Alma Mater Studiorum · Università di Bologna

**SCUOLA DI SCIENZE**
**Corso di Laurea Magistrale in Informatica**

# Utilizzo di Tecniche RAG per la Valutazione e Comparazione dei Modelli LLM in ambito medico

**Relatore**
**Chiar.mo Prof.**
**Stefano Ferretti**

**Presentata da**
**Giuseppe Pio Salcuni**

**Correlatori**
**Prof.ssa Sara Montagna**
**Dott. Ing. Gianluca Aguzzi**

**Sessione II**
**Anno Accademico 2023/2024**

*Al Maresciallo Francesco Pastore.*
*Il tuo coraggio, la tua dedizione*
*e il tuo esempio resteranno per*
*sempre impressi nel mio cuore.*
*Con gratitudine e affetto,*
*ti dedico questo traguardo.*

# Indice

# Abstract

La presente tesi esplora l'applicazione della tecnica Retrieval-Augmented Generation (RAG) combinata con diversi modelli di linguaggio di grandi dimensioni (LLM), con particolare attenzione all'ambito medico. L'obiettivo principale è analizzare come la tecnica RAG possa migliorare la qualità delle risposte generate, in particolare nell'identificazione di pattern clinici e nella personalizzazione delle strategie terapeutiche. Viene fornita una panoramica sull'evoluzione dei LLM, con un focus sulle applicazioni in ambito sanitario. Inoltre, la tesi valuta le performance dei modelli tramite la libreria RAGAS, analizzando metriche quali accuratezza, robustezza e fedeltà delle risposte in contesti clinici. Il lavoro culmina in un caso studio sulla gestione dell'ipertensione, utilizzando la tecnica RAG per migliorare la rilevazione e il monitoraggio continuo della pressione arteriosa. Si esplorano le potenzialità dell'IA nel superare le limitazioni dei metodi diagnostici tradizionali e nel supportare la personalizzazione delle terapie. I risultati evidenziano il potenziale dei LLM specializzati nell'ottimizzazione della gestione dei pazienti ipertesi, aprendo la strada a future applicazioni di IA nel settore sanitario, con impatti positivi sulla qualità della cura e l'efficienza operativa.

# Introduzione

Negli ultimi anni, l'Intelligenza Artificiale (IA) e, in particolare, i modelli di linguaggio di grandi dimensioni (LLM) hanno rivoluzionato numerosi settori, compreso quello sanitario. I LLM, grazie alla loro capacità di analizzare e generare testo in linguaggio naturale, hanno aperto nuove possibilità in ambiti quali la diagnosi automatizzata, la gestione delle informazioni cliniche e il supporto decisionale per i professionisti della salute. Tuttavia, nonostante i progressi tecnologici, l'utilizzo di questi modelli in ambito medico presenta ancora notevoli sfide, tra cui l'accuratezza delle risposte generate, la robustezza rispetto a scenari complessi e la necessità di affrontare questioni etiche legate alla privacy e alla sicurezza dei dati.

In questo contesto, la tecnica Retrieval-Augmented Generation (RAG) emerge come una promettente soluzione. La RAG combina la generazione di testo tipica dei LLM con il recupero di informazioni pertinenti da fonti esterne, consentendo ai modelli di fornire risposte non solo coerenti ma anche ancorate a dati specifici e aggiornati. Questa tecnica può mitigare alcune delle limitazioni intrinseche dei modelli di linguaggio, migliorando la qualità delle risposte generate in contesti complessi e dinamici come quello sanitario.

L'ipertensione rappresenta uno dei principali fattori di rischio per le malattie cardiovascolari, una delle principali cause di mortalità a livello globale. La diagnosi e la gestione efficace dell'ipertensione rimangono una sfida aperta, principalmente a causa della natura spesso asintomatica della condizione e delle limitazioni dei metodi diagnostici tradizionali, che possono essere influenzati da fattori esterni come l'ansia del paziente o il momento della misurazione. In questo scenario, l'integrazione dei LLM con tecnologie avanzate di IA, come il RAG, offre una soluzione innovativa per migliorare la diagnosi precoce, il monitoraggio continuo e la personalizzazione delle terapie.

Questa tesi si articola in tre principali direzioni di ricerca. In primo luogo, viene fornita una panoramica sull'evoluzione dei LLM, con particolare attenzione alle loro applicazioni nel settore sanitario, dove l'accesso rapido e preciso a informazioni cliniche aggiornate è cruciale per la gestione delle patologie croniche. Successivamente, viene analizzata in dettaglio la tecnica RAG, descrivendone l'architettura, i vantaggi e le sfide legate alla sua implementazione in contesti reali. Infine, viene presentata un'analisi comparativa delle performance di diversi modelli LLM nel contesto della gestione dell'ipertensione,

utilizzando la tecnica RAG per ottimizzare la rilevazione e il controllo della pressione arteriosa.

L'obiettivo principale di questo lavoro è dimostrare come l'integrazione tra LLM e RAG possa migliorare significativamente la qualità delle risposte cliniche, riducendo l'incidenza di errori diagnostici e migliorando l'efficacia complessiva del trattamento. Si intende, inoltre, evidenziare il potenziale di queste tecnologie nel rivoluzionare il modo in cui vengono gestite le patologie croniche, aprendo nuove prospettive per la medicina personalizzata e la sanità digitale. Il caso studio sull'ipertensione rappresenta un esempio concreto di come l'IA possa essere utilizzata per affrontare problematiche complesse e migliorare gli esiti clinici dei pazienti.

# 1 Introduzione alla GenAI e LLM

L'Intelligenza Artificiale Generativa (GenAI) si occupa della creazione di contenuti nuovi e originali attraverso l'uso di tecniche avanzate di machine learning. A differenza delle tradizionali applicazioni di intelligenza artificiale, progettate per compiti specifici come la classificazione delle immagini o il riconoscimento del parlato, i modelli generativi sono in grado di produrre asset completamente nuovi che spaziano dalla generazione di testi alla creazione di immagini, audio e video. Questi contenuti sono spesso così realistici e ben realizzati che è difficile, se non impossibile, distinguerli da quelli creati dall'essere umano. I modelli di GenAI si basano su tecniche di apprendimento automatico che utilizzano reti neurali profonde, capaci di analizzare e apprendere da enormi quantità di dati. Durante il processo di addestramento, questi modelli identificano e comprendono schemi e strutture nei dati di input. Questa fase di apprendimento consente ai modelli di rappresentare la distribuzione probabilistica sottostante dei dati stessi. Successivamente, possono generare nuovi esempi campionando da questa distribuzione, creando contenuti che riflettono le caratteristiche e le variabilità dei dati originali. L'utilità degli output generati dall'Intelligenza Artificiale Generativa dipende dalla qualità dei dati di addestramento, dall'architettura del modello, dalle metodologie di addestramento e dai prompt forniti dagli utenti. La qualità dei dati è cruciale perché modelli ben addestrati su dati diversi e completi possono comprendere e replicare meglio i pattern e le sfumature. Dati incoerenti o distorti possono portare a risultati problematici. Le metodologie di addestramento e le strategie di valutazione influiscono sulla capacità del modello di adattarsi e migliorare. Inoltre, l'architettura del modello deve essere bilanciata: un'architettura troppo semplice può non cogliere le sfumature contestuali, mentre una troppo complessa può sovrapporsi e ignorare pattern importanti.

## 1.1 LLM

Un ambito particolarmente rilevante dell'Intelligenza Artificiale Generativa è rappresentato dai Modelli di Linguaggio di Grandi Dimensioni (LLM). Questi modelli sono progettati per comprendere e generare testi in linguaggio naturale con una notevole capacità di coerenza e rilevanza. Gli LLM, come GPT (Generative Pre-trained Transformer) e altri

modelli basati su architetture simili, sono addestrati su vasti corpus di testi provenienti da diverse fonti. Questa vasta esposizione consente loro di apprendere schemi linguistici complessi e di replicare stili di scrittura, tonalità e contenuti variabili.

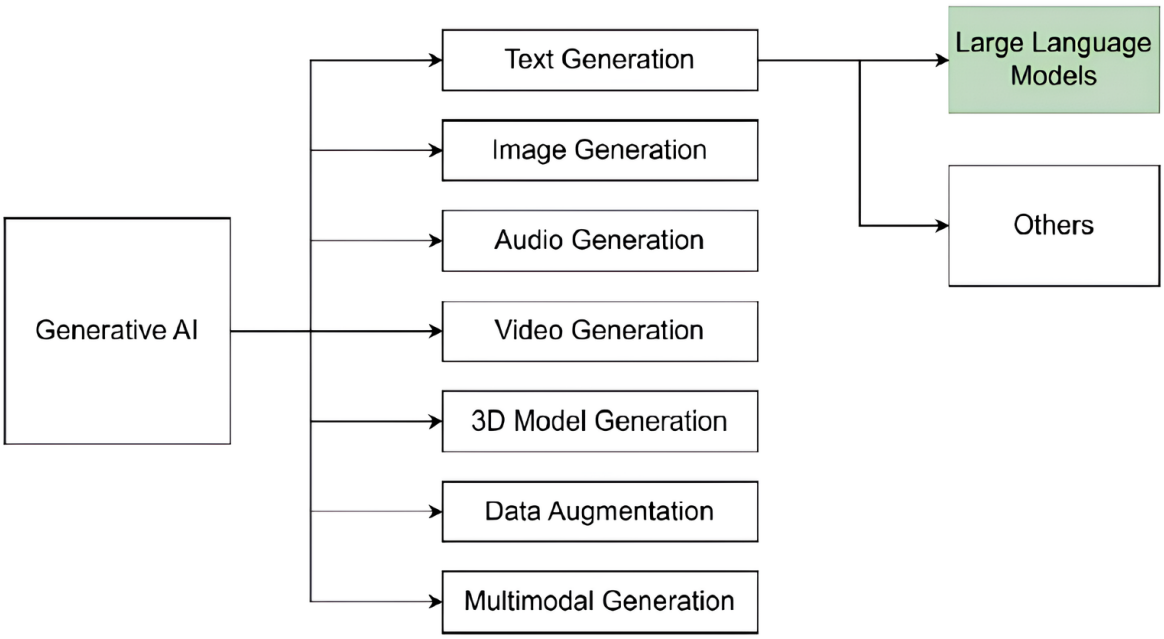


Figura 1.1: LLM come ramo della GenAI

Il processo di generazione del testo negli LLM è autoregressivo, il che significa che il modello predice e genera il token successivo in base alla sequenza di token già elaborata. Utilizzando meccanismi di attenzione, che stabiliscono le connessioni tra le parole e assicurano coerenza e pertinenza contestuale, gli LLM sono in grado di produrre testi che appaiono estremamente naturali e coerenti. Questo approccio non solo consente loro di generare articoli, storie e risposte a domande, ma anche di svolgere compiti complessi come la traduzione automatica, l'analisi del sentiment e la risposta automatica a richieste.

## 1.2 Transformer

I LLM, come i modelli GPT, si basano sull'architettura Transformer. Introdotta nel celebre articolo "*Attention is All You Need*"[22] dai ricercatori di Google nel 2017, l'architettura Transformer è progettata per elaborare e generare testi simili a quelli umani per una vasta gamma di compiti, dalla traduzione automatica alla generazione di testi per scopi generali. Al centro di questa architettura si trova il cosiddetto meccanismo di auto-attenzione (self-attention), che consente di elaborare i dati di input. A differenza dei modelli di deep learning precedenti per l'elaborazione del linguaggio naturale, come le Reti Neurali Ricorrenti (RNN) e le LSTM (Long Short-Term Memory), che elaborano i dati in modo sequenziale, i Transformer sono in grado di gestire i dati di input in parallelo. Questo non solo migliora l'efficienza del modello, ma aumenta anche la sua capacità di comprendere il contesto all'interno del linguaggio.

