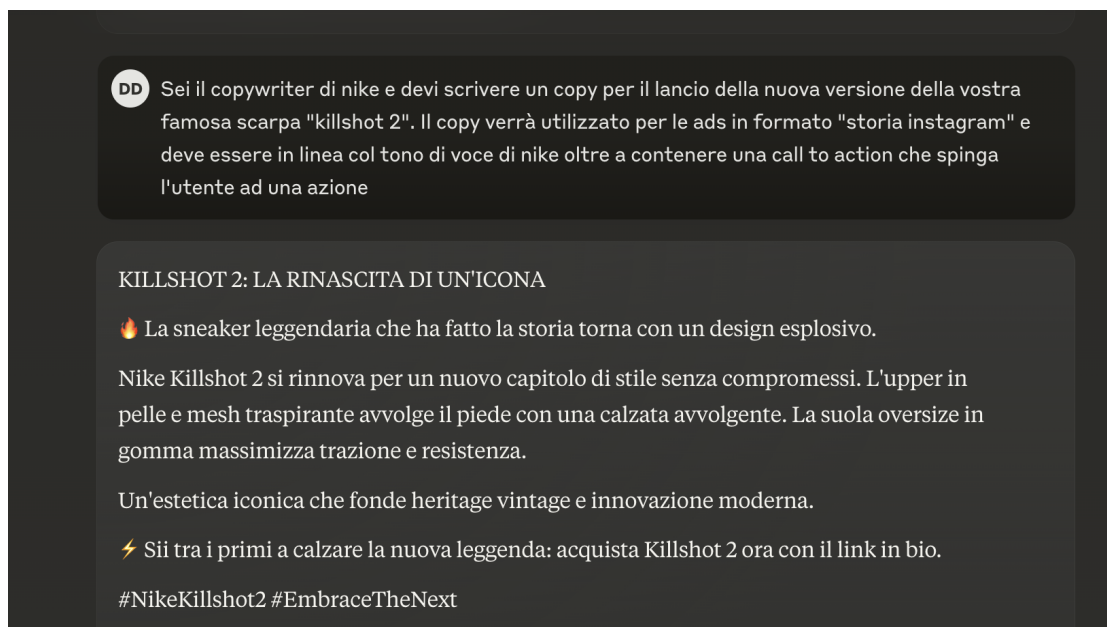


Fig.19: Un copy generato dal modello Claude-3 Sonnet in risposta al prompt strutturato in base alla guida al prompt engineering fornita da Google e che sarà utilizzato per la creazione della storia Instagram che andremo successivamente a testare.



Da queste risposte capiamo per prima cosa che il contesto da parte dei modelli non è stato interpretato nel migliore dei modi, in quanto i copy generati sono piuttosto lunghi e perciò meno adatti ai formati come storie Instagram, ma rappresenta comunque una buona base di partenza per la creazione di un buon contenuto potenzialmente più efficace. Le risposte ottenute sono state utili per la creazione, tramite Figma (uno strumento per prototipazione e design di interfacce e contenuti web), di 3 diverse storie (una per ogni modello che vogliamo analizzare) tramite Figma con un risultato finale rappresentato nella figura 20 che presenta una sequenza di 3 immagini corrispondono rispettivamente ai risultati ottenuti da ChatGPT, Google Gemini e Claude-3 Sonnet.



Fig.20: Il risultato della creazione delle storie partendo dai copy generati precedentemente realizzato tramite Figma. In ordine partendo da sinistra si vede il copy generato da ChatGpt-3.5, nell'immagine centrale quello generato da Google Gemini 1.5 e infine sulla destra quello generato da Claude-3 Sonnet.

### 3.3 Risultati

Dalla diffusione del questionario si sono raccolte un totale di 104 risposte di persone con un'età che va per il dai 18 ai 40 anni. In questo ultimo campo 62,5% ha un'età che va dai 18 ai 25 anni (molti dei quali studenti), il 30,8% dai 26 ai 32 anni e un 5,8% dai 33 ai 40 anni, suddiviso al 50% di genere maschile e 50% di genere femminile. Dell'intero campione il 57,7% afferma di utilizzare strumenti di intelligenza artificiale generativa a cadenza quotidiana o settimanale, il 20,2% mensilmente e il restante 22,1% li utilizza raramente o mai. I motivi che