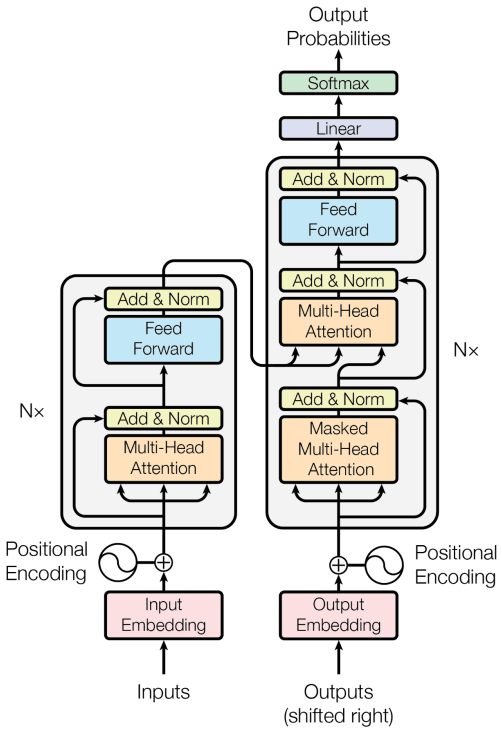


Figura 1.2: Architettura transformer

L'uso del meccanismo di auto-attenzione consente ai Transformer di pesare l'importanza di diverse parole in una sequenza e di considerare le relazioni tra le parole indipendentemente dalla loro posizione relativa. Questo approccio consente di catturare meglio le dipendenze a lungo raggio nel testo, migliorando così la qualità e la coerenza del testo generato e facilitando la gestione di compiti complessi legati al linguaggio naturale.
I componenti chiave dell'architettura Transformer sono:

- **Strato di Embedding**: Questo strato trasforma i token di input, ossia le parole, in vettori numerici che rappresentano sia il significato semantico sia il contesto del testo. L'Embedding Layer è cruciale per convertire le parole in rappresentazioni densi, catturando le loro caratteristiche semantiche e contestuali.

- **Codifica Posizionale (Positional Encoding)**: In questo passaggio si aggiungono informazioni sulla posizione di ogni token nella sequenza al suo embedding. Questo serve a compensare il fatto che il modello Transformer, a differenza di altri modelli, non processa i dati in ordine sequenziale.

- **Encoder e Decoder**: L'encoder ha il compito di elaborare il testo di input e di estrarre le informazioni contestuali, mentre il decoder si occupa di generare risposte coerenti prevedendo le parole successive nella sequenza. Questi due componenti lavorano in sinergia: l'encoder fornisce rappresentazioni dettagliate dei dati di input, che il decoder utilizza per produrre testi pertinenti e ben strutturati.

- **Meccanismo di Auto-Attenzione (Self-Attention Mechanism)**: Il meccanismo di auto-attenzione consente al modello di valutare l'importanza delle diverse parole nella sequenza di input. Questa capacità di pesare le parole in relazione al loro contesto globale è fondamentale per il linguaggio naturale, dove il significato di una parola può variare a seconda del contesto in cui si inserisce.

- **Reti Neurali Feedforward**: Sia l'encoder che il decoder sono dotati di reti neurali feedforward, che applicano trasformazioni aggiuntive ai dati elaborati dal meccanismo di auto-attenzione. Queste reti neurali permettono di catturare caratteristiche e dettagli più sottili del linguaggio naturale, migliorando la rappresentazione e l'elaborazione delle informazioni.

- **Normalizzazione dei Livelli e Connessioni Residuali (Layer Normalization & Residual Connections)**: Queste tecniche sono impiegate all'interno dei blocchi del modello per aumentare la stabilità dell'addestramento e prevenire il problema del gradiente che scompare, facilitando così l'addestramento di reti neurali più profonde. La normalizzazione dei livelli contribuisce a mantenere la stabilità durante il processo di addestramento, mentre le connessioni residue consentono al modello di preservare e integrare informazioni attraverso i vari strati.

### BERT e GPT

Come detto in precedenza l'architettura Transformer ha rivoluzionato l'approccio all'NLP. Su questa base sono stati costruiti BERT e GPT, due modelli potenti che apprendono il contesto del testo utilizzando meccanismi di attenzione in modo non supervisionato. BERT, sviluppato nei laboratori di Google AI, è un modello pre-addestrato che eccelle nell'analisi bidirezionale del testo, ossia da sinistra a destra e da destra a sinistra. Questo approccio conferisce a BERT una comprensione approfondita del linguaggio. BERT è stato addestrato utilizzando il *masked language modeling*, che prevede la previsione di parole mancanti in una frase. Con questa conoscenza, BERT ha ottenuto risultati notevoli in numerosi compiti di NLP, come l'analisi del sentiment, il riconoscimento di entità nominate e il question answering, stabilendo nuovi standard e superando i concorrenti. D'altra parte, GPT, sviluppato da OpenAI, si distingue per l'approccio unidirezionale all'elaborazione del testo. Sebbene analizzasse il testo esclusivamente da sinistra a destra, GPT ha dimostrato un'abilità straordinaria nella generazione di testo, creando frasi coerenti e contestualmente rilevanti con grande maestria. Addestrato nel campo del*causal language modeling*, GPT ha imparato a prevedere la parola successiva in una frase, affinando le sue capacità di generazione del testo. Con queste competenze, GPT ha stupito il mondo con le sue capacità di completamento e generazione del testo.

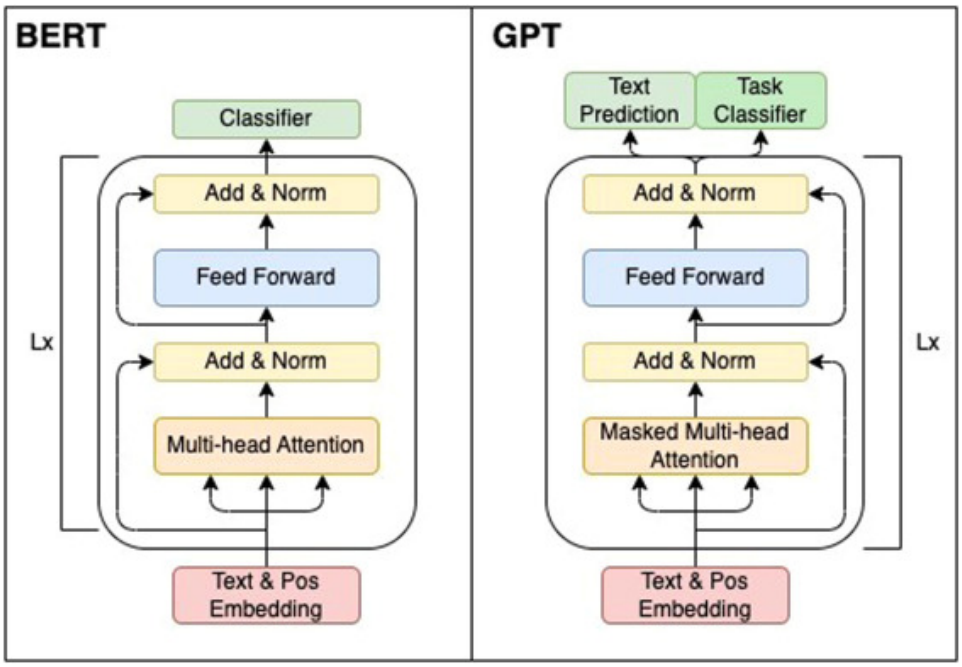


Figura 1.3: Confronto tra le architetture di BERT e GPT

Nonostante le loro differenze architetturali e nei metodi di addestramento, BERT e GPT sono adatti a diverse categorie di compiti NLP. È dunque utile confrontare le loro caratteristiche uniche:

- **Obiettivi di addestramento**: BERT è addestrato per prevedere parole mascherate basandosi sul contesto fornito da altre parole, mentre GPT è addestrato a prevedere la parola successiva in una frase, data una sequenza di parole precedenti.
- **Direzione del contesto**: Il "B" in BERT sta per "bidirezionale". BERT analizza una frase sia da sinistra a destra che da destra a sinistra, permettendo una comprensione più profonda del contesto e del significato. Al contrario, GPT utilizza un approccio unidirezionale, focalizzandosi solo sull'analisi da sinistra a destra.
- **Architettura del modello**: BERT è composto da un singolo modello encoder addestrato con due compiti principali: il masked language modeling (MLM), che prevede la previsione di parole mascherate, e il next sentence prediction (NSP). GPT, invece, è costituito da una pila di blocchi Transformer, ciascuno con più strati di auto-attenzione e strati feedforward.

- **Tecniche di addestramento**: BERT ha affinato le sue competenze attraverso il masked language modeling, mentre GPT ha eccelso nel causal language modeling, prevedendo la parola successiva in una frase.
- **Fine-tuning**: Entrambi i modelli richiedono grandi quantità di dati testuali per il fine-tuning. Nonostante la disponibilità di pesi di modelli open-source (non sempre garantita), il numero elevato di parametri rende il processo computazionalmente impegnativo e costoso. Tuttavia, BERT può essere fine-tuned utilizzando GPU su risorse cloud o server, rendendolo più accessibile e gestibile per ricercatori e sviluppatori.
- **Ambiti di applicazione**: BERT è particolarmente adatto per compiti come l'analisi del sentiment, il question answering e la classificazione del testo, dove è fondamentale comprendere le relazioni tra le parti di una frase. GPT, invece, ha dimostrato eccellenza nella generazione di testo, completamento di frasi, sintesi e traduzione linguistica.

## 1.3 Tokenizzazione

I LLM sono modelli predittivi. Questo significa che sono addestrati per prevedere elementi all'interno di una sequenza, basandosi sugli altri elementi presenti in quella sequenza. La tokenizzazione è il processo di suddivisione di questa sequenza in componenti discrete(token). Questi token, a loro volta, possono essere considerati come il vocabolario del modello, ovvero i tipi di unità che il modello è stato addestrato a riconoscere e a produrre. Per addestrare e utilizzare un LLM, è necessario prendere delle decisioni su cosa includere nel suo vocabolario.

### Tipi di Token

- **Parole**: Questo è il tipo di token più comune, dove il tokenizer suddivide il testo basandosi sugli spazi bianchi (spazi, tabulazioni, interruzioni di riga).
- **Caratteri**: In alcuni casi, i tokenizer possono suddividere il testo in singoli caratteri, specialmente per lingue che non utilizzano spazi o per compiti NLP specializzati.

- **Sub-parole**: Questo approccio è particolarmente utile per gestire parole fuori vocabolario (OOV), ossia parole che il modello non ha incontrato durante l'addestramento. I tokenizer di sub-parole suddividono le parole in unità più piccole e significative, come prefissi, suffissi o n-grammi di caratteri (sequenze di caratteri).

La tokenizzazione è essenziale perché trasforma il linguaggio umano complesso in unità che i computer possono comprendere e elaborare. Questo processo è fondamentale per vari compiti di NLP, tra cui traduzione automatica, sommario di testi, analisi del sentiment, classificazione del testo e risposta a domande.