spingono maggiormente le persone ad utilizzare questi strumenti riguardano maggiormente l'assistenza personale (scelta dal 78,6% delle persone), la traduzione linguistica (scelta dal 53,1%), la creazione di contenuti (scelta dal 38,2%) e l'apprendimento (scelta dal 35,7%). Questa ultima voce è interessante in quanto evidenzia un nuovo modo di apprendere tramite una nuova interazione che gli studenti potrebbero in futuro adottare sempre di più a supporto del loro studio. I principali benefici dell'intelligenza artificiale generativa, condivisi da maggior parte delle persone, riguardano il miglioramento dell'efficienza lavorativa, l'innovazione tecnologica e miglioramento dei servizi. Un dato che sembra essere più interessante è quello invece riguardante le problematiche riscontrate nell'utilizzo di questi software. In questo caso si vede che indipendentemente dai diversi livelli di conoscenza nel campo IA i problemi più frequenti risultano essere sempre gli stessi per quasi tutti. Quello più riscontrato riguarda principalmente la mancanza di controllo sui risultati generati, ricorrente principalmente dalle persone che si ritengono poco esperte in IA generativa ma comunque presente anche nel sottoinsieme del campione che si ritiene più esperto anche e in minor frequenza. Questo stesso discorso si verifica anche nel caso della qualità variabile dei contenuti che dove il 62,5% dell'intero campione lamenta il fatto che la qualità di questi non sia sempre la migliore e che vediamo scendere, ma essere sempre presente, da parte delle persone che si ritengono più esperte. Il terzo principale problema che si riscontra più spesso (per il 36,5% dei casi) è quello riguardante l'affidabilità e la precisione dei dati forniti ricevuti come output dal modello. Il modello più utilizzato, indifferentemente dalla conoscenza del campo delle intelligenze artificiali generative e da quanto le persone si sentissero più o meno esperte, rimane, come c'era da aspettarsi, Chat-GPT seguito dopo da Google Gemini e Microsoft Copilot lasciando indietro Meta AI e Claude che rimangono fuori dal podio. Questo è dato anche dal fatto che Meta AI non è attualmente disponibile in Italia se non tramite l'utilizzo di ulteriori servizi come VPN (Virtual Private Network ovvero un servizio fornito da terzi che ci permette di criptare la nostra navigazione di internet tramite una maschera del nostro indirizzo IP e della posizione) oppure aggregatori di intelligenze artificiali (ovvero delle piattaforme che permettono di utilizzare diversi modelli AI e che ci consentono di sfruttare le loro capacità collettive per generare risultati migliori rispetto all'utilizzo di una singola intelligenza artificiale generativa). Per quanto riguarda Claude invece è un modello che è stato rilasciato solo nel recente nel nostro paese e con un lancio sostenuto da una comunicazione praticamente assente. Anche per questi motivi è stato utilizzato più spesso solo dalle persone più interessate a questa nicchia e che si sono ritenuti maggiormente esperti del settore. Un dato interessante da approfondire lo si trova nella valutazione che le persone hanno assegnato riguardo all'efficacia dei modelli, in particolar modo se compariamo le persone più o meno esperte. Infatti, su una scala Likert con un intervallo che va da 1 a 5 dove, 1 rappresenta un livello di conoscenza in ambito intelligenza artificiale basso e 5 un livello di conoscenza alto, il 52,9% si ritiene poco

esperto assegnando un punteggio di 1 o 2, il 36,5% ritiene di saperne di più e si ritiene nella media assegnando un punteggio di 3 e il restante 12,5% ritiene di saperne, invece, molto o di essere un esperto assegnando un valore di 4 o 5. Questi dati diventano interessanti nel momento in cui andiamo a compararli con quelli raccolti sulla valutazione dell'efficacia dei diversi modelli. È stato chiesto al campione di assegnare un voto all'efficacia dei diversi tool sempre su una scala Likert simile a quella precedentemente descritta dove 1 rappresenta un livello di efficacia basso e 5 un livello di efficacia del modello AI alto. In questo caso Chat-GPT risulta essere la scelta migliore per gran parte delle persone sia più esperti che meno esperti. In secondo posto arriva Google Gemini che, come prima, indipendentemente dal livello di conoscenza del campo AI rimane nella stessa posizione. Ciò che invece cambia è la terza posizione contesa tra Microsoft Copilot e Claude dove, se guardiamo i dati nel complesso, in primo luogo troviamo allo scalino più basso del podio il modello di Microsoft con un leggero distacco su quello di Anthropic. Andando ad analizzare in maniera più approfondita questa posizione, dividendo le preferenze del pubblico meno esperto (ovvero coloro che hanno assegnato un punteggio di 1 o 2 riguardante la loro conoscenza in campo dell'intelligenza artificiale) influisce maggiormente, assieme al pubblico mediamente esperto (coloro che hanno assegnato un punteggio uguale a 3 riguardante la loro conoscenza in campo intelligenza artificiale), nel posizionamento di Microsoft Copilot che risulta essere preferito rispetto a Claude che sembra inizialmente essere lasciato fuori dal podio. La situazione invece cambia nel momento in cui si vuole prendere in considerazione la parte del campione più esperta che preferisce il modello di Anthropic su Microsoft Copilot. Un'ulteriore cosa che notiamo da questa analisi, fatta sommando i valori dati dagli utenti sull'efficacia di questi strumenti, è che più le persone si ritengono esperte più la differenza tra le preferenze dell'efficacia dei modelli si assottiglia e nel caso di Google Gemini e Claude diventa pressoché minima. Su questi risultati va fatta una precisazione: la gran parte delle risposte ricevute da parte dal pubblico meno esperto e mediamente esperto sostiene di non aver mai utilizzato Claude. Per questo motivo questa parte del campione non ha un'opinione a riguardo questo tool e perciò questa non risulta avere una rilevanza così tale quanto quella delle persone che lo hanno utilizzato e che sono ad un livello di conoscenza più avanzato ed esperto.