### Embedding

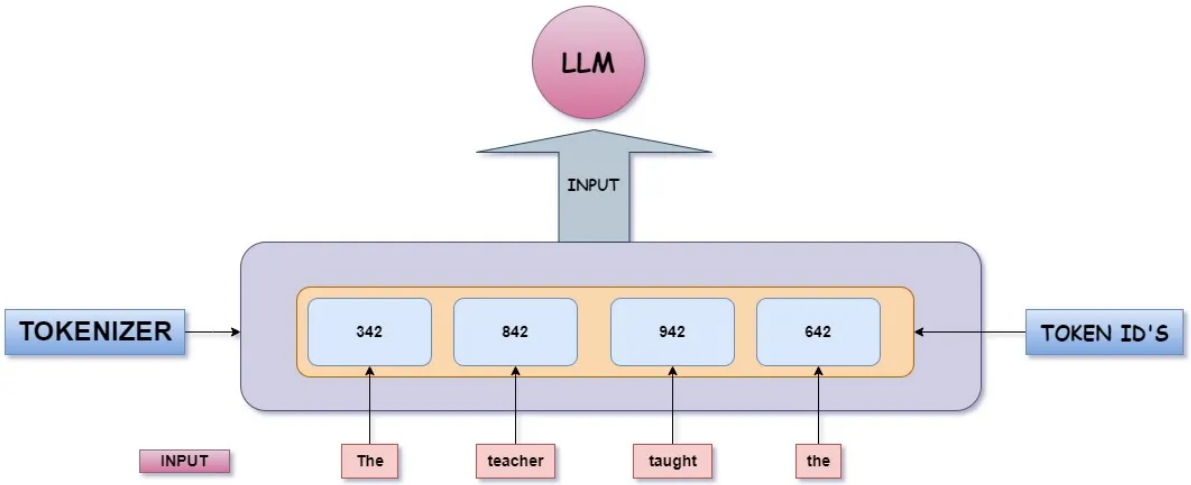


Figura 1.4: tokenizzazione con LLM

Dopo la tokenizzazione, il passo successivo è trasformare questi token in una forma che il computer possa comprendere e elaborare: è qui che entrano in gioco gli embeddings. Gli embeddings sono un metodo per tradurre i token, in una rappresentazione numerica che il computer può gestire. Questi embeddings aiutano il modello a comprendere le relazioni e il contesto, permettendo di riconoscere le connessioni tra le parole e di utilizzare tali connessioni per comprendere meglio il testo, principalmente attraverso il processo di

attenzione, che esploreremo successivamente. Ogni token riceve un ID numerico unico tramite l'embedding, che cattura il suo significato. Questa forma numerica consente al computer di valutare quanto due token siano simili, ad esempio riconoscendo che "felice" e "gioioso" sono vicini nel significato, nonostante siano parole diverse. questo passaggio è fondamentale perché aiuta il modello a comprendere il linguaggio in termini numerici, colmando il divario tra il linguaggio umano e l'elaborazione automatica. Inizialmente, ogni token viene assegnato a un insieme casuale di numeri come embedding. Con il progredire dell'addestramento del modello, ovvero man mano che il modello legge e apprende da una grande quantità di testo, questi numeri vengono aggiustati. L'obiettivo è ottimizzare questi numeri in modo che i token con significati simili abbiano set di numeri simili. Questa regolazione avviene automaticamente durante l'apprendimento del modello, che si adatta ai diversi contesti in cui i token compaiono. Sebbene il concetto di set numerici, o vettori, possa sembrare complesso, essi rappresentano semplicemente un modo efficiente per il modello di memorizzare e elaborare informazioni sui token. I vettori sono utilizzati perché costituiscono un metodo diretto per il modello per tenere traccia di come i token sono correlati tra loro. In sostanza, si tratta di ampie liste di numeri.

## 1.4 Training e Fine tuning

I LLM sono addestrati su un vasto corpus di testi con l'obiettivo di prevedere correttamente il token successivo in una sequenza. L'obiettivo è regolare i parametri del modello per massimizzare la probabilità di una previsione corretta basata sui dati osservati. Tipicamente, un modello viene addestrato su un grande dataset a scopo generale di testi provenienti da Internet. Talvolta, vengono utilizzati anche dataset più specifici per acquisire conoscenze specifiche di dominio. Questa fase è nota come fase di pre-addestramento, durante la quale il modello impara a comprendere il linguaggio ed è preparato per ulteriori perfezionamenti. Il processo di addestramento regola i pesi del modello per aumentare la probabilità di prevedere correttamente il token successivo in una sequenza. Questo aggiustamento si basa sui dati di addestramento, guidando il modello verso previsioni accurate dei token. Dopo il pre-addestramento, il modello

solitamente subisce un fine-tuning ovvero l'addestramento che ha già acquisito schemi e caratteristiche da un vasto dataset, utilizzando un dataset più piccolo e specifico per un determinato dominio. Questa metodologia è significativa perché addestrare un modello di linguaggio di grandi dimensioni da zero è estremamente costoso in termini di potenza computazionale e tempo. Utilizzare la conoscenza già incorporata nel modello pre-addestrato consente di ottenere prestazioni elevate in compiti specifici con ridotti requisiti di dati e computazione.

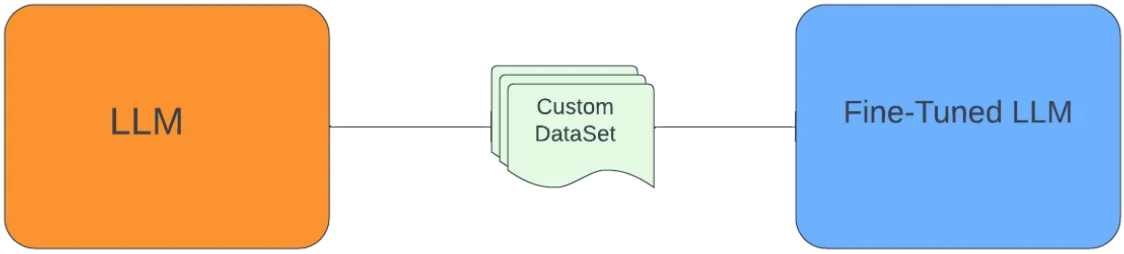


Figura 1.5: Processo di Fine Tuning

I passaggi chiave nel fine-tuning dei LLM includono:

1. **Selezione del Modello Pre-Addestrato**: Il primo passo nel fine-tuning di un LLM è la scelta accurata di un modello pre-addestrato che si allinei con l'architettura e le funzionalità desiderate. Questi modelli pre-addestrati sono generalmente modelli a scopo generico addestrati su un ampio corpus di dati non etichettati.

2. **Preprocessing del Dataset**: Una volta pronto il dataset, è necessario eseguire alcune operazioni di preprocessing per il fine-tuning, come la pulizia, la suddivisione in set di addestramento, validazione e test, e l'assicurazione della compatibilità con il modello su cui si intende effettuare il fine-tuning.

3. **Fine-Tuning**: Dopo aver selezionato il modello pre-addestrato, è necessario adattarlo al nostro dataset pre-elaborato, che è più specifico per il compito in questione. Questo dataset può essere relativo a un dominio o applicazione particolare, consentendo al modello di adattarsi e specializzarsi per quel contesto.

4. **Adattamento al Compito Specifico**: Durante il fine-tuning, i parametri del modello vengono regolati in base al nuovo dataset, aiutando il modello a comprendere meglio e generare contenuti rilevanti per il compito specifico. Questo processo conserva la conoscenza linguistica generale acquisita durante l'addestramento iniziale, mentre adatta il modello alle sfumature del dominio target.

### Metodi di Fine Tuning

Il fine-tuning di un LLM implica un processo di apprendimento supervisionato, dove un dataset etichettato viene utilizzato per ottimizzare i pesi del modello e migliorarne le prestazioni in compiti specifici. Di seguito sono descritti alcuni metodi principali di fine-tuning:

- **Full Fine-Tuning (Istruzione Fine-Tuning)**:Questo approccio mira a migliorare le prestazioni del modello addestrandolo su esempi che orientano le sue risposte a richieste specifiche. La scelta del dataset è cruciale e deve essere mirata al compito specifico, come sintesi o traduzione. Questo metodo aggiorna tutti i pesi del modello, creando una nuova versione con capacità ampliate. Tuttavia, richiede risorse significative in termini di memoria e computazione, simili a quelle necessarie per l'addestramento iniziale, per gestire gradienti, ottimizzatori e altri componenti.

- **Parameter Efficient Fine-Tuning (PEFT)**: PEFT è una tecnica più efficiente rispetto al fine-tuning completo, poiché aggiorna solo un sottoinsieme di parametri, mantenendo congelati gli altri. Questo approccio riduce il numero di parametri addestrabili e semplifica i requisiti di memoria, evitando la perdita di conoscenze precedentemente apprese. È particolarmente utile per gestire problemi di archiviazione quando si effettua il fine-tuning per più compiti. Due metodi di PEFT ampiamente utilizzati sono Low-Rank Adaptation (LoRA) e Quantized LoRA (QLoRA).

**Low-Rank Adaptation (LoRA)** è un metodo avanzato di fine-tuning che, invece di aggiornare tutti i pesi del modello, affina due matrici più piccole che approssimano la matrice di pesi principale. Queste matrici costituiscono l'adattatore LoRA, che viene integrato nel modello pre-addestrato per l'inferenza. Dopo il fine-tuning con LoRA, il

modello originale rimane invariato, mentre emerge un "adattatore LoRA" significativamente più piccolo (spesso solo una piccola percentuale della dimensione originale). Durante l'inferenza, l'adattatore LoRA viene combinato con il modello originale, riducendo i requisiti di memoria per gestire diversi compiti e casi d'uso. Invece **Quantized LoRA (QLoRA)** è una versione più efficiente di LoRA che quantizza i pesi degli adattatori LoRA a una precisione inferiore (ad esempio, 4-bit invece di 8-bit). Questo approccio riduce ulteriormente l'impronta di memoria e i requisiti di archiviazione. In QLoRA, il modello pre-addestrato viene caricato nella memoria GPU con pesi quantizzati a 4-bit, mantenendo un livello di efficacia comparabile a quello di LoRA.

### 1.4.1 RAG

L'accuratezza dei LLM può essere significativamente migliorata integrando conoscenze specifiche di dominio attraverso l'uso di documenti esterni. Questo processo consiste nell'aggiornare la base di conoscenze del modello con informazioni rilevanti, permettendo così al modello di fondare le sue risposte su dati più specifici e aggiornati. Quando viene effettuata una query, un modulo di "recupero" recupera i documenti pertinenti, i quali sono poi utilizzati per arricchire la risposta del modello. Questa metodologia è essenziale per le architetture di recupero, le quali operano come segue: inizialmente, il sistema genera una rappresentazione vettoriale della query; successivamente, questa rappresentazione viene utilizzata per effettuare una ricerca semantica all'interno di un database di documenti, confrontando i vettori e calcolando i punteggi di similarità; infine, il modello di linguaggio utilizza i documenti recuperati con il punteggio più alto come contesto per fornire la risposta finale, estraendo informazioni direttamente dai paragrafi senza aggiungere contenuti non inferibili. La generazione aumentata da recupero (Retrieval-Augmented Generation) è una tecnica avanzata che migliora le capacità dei modelli di linguaggio incorporando dati provenienti da fonti esterne. Questa tecnica combina le informazioni esterne con il contesto già presente nel prompt del modello, consentendo così di fornire risposte più accurate e pertinenti. Un'analisi dettagliata di questa tecnica sarà presentata nel Capitolo 3.