Nella successiva sezione parleremo invece della comparazione vera e propria dei testi utilizzati per la creazione di ads in formato "storia" di Instagram visti in precedenza e generati tramite l'utilizzo dei tre principali modelli che abbiamo citato e descritto nei precedenti capitoli. Questi testi sono stati presentati in maniera anonima, in modo che nessuno potesse sapere quali fossero generati quale modello e per evitare il condizionamento di espressione delle proprie preferenze nel momento di valutazione di questi copy.

I dati raccolti sulla percezione delle persone riguardo a questi tre diversi testi, ha generato dei risultati piuttosto simili tra di loro, ma che possono comunque essere utili per decretare il

modello che più si avvicina, con un singolo prompt, al risultato migliore. In primo luogo, si è voluto capire la chiarezza dei testi scritti con una scala Likert da un intervallo che va da 1 a 5, dove 1 rappresenta un risultato per niente o per nulla chiaro e 5 rappresenta un risultato molto chiaro. Tutti i testi generati hanno ricevuto dei punteggi simili ma con qualche differenza. Se si guarda in maniera più approfondita i risultati di questa domanda, si riesce a decretare già una prima classifica di che modello ha generato un testo ritenuto più chiaro rispetto ad altri. Per farlo utilizzeremo un approccio simile a quello precedentemente usato nella valutazione delle intelligenze artificiali generative, dove si erano divisi per gruppi di esperti, mediamente esperti e meno esperti per poi andare a sommare le loro valutazioni su ciascun modello (es: se un modello riceve 4 valutazioni che risultano essere rispettivamente 4, 5, 4, 3 il suo punteggio sarà la somma di questi risultando essere  $4+5+4+3=16$ ). Questa volta però lo faremo in maniera aggregata per cui non faremo distinzioni tra persone che siano più o meno esperte in questo campo. Partendo dalla terza posizione (ovvero la più bassa) troviamo il testo generato dal modello Claude di Anthropic che raggiunge un punteggio di 411 ottenendo un numero di voti tra il 4 e il 5 che si quasi equivale (è stato valutato il 34,6% delle volte come 4 e 40,4% delle volte come 5) per una media totale dei voti pari all'incirca a 3,97 su un complessivo di 5. Nella seconda posizione invece si trova Google con Gemini che totalizza un punteggio di 435 e dove il distacco tra le differenti valutazioni inizia a farsi più netto e distinguibile. Gemini infatti è riuscita a ottenere il punteggio massimo (ovvero 5) per il 44,2% del campione e un punteggio di 4 per il per il 36,5% arrivando ad una media valutazioni di 4,19. In prima posizione troviamo il modello Chat-GPT di OpenAI che arriva al punteggio totale di 460 ottenendo, di fatto, il distacco maggiore tra le differenti valutazioni. Ben 62 votanti su 104, ovvero il 59,6% (che corrispondono a più della metà) del campione gli conferisce un punteggio di 5 su 5 nella chiarezza del testo e solo 29 dei 104 votanti (che corrisponde invece al 27,8% del campione) hanno espresso una valutazione di 4 su 5, per una media voti che raggiunge il 4,44 su 5.