### 1.4.2 Prompt

Il testo fornito ai LLM, che può includere testi, immagini, numeri o tabelle, è comunemente chiamato "prompt". I prompt sono istruzioni fornite ai sistemi AI, come GPT-3 e GPT-4 di OpenAI, per generare testi che imitano il linguaggio umano. La precisione e l'accuratezza del prompt sono fondamentali per ottenere un output di alta qualità. Generalmente, prompt concisi, descrittivi e, a seconda del compito, producono risultati più efficaci, consentendo al modello di esprimere creatività e di orientarsi verso il risultato desiderato. L'uso di parole o frasi specifiche aiuta a focalizzare il modello sulla produzione di contenuti pertinenti. Per creare prompt efficaci è necessario avere uno scopo chiaro, mantenere la semplicità, utilizzare strategicamente le parole chiave e garantire l'azione. Testare i prompt prima dell'uso finale è essenziale per assicurarsi che l'output sia rilevante e privo di errori. Alcune tecniche per la creazione di un buon prompt:

1. **Usare un Linguaggio Preciso**: La chiarezza nel prompt può migliorare notevolmente l'accuratezza dell'output.

2. **Fornire un Contesto Adeguato**: Il contesto aiuta il modello a comprendere meglio l'output richiesto.

3. **Sperimentare con Varianti**: Provare diverse formulazioni del prompt per trovare l'approccio più efficace.

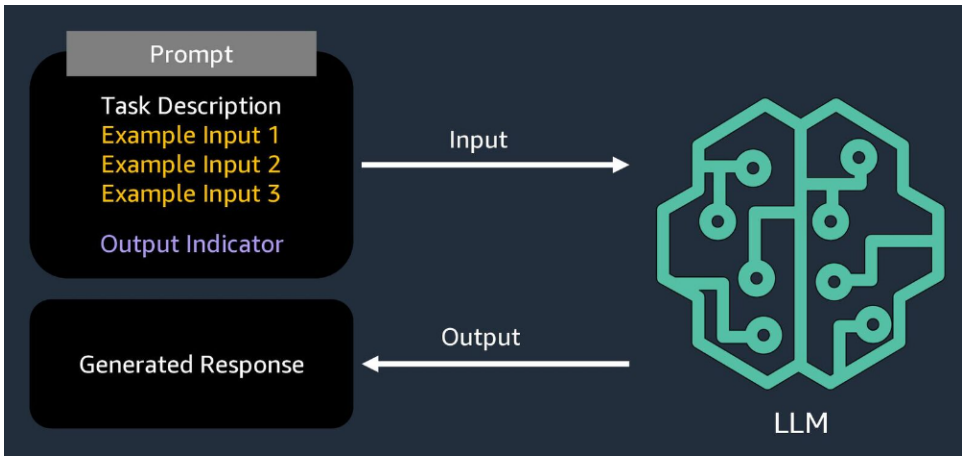


Figura 1.6: Processo prompt

### 1.4.3 Dimensione del Contesto

La dimensione del contesto, o finestra di contesto, è un aspetto cruciale dei LLM. Essa si riferisce al numero massimo di token che il modello può elaborare in una singola richiesta. La dimensione del contesto influenza la lunghezza del testo che il modello è in grado di gestire in un dato momento, e incide direttamente sulle prestazioni del modello e sui risultati che esso produce.
I diversi LLM sono progettati con dimensioni di contesto variabili. Ad esempio, il modello "gpt-3.5-turbo-16k" di OpenAI possiede una finestra di contesto capace di gestire fino a 16.000 token. Esiste un limite intrinseco al numero di token che un modello può generare. I modelli più piccoli possono gestire fino a 1.000 token, mentre modelli più grandi, come GPT-4, possono arrivare fino a 32.000 token, secondo i dati disponibili al momento della redazione di questo testo. La dimensione del contesto è significativa per diversi motivi:

- **Ambito e Complessità dell'Input**: Una finestra di contesto più ampia consente a un LLM di gestire input più dettagliati e complessi, influenzando così la sua applicabilità e il grado di precisione. Ad esempio, una finestra di contesto di 4.000 token, come quella di GPT-3.5 o Llama 2, è equivalente a sei pagine di testo, mentre una finestra di 32.000 token corrisponde a 49 pagine.

- **Coerenza**: Poiché i modelli non dispongono di memoria a lungo termine, la dimensione del contesto determina quanto input precedente il modello può ricordare. Questo influisce sulla coerenza e sull'accuratezza dell'output prodotto.

- **Accuratezza**: Una finestra di contesto più grande aumenta la possibilità che il modello fornisca una risposta rilevante, grazie a una comprensione più approfondita dell'input.

#### Dimensioni del Contesto Breve

Una finestra di contesto ridotta presenta vantaggi significativi. Essa consente tempi di risposta più rapidi, grazie alla minore quantità di dati da elaborare, migliorando così l'esperienza dell'utente. Inoltre, l'efficienza delle risorse è ottimizzata, poiché modelli con contesti brevi richiedono meno potenza computazionale, memoria ed energia, rendendo

l'implementazione più economica e accessibile su una varietà di dispositivi. Tuttavia, ci sono anche delle limitazioni. I modelli con contesti brevi soffrono della mancanza di memoria a lungo termine, limitandosi alla finestra di contesto corrente e impedendo l'accesso a informazioni precedenti. Questo può compromettere l'efficacia del modello in compiti che richiedono una comprensione estesa del contesto. Inoltre, la comprensione contestuale può essere insufficiente, portando a risposte imprecise o incomplete e richiedendo potenzialmente ripetizioni di input o tecniche di ingegneria dei prompt per migliorare la precisione.

#### Dimensioni del Contesto Lungo

Al contrario, una finestra di contesto più ampia offre vantaggi notevoli. Permette la gestione di input più estesi e complessi, rendendo il modello adatto a una gamma più ampia di applicazioni, come documenti lunghi, database estesi, grandi basi di codice e più fonti di dati. Inoltre, un contesto maggiore facilita una comprensione più completa della richiesta dell'utente, migliorando la coerenza e la pertinenza delle risposte. L'efficienza è migliorata poiché il modello può elaborare e comprendere più informazioni in un singolo passaggio, riducendo la necessità di passaggi iterativi. Tuttavia, l'adozione di finestre di contesto lunghe comporta anche delle sfide. La precisione può essere compromessa a causa di difficoltà nel mantenere l'accuratezza delle informazioni, specialmente nella parte centrale del contesto, fenomeno noto come "*missing middle*". Inoltre, l'uso di contesti estesi aumenta significativamente i requisiti computazionali e di memoria, comportando tempi di risposta più lunghi e impatti negativi sulla performance in tempo reale delle applicazioni. Infine, l'addestramento e l'implementazione di modelli con finestre di contesto ampie richiedono risorse elevate, limitando l'accessibilità e la praticabilità per organizzazioni con risorse limitate e dispositivi con capacità computazionale inferiore.

## 1.5 Allucinazioni e Biases

Le "*allucinazioni*" nei sistemi di intelligenza artificiale si riferiscono ai casi in cui questi sistemi generano output sia esso testo, immagini o altri tipi di dati che non sono in linea con la realtà o le informazioni fornite. Ad esempio, se ChatGPT produce una

risposta convincente ma errata dal punto di vista fattuale, si manifesta una discrepanza tra l'output dell'IA e la conoscenza reale o il contesto. Nei LLM, le allucinazioni si verificano quando il modello crea risposte che non corrispondono a fatti verificabili o al contesto fornito. Questo fenomeno può contribuire alla diffusione di disinformazione, un problema particolarmente grave in settori cruciali come la sanità e l'istruzione, dove l'accuratezza delle informazioni è fondamentale. Inoltre, i bias presenti nei LLM possono portare a risultati che favoriscono determinati punti di vista o rinforzano stereotipi e discriminazioni dannose. Questo indica la tendenza dei modelli a produrre output o decisioni che riflettono i pregiudizi presenti nei dati di addestramento. Ad esempio, se i dati di addestramento provengono prevalentemente da una determinata regione, il modello potrebbe mostrare una parzialità verso il linguaggio, la cultura o le prospettive di quella regione. Inoltre, i bias incorporati nei dati di addestramento come quelli legati al genere o alla razza possono manifestarsi nelle risposte dell'IA, portando a output biasati o discriminatori. Affrontare le allucinazioni e i pregiudizi nei sistemi di intelligenza artificiale richiede un approccio complesso e articolato. Questo include il miglioramento del processo di addestramento del modello, l'uso di tecniche di verifica e l'assicurarsi che i dati di addestramento siano diversificati e rappresentativi. Trovare un equilibrio tra massimizzare il potenziale del modello e minimizzare questi problemi rimane una sfida significativa. Tuttavia, è sorprendente notare che queste "*allucinazioni*" potrebbero avere un valore nelle aree creative come la scrittura di fiction, permettendo la creazione di contenuti nuovi e innovativi. L'obiettivo finale è sviluppare LLM potenti ed efficienti, ma anche affidabili, equi e giusti, massimizzando i benefici di questa tecnologia mentre si minimizzano i rischi, assicurando che i vantaggi siano accessibili a tutti.

## 1.6 Modelli LLM più importanti

Esistono numerosi modelli di linguaggio di grande rilevanza, e il numero di modelli significativi è in continua crescita, rendendo impraticabile una lista completa e aggiornata. La rapida evoluzione dei LLM implica che ogni elenco diventi rapidamente obsoleto. L'elenco proposto si concentra sui LLM più significativi e rilevanti, non necessariamente sui modelli con le migliori performance sui benchmark, anche se molti di essi eccellono in

questo ambito.Di seguito è presentata una panoramica dei principali LLM attualmente disponibili, evidenziandone le caratteristiche distintive e le applicazioni principali.

- **ChatGPT**: OpenAI ha rivoluzionato il panorama dei modelli di linguaggio con la sua serie GPT, che include alcune delle tecnologie più avanzate nel campo dell'intelligenza artificiale. *GPT-3 (2020)* è un modello con 175 miliardi di parametri, noto per le sue avanzate capacità di generazione del linguaggio grazie alla sua architettura di trasformatore decoder-only. *GPT-3.5 (2021)*, un aggiornamento di GPT-3, migliora la comprensione del linguaggio tramite affinamento basato su feedback umano ed è il modello che alimenta ChatGPT. *GPT-4 (2023)* segna un progresso significativo con capacità multimodali, gestendo testo e immagini, e prestazioni comparabili a quelle umane in vari test. *GPT-4o (2024)*, successore di GPT-4, offre risposte più rapide e una maggiore interattività multimodale, includendo anche capacità avanzate di comprensione del contesto e delle emozioni.

- **Claude**: Claude, sviluppato da Anthropic, si distingue per il suo approccio di intelligenza artificiale costituzionale, che guida gli output dell'AI secondo principi che garantiscono utilità, innocuità e accuratezza. L'ultima iterazione, Claude 3.5 Sonnet, migliora la comprensione delle sfumature, dell'umorismo e delle istruzioni complesse rispetto alle versioni precedenti e opera a una velocità doppia rispetto a Claude 3 Opus.

- **Falcon 40B**: Sviluppato dal Technology Innovation Institute, Falcon 40B è un modello basato su trasformatore e decoder causale. È open source e allenato su dati in inglese. Esistono varianti più piccole del modello: Falcon 1B e Falcon 7B (1 miliardo e 7 miliardi di parametri).

- **Gemini**: Sviluppato da Google, è una famiglia di LLM che alimenta il chatbot omonimo, sostituendo il modello Palm. I modelli Gemini sono multimodali, in grado di gestire testo, immagini, audio e video, e sono integrati in numerose applicazioni e prodotti Google. Disponibile in tre varianti (Ultra, Pro e Nano), Gemini supera GPT-4 nella maggior parte dei benchmark valutati.