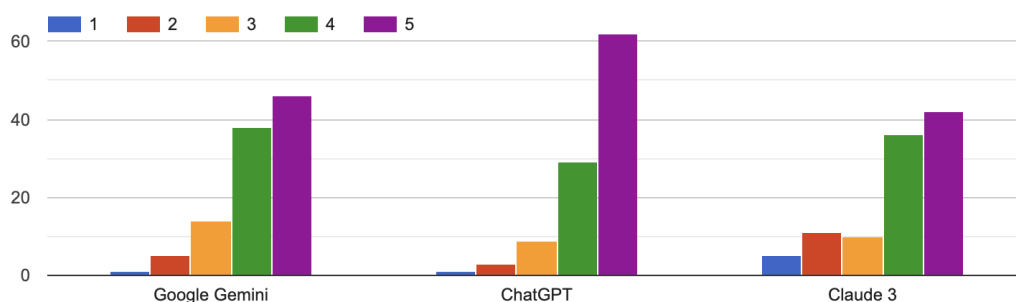


Fig.21: Un grafico che illustra i punteggi ottenuti dai tre modelli sulla valutazione della chiarezza del testo su una scala Likert a con un intervallo da 1 (poco chiaro) a 5 (molto chiaro)

Da questi risultati capiamo che i modelli analizzati riescono senza problemi nella generazione di testi chiari (anche perché hanno subito un processo di fine-tuning apposito a questa tipologia di compiti) senza troppi problemi. Già da questa primo sguardo possiamo definire una prima classifica dei modelli che vede in cima e come vincitrice ChatGPT e il discorso non cambia moltissimo nel momento in cui si vuole analizzare anche la sezione successiva, dove si chiede al campione di valutare, sempre su una scala Likert con un intervallo da 1 (poco creativo) a 5 (molto creativo), il livello di creatività del copy presente nelle immagini viste. Partendo sempre dal gradino più basso del podio troviamo, come nel caso precedente, Claude 3 che totalizza un punteggio di 326. Questo tra tutti risulta essere il meno creativo e che ha raccolto il maggior numero di valutazioni basse (il 19,2% del campione lo ha valutato con un punteggio di 1 o 2) e mediocri (il 53% del campione lo ha valutato con un punteggio pari a 3). Solo il 27.8% valuta, come punteggio di 4 o superiore, il testo generato dal modello di Anthropic come molto creativo. Se si vuole fare una media dei voti ricevuti come per la chiarezza del testo Claude 3 arriva a 3,14 su 5. La situazione non cambia di molto anche nel momento in cui si si sposta al secondo posto, dove troviamo sempre Google Gemini, che supera di un nonnulla Claude. Gemini arriva di una decina di punti davanti al terzo posto, dove la somma delle valutazioni raggiunge un punteggio totale di 337. Anche in questo caso si vede un elevato numero di valutazioni basse (il 14,4% del campione ha assegnato una valutazione pari 1 o 2 riguardante la creatività del contenuto scritto) e mediocri (il 51% del campione ha assegnato una valutazione pari a 3 riguardante la creatività del contenuto scritto) e una cifra non di molto più elevata valutazioni alte (il 34,6% del campione ha assegnato una valutazione pari a 4 o superiore alla creatività del contenuto scritto). Non distaccandosi troppo dal gradino inferiore del podio anche la media rimane simile e raggiunge il 3,22 su 5. Ciò porta al primo posto con la valutazione più alta nuovamente ChatGPT che con un solido punteggio di 358 mantiene la sua posizione davanti a Google Gemini e Claude 3. In questo caso meno di un decimo delle persone che hanno espresso una valutazione gli assegnano un valore basso (il 9,6% del campione ha assegnato una valutazione pari 1 o 2 riguardante la creatività del contenuto scritto). Come al solito però rimane una maggioranza di valutazioni mediocri che toccano la cifra di 49 su 104 (il 47,1% del campione ha quindi assegnato una valutazione pari a 3 riguardante la creatività del contenuto scritto), che questa volta si avvicina molto quella delle valutazioni alte che arriva a 45 su 104 (il 43,3% del campione ha quindi assegnato una valutazione pari a 4 o superiore alla creatività del contenuto scritto). Con una media di 3,44 su 5, ChatGPT conferma nuovamente la sua posizione al primo posto. Da questa ultima analisi possiamo già intravedere una discesa importante delle medie e dei punteggi ottenuti da ciascuna delle intelligenze artificiali viste. Claude 3 di Anthropic, per esempio cala da una media ottenuta sulla chiarezza del testo di 3,97 ad una media di 3,14 (con un punteggio totale dei voti di 411 nella chiarezza del contenuto e di 326 per la creatività), Google Gemini passa

da una media di 4,18 a una di 3,22 (con un punteggio totale dei voti di 435 nella chiarezza del contenuto e di 337 per la creatività) e ChatGPT da 4,44 a 3,48 (con un punteggio totale dei voti di 460 nella chiarezza del contenuto e di 358 per la creatività). La creatività, quindi, non risulta essere un punto forte di questa tecnologia, che molti appunto valutano o percepiscono come una sua caratteristica per lo più mediocre e poco eccelsa. Vi è comunque la possibilità di creare dei testi più specifici tramite l'approccio di chain of thoughts prompting descritto precedentemente, che potrebbe darci più possibilità di generare un copy più creativi e maggiormente desiderati aumentando la probabilità di successo di questi.