- **Gemma**: È una serie di modelli di linguaggio open source di Google, allenati con le stesse risorse di Gemini. Gemma è disponibile in due dimensioni: 2 miliardi e 7 miliardi di parametri. Questi modelli possono essere eseguiti localmente su computer personali e superano i modelli Llama 2 di dimensioni simili in diversi benchmark.

- **PaLM**: Il Pathways Language Model (PaLM) di Google è un modello basato su trasformatore con 540 miliardi di parametri, alimentando il chatbot Bard. PaLM è specializzato in compiti di ragionamento come coding, matematica, classificazione e risposta a domande. PaLM eccelle anche nel decomporre compiti complessi in sotto-compiti più semplici.

- **Llama**: Large Language Model Meta AI(Llama) è un modello di Meta, l'azienda madre di Facebook e Instagram. Inizialmente rilasciato a ricercatori e sviluppatori approvati, è ora open source. L'ultimo modello rilasciato è *Llama 3.1*, disponibile in varianti con 8, 70 e 405 miliardi di parametri. Il modello precedente, Llama 2, rimane disponibile con configurazioni di 7, 13 e 70 miliardi di parametri, sebbene le sue prestazioni siano inferiori rispetto alla più recente versione Llama 3.1.

- **Mistral**: È un modello di linguaggio con 7 miliardi di parametri che supera i modelli di dimensioni simili, come Llama, in tutti i benchmark valutati. Include anche un modello specializzato per seguire le istruzioni e viene rilasciato sotto la licenza Apache 2.0.

- **Orca**: Sviluppato da Microsoft, mira a migliorare i progressi fatti dai modelli open source imitando i processi di ragionamento degli LLM. Ottiene prestazioni comparabili a GPT-4 con significativamente meno parametri e si comporta bene in molti compiti rispetto a GPT-3.5. Orca è costruito sulla versione da 13 miliardi di parametri di Llama.

- **Phi-1**: È un modello di linguaggio basato su trasformatore di Microsoft, con solo 1,3 miliardi di parametri. Allenato per quattro giorni su dati di alta qualità, Phi-1 rappresenta una tendenza verso modelli più piccoli ma altamente capaci, specializzati nel coding Python.

# 2 Applicazione dei LLM in ambito sanitario

## 2.1 Salute del paziente

I modelli di linguaggio possono potenziare la salute dei pazienti rafforzando competenze mediche fondamentali, come la conoscenza fattuale e le abilità di comunicazione interpersonale. Strumenti come ChatGPT possiedono una notevole comprensione semantica nel campo medico e dimostrano capacità di ragionamento clinico, come evidenziato dalla loro performance negli esami di abilitazione medica. Un ulteriore perfezionamento di questi strumenti, attraverso addestramenti supplementari con domande stilizzate sul modello degli esami di licenza medica e risposte selezionate da esperti clinici, può affinare il ragionamento e la comprensione medica. Attualmente, GPT-4 rappresenta il modello con la conoscenza più avanzata nel dominio medico. Tuttavia, è importante considerare che questi strumenti possono riprodurre bias esistenti e perpetuare disuguaglianze legate a razza, genere, orientamento sessuale e status socio-economico. I LLM possono inoltre migliorare la comunicazione tra personale sanitario e pazienti grazie alla loro capacità di semplificare il linguaggio. Essi possono essere consultati dai pazienti in qualsiasi momento, superando le limitazioni temporali degli esperti sanitari e rendendo il contatto più agevole e confortevole. Questi vantaggi sono particolarmente evidenti in contesti caratterizzati da stigma sociale, come le dipendenze o le malattie sessualmente trasmissibili. Strumenti digitali sviluppati con l'avvento degli smartphone, come *First Derm*, un'applicazione di teledermoscopia, e *Pahola*, un chatbot per l'orientamento sul consumo di alcol, offrono supporto a distanza. Sebbene vi siano ancora limiti tecnici e una parziale accettazione da parte dei professionisti sanitari, i rapidi progressi nella tecnologia dei modelli di linguaggio potrebbero contribuire a superare queste barriere. È fondamentale riconoscere che, nonostante le loro capacità avanzate, tali strumenti non possono sostituire l'empatia umana, soprattutto in situazioni emotivamente complesse. Questi modelli possono inoltre semplificare la gestione della documentazione clinica, convertendo informazioni non strutturate in dati strutturati. Questa capacità non solo

riduce il carico amministrativo per i professionisti sanitari, ma migliora anche l'efficienza nella gestione dei dati clinici e nella comunicazione con i pazienti.

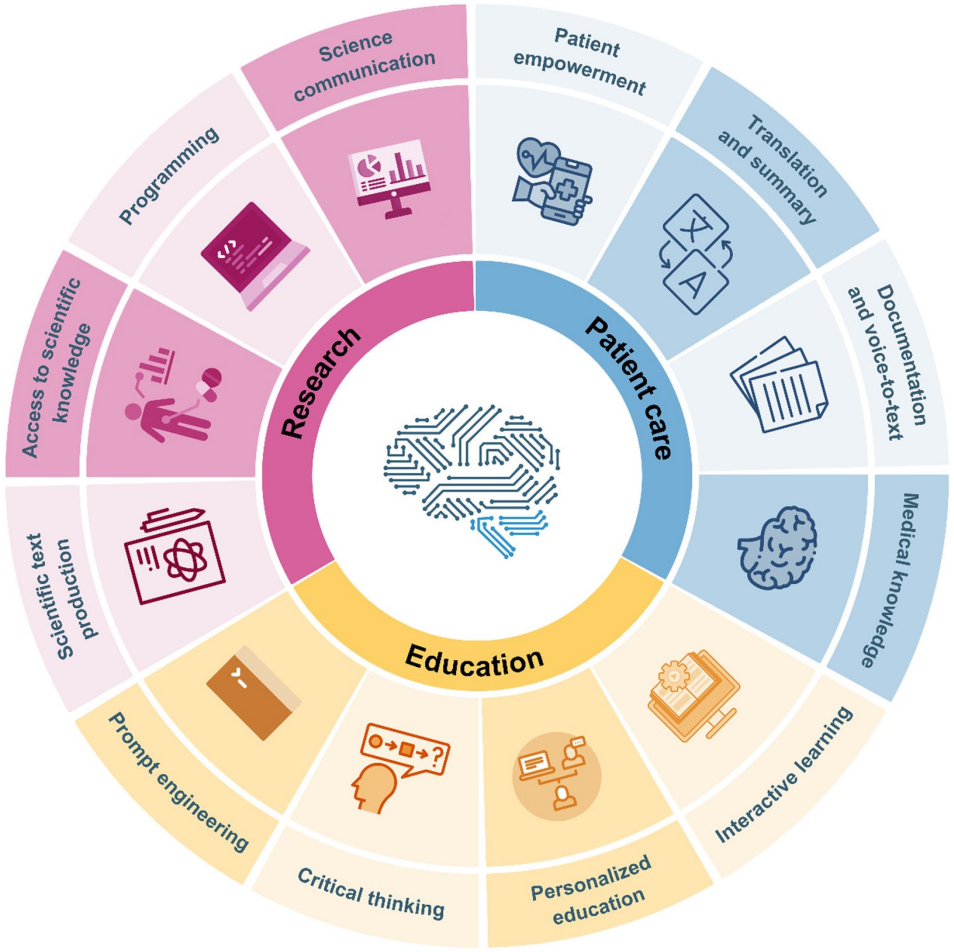


Figura 2.1: LLM in ambito medico. *Immagine concessa sotto licenza Creative Commons Attribution (CC BY) da [6]*

Come detto in precedenza, i LLM utilizzano algoritmi avanzati di intelligenza artificiale per generare testi che somigliano a quelli scritti dall'uomo. Addestrati su enormi quantità di testo provenienti da diverse fonti, questi modelli possono rispondere a domande, fornire riassunti e traduzioni. Il loro utilizzo però solleva anche interrogativi importanti riguardo alla precisione delle informazioni, alla privacy dei dati e ai possibili bias. Tuttavia data la loro capacità di elaborare e comprendere concetti complessi, i LLM hanno il potenziale

per apportare significativi cambiamenti nei settori dell'assistenza ai pazienti, della ricerca medica e della formazione medica. Nello specifico esploreremo come gli LLM potrebbero influenzare questi ambiti chiave della medicina.

### 2.1.1 Cura del paziente

Nel settore sanitario, una comunicazione chiara ed efficace è essenziale per garantire un'alta qualità della cura dei pazienti. Essa influisce profondamente sulle dinamiche tra pazienti e operatori sanitari, sulla soddisfazione dei pazienti e sui risultati clinici complessivi. La precisione nella comprensione del linguaggio, sia parlato che scritto, è cruciale per facilitare lo scambio di informazioni tra i professionisti medici riguardo ai dati dei pazienti, alle diagnosi e alle opzioni terapeutiche [16].
In questo contesto, i LLM emergono come strumenti innovativi per migliorare la cura dei pazienti. I principali ambiti di utilizzo di questi strumenti includono:

- **Elaborazione Intelligente delle Note Cliniche**: Le tecnologie avanzate di elaborazione del linguaggio naturale possono apportare notevoli miglioramenti nella gestione e nell'estrazione di dati da grandi volumi di note cliniche. Questi strumenti sono in grado di automatizzare la trascrizione e la documentazione delle interazioni con i pazienti, sia in formato audio che testuale. Tale automazione non solo allevia il carico amministrativo per i professionisti sanitari, ma migliora anche la qualità complessiva delle cure, permettendo agli operatori di concentrarsi maggiormente sull'interazione e sulla cura personalizzata dei pazienti. Questi strumenti possono essere addestrati per riconoscere ed estrarre informazioni cruciali, come nomi dei pazienti, date, condizioni mediche e farmaci, utilizzando tecniche di Riconoscimento di Entità Nominate (NER). Inoltre, possono estrarre dati temporali relativi alla progressione del trattamento e ai cambiamenti nelle condizioni del paziente. La categorizzazione delle note cliniche diventa così più gestibile, facilitando l'organizzazione e il recupero delle informazioni. La redazione di riassunti sintetici delle note estese migliora l'efficienza nella revisione delle informazioni senza necessità di consultare l'intero documento. Infine, la capacità di automatizzare l'anonimizzazione delle note cliniche garantisce la conformità alle normative sulla privacy e

sulla protezione dei dati. Recenti sviluppi includono framework come DeID-GPT, progettati per assicurare l'adesione alle regolazioni HIPAA.

- **Supporto alle Decisioni Cliniche**: I LLM possono supportare la diagnosi identificando modelli anomali nelle note cliniche e fornendo spiegazioni dettagliate che possono migliorare l'accordo tra i medici. Studi recenti hanno dimostrato che le spiegazioni generate possono aumentare il tasso di accordo tra i medici su ipotesi diagnostiche, sebbene possano verificarsi errori compresi tra il 5% e il 30%, dovuti alla tendenza dei modelli a integrare tutti i sintomi con l'ipotesi diagnostica corrente. Questi strumenti rappresentano un'innovazione significativa nel supporto ai pazienti, offrendo interazioni personalizzate ed efficienti. Possono superare le barriere linguistiche e facilitare la comunicazione tra operatori sanitari e pazienti attraverso la semplificazione, la traduzione in altre lingue e la sintesi di informazioni mediche complesse. Inoltre, possono funzionare come interfacce di linguaggio naturale, consentendo ai pazienti di porre domande e ricevere risposte basate sulla loro situazione clinica.

- **Assistenti Medici Virtuali e Chatbot Sanitari**: Gli assistenti virtuali basati su LLM possono gestire compiti come la pianificazione degli appuntamenti, la gestione delle richieste di prescrizione e il routing delle chiamate ai vari reparti. Questi strumenti riducono il carico amministrativo, permettendo ai professionisti sanitari di concentrarsi maggiormente sulla cura dei pazienti. Inoltre, possono analizzare i dati dei pazienti per prevedere situazioni critiche e ottimizzare l'allocazione delle risorse nelle sale di emergenza e nella previsione delle riammisioni ospedaliere. Questo approccio proattivo contribuisce a una gestione più efficace delle risorse sanitarie. Inoltre, questi strumenti possono elaborare piani di salute mentale personalizzati, considerando le caratteristiche individuali e la storia clinica del paziente. Tali piani possono includere raccomandazioni su terapie, strategie di coping e aggiustamenti ai farmaci.

- **Radiologia e Imaging**: Integrando dati visivi e testuali, i modelli di linguaggio visivo-multimodale mostrano un notevole potenziale per migliorare l'analisi delle immagini mediche. I radiologi possono trarre vantaggio da questi modelli, che facili-

tano l'identificazione precoce di anomalie nelle immagini mediche e contribuiscono alla generazione di interpretazioni diagnostiche più precise e complete, avanzando così l'accuratezza e l'efficienza dei processi diagnostici nel campo dell'imaging medico.

- **Sintesi Automatica dei Rapporti Medici dai Dati di Imaging**: La generazione automatica di rapporti medici a partire dalle immagini è cruciale per semplificare compiti complessi e soggetti a errori, affrontati da patologi e radiologi. Questo campo emergente all'intersezione tra sanità e IA mira ad alleggerire il carico sui professionisti medici esperti e migliorare l'accuratezza dei meno esperti. L'integrazione dell'IA con l'imaging medico facilita la redazione automatica dei rapporti, comprendente rilevamenti anomali, osservazioni normali rilevanti e storia del paziente. I primi sforzi hanno impiegato reti neurali basate sui dati, combinando modelli convoluzionali e ricorrenti per rapporti di singole frasi; tuttavia, sono emerse limitazioni nel catturare la complessità dei casi medici reali. I recenti progressi hanno sfruttato LLM come ChatCAD, che ha consentito applicazioni più sofisticate. ChatCAD migliora le reti CAD per immagini mediche, portando a significativi miglioramenti nella generazione dei rapporti. ChatCAD+ affronta ulteriormente le discrepanze stilistiche, garantendo universalità e affidabilità attraverso un sistema di recupero dei modelli per coerenza con l'expertise umana.

### 2.1.2 Ricerca

Fornire un'assistenza sanitaria di alta qualità richiede che i medici integrino costantemente le ultime evidenze scientifiche nei loro processi decisionali. Inoltre, i medici sono spesso coinvolti in attività di ricerca preclinica, traslazionale e clinica. Una comunicazione efficiente dei risultati della ricerca, attraverso pubblicazioni scritte e presentazioni orali in conferenze, è essenziale affinché tali risultati raggiungano le comunità mediche e scientifiche appropriate e, in ultima analisi, vengano adottati nella pratica clinica. I LLM avranno probabilmente un impatto significativo sulla ricerca medica nel prossimo futuro. Tuttavia, sebbene abbiano il potenziale per democratizzare l'accesso alle evidenze scientifiche, potrebbero anche contribuire alla diffusione di informazioni errate e facili-

tare comportamenti scientificamente scorretti. In questo contesto, i LLM potrebbero influenzare l'accesso al sapere scientifico attraverso:

- **Accesso al sapere scientifico**: La rapida evoluzione della ricerca scientifica genera un numero crescente di pubblicazioni, spesso di qualità variabile, che rendono difficile per i ricercatori rimanere aggiornati. I LLM hanno dimostrato un'efficienza notevole nella revisione di volumi estesi di letteratura medica. Potrebbero facilitare l'accesso alle conoscenze scientifiche sintetizzando concetti e studi esistenti, riducendo così la necessità di consultare numerose risorse. Tuttavia, la qualità delle sintesi dipende dai dati di addestramento e, poiché i LLM non vengono aggiornati in tempo reale, potrebbero non riflettere gli sviluppi scientifici più recenti. In un campo in rapido sviluppo come la sanità, mantenere la propria conoscenza al passo con le ultime innovazioni è fondamentale, e i LLM possono svolgere un ruolo cruciale nel garantire che la pratica sanitaria rimanga all'avanguardia nell'innovazione e nella cura basata su evidenze.

- **Produzione di testi scientifici**: I LLM possono generare testi scientifici adattando contenuti, linguaggio e stile, arrivando a produrre abstract che difficilmente si distinguono da quelli redatti da umani. Tuttavia, l'uso dei LLM nella scrittura scientifica richiede ancora revisioni significative a causa di errori, superficialità e ripetitività nei contenuti generati. Sebbene i LLM possano rivoluzionare la comunicazione scientifica, c'è il rischio che la loro adozione possa compromettere la qualità delle pubblicazioni, rendendo più difficile verificare l'autenticità e l'accuratezza delle informazioni riportate. È quindi essenziale stabilire linee guida per l'uso dei LLM nella redazione scientifica.

- **Programmazione informatica**: I LLM, addestrati su vari linguaggi di programmazione, possono assistere nella scrittura e nel debugging del codice, traducendo linguaggi e generando codice a partire da input in linguaggio naturale. Pur essendo talvolta imprecisi, questi modelli possono fornire soluzioni su richiesta, supportando ricercatori e clinici, anche con limitate competenze di programmazione, nello svolgimento di compiti tecnici complessi, migliorando così la loro efficienza e capacità di testare ipotesi scientifiche.

- **Riproducibilità**: La riproducibilità è cruciale per mantenere elevati standard nella ricerca scientifica. Tuttavia, le continue modifiche e aggiornamenti dei LLM possono complicare la riproduzione coerente dei risultati, poiché diversi modelli e versioni possono produrre risposte variabili. È quindi importante documentare con precisione i prompt e le versioni dei modelli utilizzati, e sviluppare soluzioni open-access per il controllo delle versioni, al fine di garantire la riproducibilità delle ricerche nel tempo.

- **Scoperta di Farmaci**: I modelli di linguaggio di grandi dimensioni hanno un impatto significativo nella facilitazione della scoperta di farmaci grazie alla loro capacità di analizzare strutture molecolari complesse, identificare composti promettenti con potenziale terapeutico e prevedere l'efficacia e la sicurezza di questi candidati. In uno studio corrispondente, gli autori hanno esplorato l'uso di modelli di linguaggio biochimici pre-addestrati per avviare modelli di generazione di molecole mirate, confrontando strategie di avvio a una o due fasi e valutando la generazione di composti utilizzando la ricerca beam e il campionamento. I risultati hanno dimostrato che i modelli avviati con calore hanno superato i modelli di base e che la strategia a una fase ha mostrato una migliore generalizzazione in termini di valutazione del docking e metriche di benchmark, mentre la ricerca beam si è rivelata più efficace del campionamento per valutare la qualità dei composti.

### 2.1.3 Formazione medica

La formazione ha subito profonde trasformazioni con l'emergere di nuove tecnologie. Analogamente, la disponibilità di un'enorme quantità di informazioni tramite internet e dispositivi intelligenti ha reso la memorizzazione meno cruciale nell'educazione medica, portando gli educatori a enfatizzare maggiormente il pensiero critico, il dibattito e la discussione, competenze che rimangono indispensabili. I LLM avranno probabilmente un impatto significativo sull'evoluzione delle metodologie educative, poiché possono facilitare il ragionamento. Nello specifico:

- **Usi benefici dei LLM nell'educazione**: I LLM, se utilizzati con responsabilità, possono arricchire le strategie educative in diversi modi. Possono generare rias-

sunti, presentazioni, traduzioni e spiegazioni su una vasta gamma di argomenti, con la possibilità di adattare profondità, tono e stile dell'output. Questi strumenti possono anche essere impiegati per creare simulazioni didattiche interattive, come simulazioni di conversazioni con pazienti fittizi, consentendo agli studenti di praticare la raccolta di anamnesi o la valutazione di diagnosi e piani di trattamento.

- **Impatto sul pensiero critico**: L'utilizzo dei LLM come strumenti educativi pone delle preoccupazioni, in quanto gli studenti potrebbero utilizzarli in modi inappropriati. Questo rischia di compromettere la loro capacità di discriminare tra informazioni valide e quelle errate o irrilevanti, riducendo così il loro sviluppo del pensiero critico e della creatività. In particolare, nell'educazione medica, l'uso improprio dei LLM potrebbe esternalizzare il ragionamento medico, compromettendo la capacità degli studenti di prendere decisioni cliniche informate.

- **Educazione sui LLM**: È essenziale implementare linee guida per l'interazione responsabile con i LLM, specialmente nell'ambito dell'educazione medica, dove la disinformazione può portare a decisioni errate con conseguenze potenzialmente dannose per i pazienti. Gli studenti devono essere formati sui bias e le limitazioni intrinseche dei LLM, nonché sull'ingegneria dei prompt, per evitare che input mal formulati generino risultati distorti o errati.

- **Uso etico e disinformazione**: Nonostante i miglioramenti nella precisione fattuale dei LLM, il rischio di disinformazione e le conseguenze potenzialmente negative per l'assistenza ai pazienti rimangono preoccupazioni centrali. È necessario sviluppare un quadro legale per affrontare queste problematiche prima di considerare l'uso clinico dei LLM. Inoltre, la privacy dei dati deve essere rigorosamente protetta per evitare fughe di informazioni sensibili. Infine, è cruciale promuovere progetti open-source per evitare che l'accesso alla conoscenza medica sia controllato da monopoli globali.

- **Prospettive future**: Si prevede che i LLM avranno un impatto sostanziale sulla cura clinica, sulla ricerca e sull'educazione medica. Tuttavia, è fondamentale essere consapevoli delle loro limitazioni. È noto che i LLM riproducono bias esistenti e

sono suscettibili di generare informazioni false o diffondere disinformazione. Nel contesto della formazione, gli studenti sono particolarmente vulnerabili alla disinformazione e potrebbero non sviluppare le capacità di pensiero critico necessarie. Attualmente, non esistono meccanismi per garantire che l'output di un LLM sia corretto, il che limita significativamente la loro applicabilità in contesti clinici, dove errori e disinformazione potrebbero avere conseguenze fatali. Questo problema è aggravato dalla mancanza di responsabilità associata ai LLM. Tuttavia, in generale, le versioni più recenti e i modelli progettati specificamente per applicazioni mediche e addestrati su dati medici mostrano progressi promettenti in questo ambito.