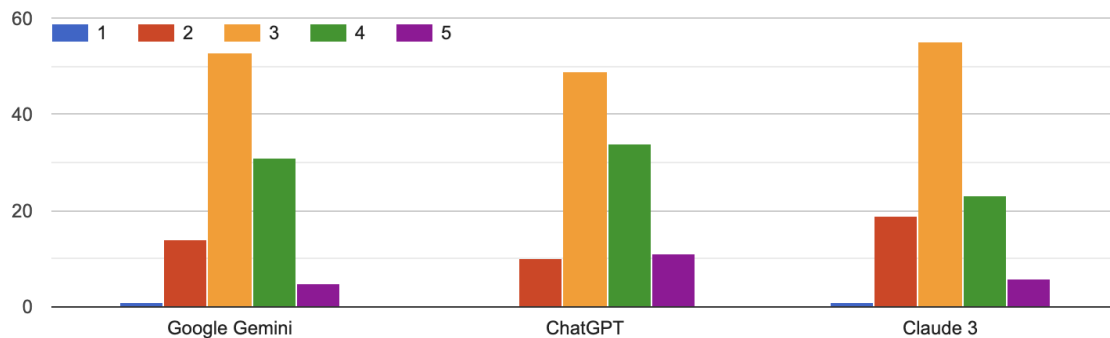


Fig.22: Un grafico che illustra i punteggi ottenuti dai tre modelli sulla valutazione della creatività del testo su una scala Likert a con un intervallo da 1 (poco creativo) a 5 (molto creativo)

Un ultimo aspetto che si vuole analizzare è l'engagement che questi testi possono generare. Anche in questo caso si misura il grado di interesse su una scala Likert con un intervallo tra 1 e 5 dove 1 rappresenta una bassa propensione all'informarsi maggiormente sul prodotto mentre 5 rappresenta una alta propensione all'ottenere più informazioni riguardo ciò che si ha appena letto. Partendo sempre dal gradino più basso e andando a salire troviamo come di consueto il modello di Anthropic che sommando i punteggi totali arriva ad un massimo di 313, con una maggioranza di valutazioni pari a 3 che arriva al 47,1% del totale (49 persone su 104). Il 21,1% (22 persone su 104) invece si dice poco interessato all'approfondimento del prodotto dopo aver letto il copy e ha assegnato una valutazione pari o inferiore a 2. Il 31,7% (33 persone su 104) del campione, invece, si dichiara propenso all'informarsi maggiormente su tale articolo dopo la lettura dei testi, assegnando un punteggio pari o superiore a 4. La media che ne scaturisce da questi punteggi per Claude 3 risulta così essere 3 su 5 per quanto riguarda il coinvolgimento. Anche al secondo posto la situazione non cambia ritrovando sempre Google Gemini che, come nel caso scorso, arriva al punteggio di 337. Questa volta però la maggioranza di valutazioni si trova nella parte più alta dove il 48,1%, infatti, ha espresso di essere più propenso ad informarsi maggiormente sul prodotto (50 persone su 104 hanno espresso una valutazione di 4 o superiore). Il 31,7% delle persone invece dopo aver

letto i testi valuta la propria propensione a informarsi maggiormente sul prodotto con una valutazione pari a 3 (33 persone su 104). Il restante 20,2% afferma di non essere ulteriormente interessato lasciando una valutazione uguale a 2 o inferiore (21 su 104). Ciò ci porta alla media totale dei voti che, come prevedibile, risulta essere uguale alla domanda precedente riguardante la creatività. Con 3.24 su un totale di 5 l'intelligenza artificiale generativa di Google rimane in seconda posizione. ChatGPT rimane al primo posto del podio con una differenza di 21 punti e un punteggio di 358 ottenendo il maggior numero di valutazioni massime ovvero 13 (ovvero il 12,5% del campione, che ha espresso una valutazione uguale a 5). Più della metà del campione, ovvero, il 55,8% (58 persone su 104) ha espresso una valutazione positiva e si reputa interessata a ricevere più informazioni sul prodotto assegnando un voto maggiore o uguale a 4. Il 25% (27 persone su 104) si reputa invece indifferente e ha assegnato un punteggio pari a 3. Chi invece boccia il testo di ChatGPT è il 18,2% del campione (19 persone su 104) che non si reputa coinvolto emotivamente e che ha manifestato poca propensione e interesse nell'informarsi maggiormente riguardo il prodotto. In media l'IA di OpenAI totalizza un 3,24 su 5 che gli basta per prendere il posto superiore del podio anche in questa occasione. Come nella creatività anche nell'engagement non si sono ottenuti punteggi interessanti e degni di una qualsiasi nota. La cosa che possiamo però analizzare è che questi non sono stati punteggi negativi ma nella media e che si possono perfezionare già con l'aiuto umano e qualche prompt in più per indirizzare meglio l'algoritmo al fine di massimizzare le probabilità di ottenere il risultato desiderato.