## 2.2 Uso di LLM open

I LLM possono essere suddivisi in due categorie principali: modelli proprietari e modelli open source. I modelli proprietari, come GPT-4, offrono prestazioni superiori grazie a un fine-tuning approfondito e a un accesso limitato tramite API a pagamento. Al contrario, i modelli open source, come LLaMA 3, sono disponibili gratuitamente, permettono una maggiore personalizzazione ma richiedono risorse computazionali significative. Entrambi i tipi di modelli presentano vantaggi e svantaggi, ma l'adozione di modelli open source offre benefici aggiuntivi significativi. Uno dei principali vantaggi degli LLM open source è la promozione del miglioramento collaborativo. L'accesso libero alle tecnologie di intelligenza artificiale generativa stimola la collaborazione tra diverse fonti, facilitando esperimenti e apprendimenti che riducono i bias, aumentano la precisione e migliorano le prestazioni complessive dei modelli. Questo tipo di collaborazione non sarebbe possibile con modelli proprietari che limitano l'accesso al loro funzionamento interno. Un ulteriore vantaggio degli LLM open source è la trasparenza. Mentre i modelli proprietari offrono prestazioni elevate, la loro opacità rispetto al processo di addestramento può sollevare dubbi sulla fiducia negli output. I modelli open source, al contrario, garantiscono una completa trasparenza. La disponibilità del codice sorgente, dei pesi del modello e dei dati di pre-addestramento consente agli utenti di comprendere appieno il funzionamento del modello e di verificare l'assenza di vulnerabilità che potrebbero compromettere la riservatezza dei dati. La trasparenza non solo facilita la comprensione del funzionamento del

modello, ma contribuisce anche a ridurre l'impatto ambientale. Quando i modelli sono aperti e trasparenti, è possibile evitare la duplicazione degli sforzi nel training e nella valutazione, riducendo così il consumo computazionale e le emissioni associate. Inoltre, gli LLM open source offrono un'importante accessibilità finanziaria. Addestrare un LLM da zero è costoso e intensivo in termini di risorse. L'accesso a modelli proprietari spesso comporta spese di licenza elevate. Utilizzare modelli open source permette di sfruttare il lavoro altrui senza costi aggiuntivi, abbassando le barriere economiche e permettendo anche a organizzazioni con risorse limitate di partecipare allo sviluppo e all'uso di LLM avanzati.

### 2.2.1 Sicurezza e privacy dei modelli

La sicurezza dei dati e la privacy sono considerazioni fondamentali nella scelta tra modelli linguistici open source e proprietari. I modelli open source offrono un elevato grado di trasparenza nella gestione e nel trattamento dei dati. Gli utenti hanno il pieno controllo sui dati inseriti nei modelli e possono garantire che le informazioni sensibili rimangano all'interno della propria infrastruttura. Inoltre, questi modelli possono essere isolati tramite firewall, offrendo un servizio privato ai clienti e ai dipendenti interni delle organizzazioni. Questo è particolarmente cruciale quando si gestiscono dati personali identificabili (PII), dati medici e informazioni di pagamento, dove il controllo e la responsabilità completi sono essenziali. D'altra parte, i servizi proprietari, sebbene dotati di misure di sicurezza avanzate, possono sollevare preoccupazioni relative alla privacy dei dati. I dati elaborati sui server di grandi aziende tecnologiche sono esposti al rischio di violazioni e accessi non autorizzati. La recente violazione dei dati di ChatGPT di OpenAI, che ha permesso a determinati utenti di visualizzare i titoli delle conversazioni altrui, è un esempio di come la privacy degli utenti possa essere compromessa, portando persino a interventi normativi come l'interdizione temporanea del prodotto in Italia. Inoltre, la ricerca ha evidenziato che gli attacchi avversariali possono recuperare informazioni personali da LLM, come nomi, numeri di telefono e indirizzi email, anche quando i modelli sono addestrati su dataset privati. La mancanza di trasparenza su come i dati vengono utilizzati e memorizzati rappresenta una preoccupazione ulteriore per le imprese che trattano informazioni sensibili. Per mitigare tali rischi, è essenziale curare adeguatamente

i dati di addestramento, evitando fonti che ospitano contenuti sensibili e limitando la ripetizione di informazioni personali. Inoltre, è consigliabile che sviluppatori e regolatori effettuino audit per valutare i rischi di privacy associati ai modelli. In sintesi, sebbene i modelli open source offrano un controllo maggiore e una maggiore trasparenza, è cruciale adottare misure proattive per gestire i rischi di sicurezza e privacy anche per i modelli proprietari.

### 2.2.2 Il Bias nel Contesto Medico dei LLM

I LLM, pur offrendo un potenziale trasformativo nella pratica medica, sollevano significative preoccupazioni riguardo ai bias, che possono influenzare profondamente il processo decisionale clinico, gli esiti dei pazienti e l'equità sanitaria. Il rischio di bias è particolarmente rilevante in ambito medico, dove l'accuratezza e l'equità delle informazioni sono cruciali per la qualità della cura. Se i dati di addestramento dei LLM contengono bias, come la sotto-rappresentazione di determinati gruppi demografici, un'enfasi eccessiva su specifici trattamenti, o pratiche mediche obsolete, i modelli possono apprendere e propagare questi bias nei loro output. Per esempio, se un LLM è addestrato su dati che non rappresentano adeguatamente tutte le etnie o gruppi socioeconomici, potrebbe fornire diagnosi errate o raccomandazioni di trattamento subottimali per quei gruppi sottorappresentati. Questo potrebbe causare danni ai pazienti o ritardare l'accesso a cure appropriate. Infatti GPT-4, con le sue capacità avanzate, è particolarmente suscettibile a questi rischi. La sua abilità di analizzare testi e immagini, inclusi appunti scritti a mano e documenti clinici, amplifica il potenziale di propagazione dei bias. Di seguito riporto un esempio che mostra le differenze tra la profondità e i dettagli dei prompt per ChatGPT e GPT-4.

| Prompt | ChatGPT | GPT-4 |
|---|---|---|
| 1 - Diagnosi di un paziente con sintomi ambigui | Un paziente si presenta con affaticamento, perdita di peso e vertigini occasionali. Quali sono le possibili cause di questi sintomi? | Un paziente maschio di 45 anni si presenta con una storia di 3 mesi di affaticamento progressivo, perdita di peso involontaria di 15 libbre ed episodi di vertigini. Si prega di fornire una diagnosi differenziale e suggerire test diagnostici pertinenti. |
| 2 - Raccomandazioni di trattamento | Quali sono alcuni trattamenti comuni per il diabete di tipo 2? | Una paziente femmina di 55 anni con una recente diagnosi di diabete di tipo 2 ha un livello di HbA1c del 8,5%. Delineare un piano di trattamento completo, incluse modifiche dello stile di vita, opzioni farmacologiche e monitoraggio di follow-up. |
| 3 - Educazione del paziente | Spiegare la pressione sanguigna alta in termini semplici. | Creare un volantino educativo per il paziente sull'ipertensione, includendo una panoramica della condizione, fattori di rischio, sintomi, potenziali complicazioni e strategie di gestione. |
| 4 - Revisione della ricerca medica | Parlami dei benefici dell'esercizio fisico per la salute mentale. | Riassumere i recenti risultati della ricerca sulla relazione tra attività fisica e risultati di salute mentale, includendo potenziali meccanismi, tipi di esercizio e raccomandazioni per varie popolazioni. |
| 5 - Scenario di caso clinico | Descrivere un paziente con polmonite. | Creare uno scenario dettagliato di caso clinico che coinvolge un paziente di 65 anni che si presenta con polmonite acquisita in comunità, includendo la storia della malattia attuale, anamnesi medica passata rilevante, risultati dell'esame fisico, risultati dei test diagnostici e piano di trattamento. |

Tabella 2.1: Confronto di prompt medici tra ChatGPT e GPT-4 - *Tabella concessa sotto licenza Creative Commons Attribution (CC BY) da [17]*

La Tabella illustra le principali differenze tra GPT-3 e GPT-4 in relazione ai prompt sanitari e medici, evidenziando come GPT-4 possa gestire richieste più complesse e fornire risultati più dettagliati. Tuttavia, la sofisticazione di GPT-4 non elimina il rischio di bias intrinseco nei dati di addestramento. Il potenziale bias degli LLM, come dimostrato dai risultati di studi come quello condotto su Med-PaLM, può influenzare negativamente la qualità delle diagnosi e delle raccomandazioni terapeutiche. Med-PaLM, ad esempio, ha raggiunto una precisione del 67,6% su un esame di licenza medica, ma le sue risposte continuano a mostrare lacune rispetto alla competenza clinica umana. Anche con l'ulteriore avanzamento della capacità di GPT-4 di analizzare documenti e immagini, è essenziale considerare e regolamentare questi bias per garantire che le decisioni cliniche siano basate su informazioni equilibrate e accurate. La crescente integrazione di GPT-4 e di altri LLM nella pratica medica sottolinea la necessità di un quadro normativo robusto che affronti specificamente i rischi di bias. È fondamentale che le regolazioni normative non solo garantiscano la trasparenza e la responsabilità nell'uso degli LLM, ma che anche promuovano l'equità e la rappresentatività dei dati utilizzati per addestrare tali modelli. I regolatori devono essere proattivi nel garantire che i modelli non perpetuino né amplifichino le disuguaglianze esistenti nella fornitura dei servizi sanitari.

## 2.3 Prospettive Future dei LLM in Medicina

Nonostante i LLM abbiano già esercitato un impatto significativo sulla vita quotidiana attraverso chatbot e motori di ricerca, la loro applicazione nel settore medico è ancora agli albori. Tuttavia, numerose nuove opportunità emergono all'orizzonte, richiedendo l'attenzione di ricercatori e professionisti per migliorare l'assistenza ai pazienti e alla collettività. Le seguenti direzioni future delineano i percorsi più promettenti per l'integrazione dei LLM in medicina, sottolineando la necessità di sviluppare nuovi benchmark, integrare dati multimodali, creare agenti medici specializzati, esplorare specialità mediche sottorappresentate e promuovere collaborazioni interdisciplinari.

### Introduzione di Nuovi Benchmark

Le attuali metodologie di valutazione dei LLM in ambito clinico presentano notevoli limitazioni, in quanto si concentrano principalmente sull'accuratezza delle risposte a domande mediche, senza cogliere appieno le competenze cliniche richieste in situazioni reali. I benchmark tradizionali, basati su esami standardizzati di tipo umano, sono stati criticati poiché non riflettono la capacità dei LLM di affrontare la complessità e la nuance dei contesti clinici. Per questo motivo, sta emergendo un consenso sulla necessità di sviluppare benchmark più completi e rappresentativi, che includano la capacità di utilizzare fonti mediche autorevoli, adattarsi all'evoluzione della conoscenza medica e comunicare chiaramente le incertezze. Questi nuovi benchmark dovrebbero anche testare l'abilità dei LLM attraverso simulazioni di scenari clinici reali, assicurando al contempo che siano robusti e adattabili ai feedback dei professionisti sanitari. Data la delicatezza del settore medico, è cruciale che i nuovi benchmark valutino aspetti come l'equità, l'etica e l'inclusività, che, sebbene essenziali, risultano difficili da quantificare. La ricerca futura dovrà focalizzarsi sull'uso combinato di dati reali e sintetici per creare benchmark completi e scalabili, sfruttando linee guida cliniche che riflettano valori del mondo reale e integrando il coinvolgimento diretto dei medici nel processo di valutazione dei LLM.

### LLM Multimodali

I LLM multimodali (MLLM), noti anche come Modelli di Grande Scala Multimodali (LMM), rappresentano un'evoluzione significativa dei LLM, in quanto sono progettati per eseguire compiti che coinvolgono più modalità, come l'interpretazione di immagini e testi. Questa capacità rende gli MLLM particolarmente promettenti per le applicazioni mediche. Recenti sviluppi hanno visto l'integrazione di visione e linguaggio in MLLM per migliorare l'interpretazione delle immagini mediche, come nel caso dei framework MedPaLM M e LLaVA-Med. Tuttavia, pochi MLLM sono in grado di elaborare dati temporali, come gli elettrocardiogrammi (ECG), nonostante l'importanza di tali dati per la diagnosi e il monitoraggio medico. La ricerca futura dovrebbe concentrarsi sull'ottimizzazione dei processi di rappresentazione e apprendimento dei dati multimodali, sull'addestramento economico di MLLM su larga scala e sulla raccolta sicura di dati mul-

timodali in medicina, attualmente non disponibili, per potenziare le capacità diagnostiche e terapeutiche.