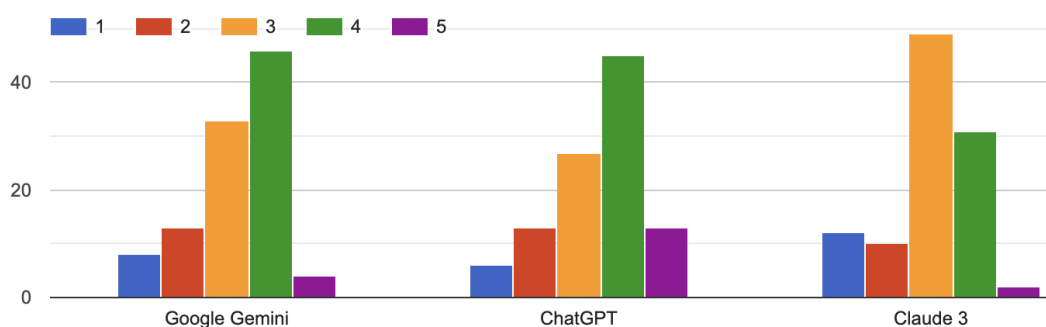


Fig.22: Un grafico che illustra i risultati ottenuti dai tre modelli sul livello di propensione del campione all'informarsi maggiormente dopo aver letto i testi, misurato su una scala Likert a con un intervallo da 1 (poco propenso) a 5 (molto propenso)

Rimane un'ultima domanda da esaminare ovvero quella nel quale si chiede agli utenti, quali dei testi appena letti fosse stato scritto da un'intelligenza artificiale generativa, per decretare il modello in grado di avvicinarsi maggiormente a un lavoro scritto da un umano e se chi utilizza maggiormente strumenti di intelligenza artificiale generativa riesce a distinguere un testo scritto dall'IA o meno. Se si vogliono guardare i dati in maniera aggregata e senza distinzioni di alcun tipo, il 32,7% del campione (ovvero 34 persone su 104) vede il post di Claude 3 il più

vicino a quello scritto da una IA. Il 29,8% (ovvero 31 persone su 104) del campione invece ritiene che tutti e tre i testi siano stati scritti tramite strumenti di intelligenza artificiale. Il testo di Google Gemini invece è stato votato dal 17,3% del campione (ovvero 18 persone su 104) mentre il 14,4% (ovvero 15 persone su 104) ha optato per il copy generato da ChatGPT. Il rimanente 5,8% (ovvero 6 persone su 104) reputa che nessuno dei tre testi letti sia stato generato da una IA.

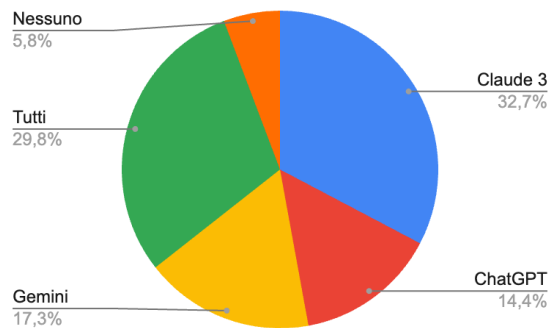


Fig.23: Un grafico a torta che illustra i risultati ottenuti al momento in cui si chiedeva, ai partecipanti del questionario, quale dei testi fosse stato scritto da una IA e dove il 32,7% del campione dichiara che il testo di Claude 3 è quello più sospetto ad esser creato con determinati strumenti.

Una cosa interessante, emersa dall'analisi invece delle risposte ricevute da chi si ritiene più esperto e utilizza più spesso questa tecnologia, è il fatto che più persone si siano rese conto che tutti i testi sono stati scritti tramite l'utilizzo dell'intelligenza artificiale. Si vuole prendere in considerazione particolarmente la parte di campione che ha espresso una valutazione di tre o superiore nel momento in cui gli è stato chiesto di esprimere il suo livello di conoscenza di certi strumenti. In primo luogo, questo sottoinsieme più esperto del campione totale esprime per il 35,5% la sua preferenza per l'opzione che indica che tutti i testi sono stati scritti da una IA superando ma non di molto l'opzione che dava come creato da determinati strumenti solo il modello Claude 3 che viene votato dal 33,3% del campione con maggiore conoscenza in ambito IA. Per il resto la classifica e le valutazioni ottenute dagli altri due modelli non differiscono di molto rispetto all'osservazione aggregata del totale del campione. Ciò che invece scende è la percentuale di persone che reputa che nessuno dei tre testi sia stato scritto da una intelligenza artificiale che risulta essere solo il 3,9% del sottoinsieme del campione più piccolo rispetto al 5,8% del campione più grande.



# Conclusioni

In conclusione, possiamo trarre una classifica generale dei modelli che sono stati più efficaci rispetto ad altri che vede il modello di Anthropic ossia Claude 3 in ultima posizione, Google Gemini in seconda e in prima quello rilasciato da OpenAI, ChatGPT il quale testo generato è risultato più chiaro, creativo e coinvolgente oltre che meno sospettoso di essere stato creato da uno strumento di intelligenza artificiale. Con questo studio non si vuole dare una classifica definitiva, e non si vogliono definire delle regole per la scelta del modello migliore da utilizzare, perché questo varia in base alle proprie esigenze e preferenze personali. Questo studio vuole mettere in evidenza la capacità di questi LLM di capire e interpretare i contenuti in cui gli chiediamo di agire e determinare quale, con un singolo prompt, riesce a generare la migliore risposta iniziale che servirà come base per un futuro sviluppo del copy finale. E questo a dirlo sono stati anche coloro che hanno risposto al questionario appena analizzato, che lasciando alcune risposte qualitative al fine di questo hanno espresso la loro idea sull'utilizzo dell'intelligenza artificiale generativa nel lavoro. Secondo loro il risultato ottenuto risulta essere una buona base di partenza ma troppo lungo per essere inserito in un formato come una storia di Instagram e necessitano sempre e comunque di un perfezionamento da parte di un umano. Alcuni invece scrivono che l'utilizzo di questa tecnologia, andrebbe formato e in modo da dare possibilità al maggior numero di persone di sfruttare al meglio tali strumenti e che questi non devono essere il sostituto della creatività e "mano" umana. Quello che si percepisce quando si parla di intelligenza artificiale nell'ultimo anno, è spesso una preoccupazione per quelli che ne saranno gli utilizzi e sviluppi futuri. Non è un segreto il fatto che molti esprimono la loro idea su quale sia il prossimo lavoro che scomparirà, passando da un estremo all'altro. C'è chi afferma che i primi lavoratori a dileguarsi saranno i giornalisti o i copywriter. Chi invece è convinto che i primi a sparire siano proprio i programmatori e gli sviluppatori. Queste ultime due frasi forse sembrano un ossimoro, e spesso sono associate ad altre affermazioni in modo da creare molto rumore attorno al tema intelligenza artificiale. Quello stesso tema che mette spesso in dubbio gli studenti che oggi stanno scegliendo la carriera da intraprendere un indomani. Qualcuno invece è più ottimista e vede l'avvento di questa nuova tecnologia come un'opportunità per far crescere il proprio business o, perché no, crearne uno di nuovo. Tengo a specificare che quello che segue è un elaborato personale, frutto dei miei pensieri riguardante il tema trattato da questo elaborato. Pensando ad ogni tipo di nuova tecnologia spesso all'inizio ci si è chiesti se potesse essere buona o meno per noi umani. Prendendo per esempio il nucleare, questo offre delle interessanti opportunità (come nell'utilizzo di questa tecnologia per la creazione di energia in maniera più efficiente) ma porta con sé anche delle minacce (come l'utilizzo di tale tecnologia in ambiti militari). Quando ci si avvicina ad una



nuova tecnologia si è spesso quindi davanti a entrambe le facce della medaglia, una buona e una cattiva, esattamente come nel caso di oggi dell'intelligenza artificiale. L'impatto che questa avrà sulla società in futuro sarà determinato in base a come la utilizzeremo. In definitiva per creare una migliore AI per il futuro dovremmo forse essere, per prima cosa, migliori noi stessi. L'intelligenza artificiale generativa è spesso monotona e non ha una personalità definita. Spetta a noi riuscire o meno a distinguerci per la nostra creatività e per le nostre idee, facendo emergere l'unicità che ognuno possiede.

# Bibliografia

- Acar, Oguz A., 25 aprile 2024, "Beyond Prompt Engineering: Skills Marketers Need to Deploy Generative AI Successfully"
  - <https://doi.org/10.2478/nimmir-2024-0003>
- Anthropic, "Model Card Evaluations for Claude Model"
  - <https://www-cdn.anthropic.com/files/4zrzovbb/website/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf>
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜". In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623.
  - <https://doi.org/10.1145/3442188.344592>
- Fui-Hoon Nah, Fiona, Ruilin Zheng, Jingyuan Cai, Keng Siau & Langtao Chen, 21 Luglio 2023 "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration, Journal of Information Technology Case and Application Research"
  - <https://doi.org/10.1080/15228053.2023.2233814>
- Gemini Team Google, Dicembre 2023, "Gemini: A Family of Highly Capable Multimodal Models"
  - [https://storage.googleapis.com/deepmind-media/gemini/gemini\\_1\\_report.pdf](https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf)
- Hendrycks, Dan, 12 Gennaio 2021, "Measuring Massive Multitask Language Understanding"
  - <https://doi.org/10.48550/arXiv.2009.03300>
- Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, Alex Hanna, 26 Novembre 2021, "AI and the Everything in the Whole Wide World Benchmark"
  - <https://doi.org/10.48550/arXiv.2111.15366>

- Kshetri, Nir, Yogesh K. Dwivedi, Thomas H. Davenport, Niki Panteli 17 Febbraio 2024, "Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda"
  - <https://doi.org/10.1016/j.ijinfomgt.2023.102716>
- Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al., 9 Febbraio 2023, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models."
  - <https://doi.org/10.1371/journal.pdig.0000198>
- Meta AI, 27 Febbraio 2023, "LLaMA: Open and Efficient Foundation Language Models"
  - <https://doi.org/10.48550/arXiv.2302.13971>
- Microsoft, Nvidia 4 Febbraio 2022, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model"
  - <https://doi.org/10.48550/arXiv.2201.11990>
- Microsoft, 23 Maggio 2024, "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone"
  - <https://doi.org/10.48550/arXiv.2404.14219>
- Open AI, 4 Marzo 2022, "Training language models to follow instructions with human feedback"
  - <https://doi.org/10.48550/arXiv.2203.02155>
- Open AI, 15 Marzo 2023, "GPT-4 Technical report"
  - <https://doi.org/10.48550/arXiv.2303.08774>
- Open AI, 20 Luglio 2017, "Proximal Policy Optimization (PPO)"
  - <https://doi.org/10.48550/arXiv.1707.06347>
- Parnami, Archit, Minwoo Lee, 7 Marzo 2022, "Learning from Few Examples: A Summary of Approaches to Few-Shot Learning"
  - <https://doi.org/10.48550/arXiv.2203.04291>
- Rane, Nitin, 1 Settembre 2023, "Role and Challenges of ChatGPT and Similar Generative Artificial Intelligence in Human Resource Management".
  - <http://dx.doi.org/10.2139/ssrn.4603230>
- Singhal, K., Azizi, S., Tu, T. et al., 12 luglio 2023, "Large language models encode clinical knowledge. Nature 620, 172–180 (2023)".
  - <https://doi.org/10.1038/s41586-023-06291-2>

- Sætra, Henrik Skaug, 14 Settembre 2023, "Generative AI: Here to stay, but for good?"
  - <https://doi.org/10.1016/j.techsoc.2023.102372>
- Ye, Ke, 24 Gennaio 2024, "SpacTor-T5: Pre-training T5 Models with Span Corruption and Replaced Token Detection"
  - <https://doi.org/10.48550/arXiv.2401.13160>

# Sitografia

- Chen, Michael, 6 Dicembre 2023, Oracle, "What is AI model Training & Why Is It Important"
  - <https://www.oracle.com/it/artificial-intelligence/ai-model-training/#types-of-ai-model-training>
- Content marketing Institute, "What is content marketing"
  - <https://contentmarketinginstitute.com/what-is-content-marketing/#:~:text=Content%20marketing%20is%20a%20strategic,to%20drive%20profitable%20customer%20action.>
- Dencheva V., Marzo 2023 "Share of marketers using generative artificial intelligence (AI) in their companies in the United States as of March 2023"
  - <https://www.statista.com/statistics/1388390/generative-ai-usage-marketing/>
- Google Cloud, 20 marzo 2024, "Punti di forza e limiti del modello Gemini"
  - <https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/strengths-limits?hl=it#:~:text=Limitazioni%20dei%20Gemini,-I%20modelli%20Gemini&text=Pu%C3%B2%20avere%20difficolt%C3%A0%20con%20una,nella%20comprensione%20delle%20immagini%20ruotate.&text=Pu%C3%B2%20fornire%20solo%20approssimazioni%20approssimative,soprattutto%20per%20gli%20oggetti%20oscurati>
- Oracle, 2022 "what is machine learning"
  - <https://www.oracle.com/artificial-intelligence/machine-learning/what-is-machine-learning/>
- Robertson, Adi, 21 Febbraio 2024, "Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis"
  - <https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical>
- Sarker, Iqbal H., 18 Agosto 2021 "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions"
  - <https://link.springer.com/article/10.1007/s42979-021-00815-1>