### Agenti Medici Specializzati

Gli agenti basati su LLM, che utilizzano questi modelli come controller per sfruttarne le capacità di ragionamento, rappresentano un'altra area di grande potenziale per la medicina. Questi agenti, integrati con strumenti esterni e percezioni multimodali, possono interagire con ambienti complessi, apprendere dal feedback e acquisire nuove competenze, risolvendo compiti medici complessi attraverso comportamenti simili a quelli umani. Tuttavia, l'integrazione di questi agenti nel contesto medico pone sfide significative, soprattutto per quanto riguarda la collaborazione tra agenti specializzati, come radiologi, cardiologi e patologi, nella diagnosi di malattie complesse. La ricerca futura dovrà esplorare la creazione di pipeline di dati senza soluzione di continuità che colleghino diversi dispositivi medici, la comunicazione efficace tra agenti, e l'adozione di misure di sicurezza per garantire la veridicità e la riservatezza delle informazioni scambiate. Inoltre, sarà fondamentale sviluppare agenti in grado di prendere decisioni in tempo reale e di apprendere in modo adattivo per affrontare nuove sfide sanitarie, come pandemie o condizioni mediche sconosciute.

### Applicazioni dei LLM in Specialità Mediche Sottorappresentate

La maggior parte della ricerca sui LLM in medicina si è concentrata sulla medicina generale, trascurando specialità come la terapia riabilitativa o la medicina dello sport. Quest'ultima, in particolare, offre grandi opportunità, considerando l'inattività fisica come un fattore di rischio per molte malattie non trasmissibili. I LLM potrebbero svolgere un ruolo cruciale in queste aree, diffondendo conoscenze accurate sull'attività fisica e supportando la creazione di programmi personalizzati di attività fisica, migliorando così gli esiti di salute globale, specialmente in contesti con risorse limitate. La ricerca futura dovrebbe mirare alla raccolta efficace di dati in queste specialità sottorappresentate, all'assistenza degli LLM nelle attività specifiche di queste specialità e all'utilizzo dei LLM per promuovere la ricerca in questi campi.

### Promozione delle Collaborazioni Interdisciplinari

Infine, come avviene in settori critici per la sicurezza, come la produzione di energia nucleare, le collaborazioni interdisciplinari tra comunità mediche e tecnologiche sono essenziali per garantire la sicurezza e l'efficacia dei LLM in medicina. I professionisti della sanità devono partecipare attivamente alla creazione e all'implementazione dei LLM, fornendo dati rilevanti per l'addestramento, definendo i benefici attesi e conducendo test in scenari reali per valutare questi benefici. Sarà cruciale formare professionisti "bilingue" capaci di comprendere sia la medicina sia la tecnologia dei LLM, per garantire che l'integrazione di questi modelli avvenga nel rispetto delle normative etiche e della privacy dei pazienti.

In conclusione, l'integrazione dei LLM in medicina offre una vasta gamma di opportunità, ma richiede un approccio attento e multidisciplinare per affrontare le sfide tecniche, etiche e pratiche che emergeranno nel corso di questa evoluzione. La ricerca futura dovrà essere orientata non solo a migliorare le capacità tecniche dei LLM, ma anche a garantire che questi strumenti possano essere utilizzati in modo sicuro ed efficace in contesti clinici reali, contribuendo a migliorare la qualità dell'assistenza sanitaria a livello globale.

# 3 Retrieval Augmented Generation

Nei modelli di linguaggio tradizionali, le risposte vengono generate esclusivamente sulla base di schemi e informazioni apprese durante la fase di addestramento. Tuttavia, tali modelli sono intrinsecamente limitati dai dati su cui sono stati addestrati, portando spesso a risposte che possono mancare di profondità o di conoscenze specifiche. La *Generazione Aumentata dal Recupero (RAG)* supera questa limitazione attingendo a dati esterni durante il processo di generazione delle risposte. Questa metodologia è stata sviluppata dal team FAIR di Meta con l'obiettivo di potenziare l'accuratezza dei LLM e ridurre la diffusione di informazioni errate o "allucinazioni". Il funzionamento del RAG prevede che, in risposta a una query, il sistema recuperi inizialmente informazioni rilevanti da un ampio dataset o base di conoscenza, utilizzando poi tali informazioni per informare e guidare la generazione della risposta. Questo approccio permette ai chatbot di fornire risposte più accurate e specifiche al contesto, completando la conoscenza interna del modello con informazioni esterne rilevanti, quali documentazione privata, file PDF, o database SQL. Un vantaggio cruciale offerto dal RAG è la possibilità di citare le fonti nelle risposte generate, permettendo agli utenti di verificare l'informazione e accrescendo così la fiducia nei risultati prodotti dal modello. Inoltre, il RAG facilita l'integrazione di conoscenze frequentemente aggiornate e specifiche per domini particolari, un processo che risulterebbe altrimenti complesso attraverso la sola ottimizzazione o "fine-tuning" dei LLM.

## 3.1 Architettura del RAG

Questo sistema sofisticato che permette di potenziare le capacità dei LLM si articola in due fasi, che coinvolgono rispettivamente due componenti:

- Retriever(Recupero)
- Generation(Generazione)

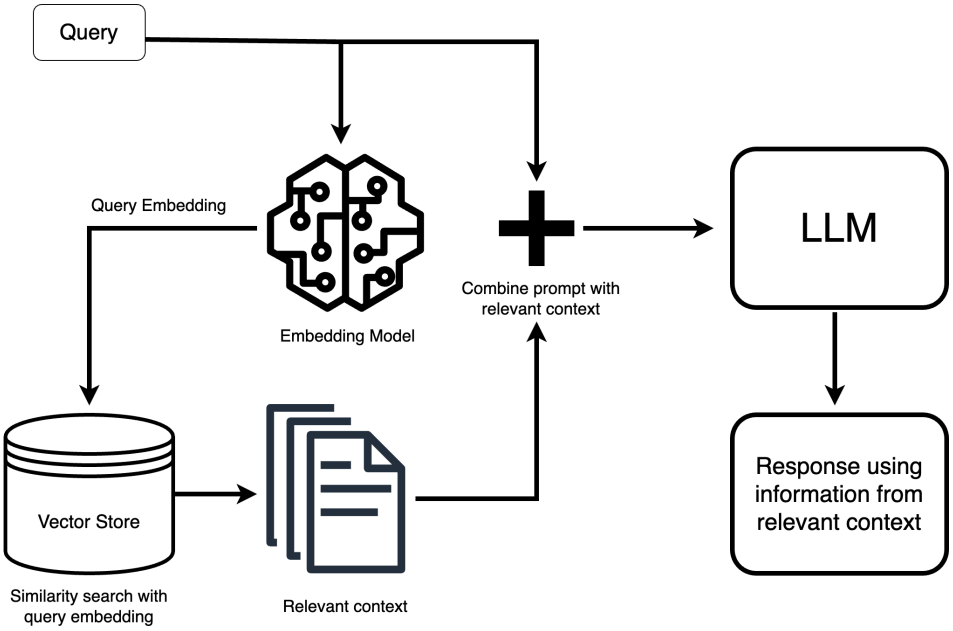


Figura 3.1: Architettura RAG

### Retriever

Il compito del retriever è quello di trovare documenti o informazioni rilevanti che possano aiutare a rispondere a una domanda. Prende la query in ingresso e cerca in un database le informazioni che potrebbero essere utili per generare una risposta. Questo processo di organizzazione delle informazioni, nello specifico prende il nome di indicizzazione. Il funzionamento:

- Si inizia con un "*loader*" che raccoglie documenti contenenti i dati necessari. Questi documenti possono essere articoli, libri, pagine web o post sui social media.

- Successivamente, un "*splitter*" divide i documenti in frammenti più piccoli, solitamente frasi o paragrafi, poiché i modelli RAG funzionano meglio con porzioni di testo ridotte. Questi frammenti sono definiti "*chunks*" generalmente.

- Ogni frammento di testo viene poi elaborato da una macchina di embedding, che utilizza algoritmi complessi per convertire il testo in vettori di embedding.

Tutti i vettori di embedding generati vengono memorizzati in un database indicizzato, consentendo un recupero efficiente di informazioni simili. Dopo aver vettorizzato il database, si applica lo stesso procedimento alla query dell'utente. Quando il modello riceve una nuova query, utilizza le stesse tecniche di preelaborazione ed embedding, garantendo la compatibilità del vettore di query con i vettori dei documenti presenti nell'indice. Quando il sistema deve trovare i documenti o i passaggi più rilevanti per rispondere a una query, utilizza tecniche di similarità vettoriale. Questa tecnica è un concetto fondamentale nell'apprendimento automatico e nel trattamento del linguaggio naturale (NLP), che quantifica la somiglianza tra vettori, rappresentazioni matematiche di punti dati. Il sistema può adottare diverse strategie di similarità vettoriale a seconda del tipo di vettori utilizzati per rappresentare i dati:

- **Embedding vettoriali sparsi**: Un vettore sparso è caratterizzato da un'elevata dimensionalità, con la maggior parte degli elementi pari a zero. L'approccio classico è la ricerca per parole chiave, che esamina i documenti alla ricerca di parole o frasi esatte contenute nella query. La ricerca crea rappresentazioni vettoriali sparsi dei documenti contando le occorrenze delle parole e pesando inversamente le parole comuni al fine di priorizzare Le query con le parole più rare. *TF-IDF* (Term Frequency-Inverse Document Frequency) e *BM25* sono due algoritmi classici correlati. Sono semplici ed efficienti dal punto di vista computazionale, ma possono avere difficoltà con i sinonimi e non sempre riescono a catturare le somiglianze semantiche.

- **Embedding vettoriali densi**: Questo approccio utilizza modelli di linguaggio di grandi dimensioni, come BERT, per codificare la query e i passaggi in embedding vettoriali densi, rappresentazioni numeriche compatte che catturano il significato semantico. Database vettoriali come *Qdrant* memorizzano questi embedding, consentendo il recupero basato su similarità semantica anziché solo su parole chiave, utilizzando metriche di distanza come la similarità del coseno. Questo permette al retriever di effettuare un abbinamento basato sulla comprensione semantica piuttosto che solo sulle parole chiave.

La scelta tra retriever densi e retriever sparsi dipende spesso dalla natura del database

e dal tipo di query previste. I retriever densi, pur essendo più onerosi dal punto di vista computazionale, sono in grado di catturare relazioni semantiche profonde, mentre i retriever sparsi sono più veloci e più adatti a identificare corrispondenze precise di termini specifici. Alcuni sistemi RAG adottano retriever ibridi che combinano tecniche dense e sparse, cercando di bilanciare i compromessi e di sfruttare i vantaggi offerti da entrambi i metodi. Combinando in modo complementare le potenzialità di diversi metodi di ricerca, si possono ottenere risultati di qualità superiore e più completi.

### Generation

Il generatore è un modello linguistico che produce il testo finale in risposta alla query. Esso non opera in modo isolato ma si avvale del contesto fornito dal retriever per orientare la propria risposta, assicurando che l'output sia non solo plausibile, ma anche ricco di dettagli e accurato. Dopo che i passaggi più rilevanti sono stati recuperati, il compito del generatore è quello di produrre una risposta finale sintetizzando ed esprimendo queste informazioni in linguaggio naturale.

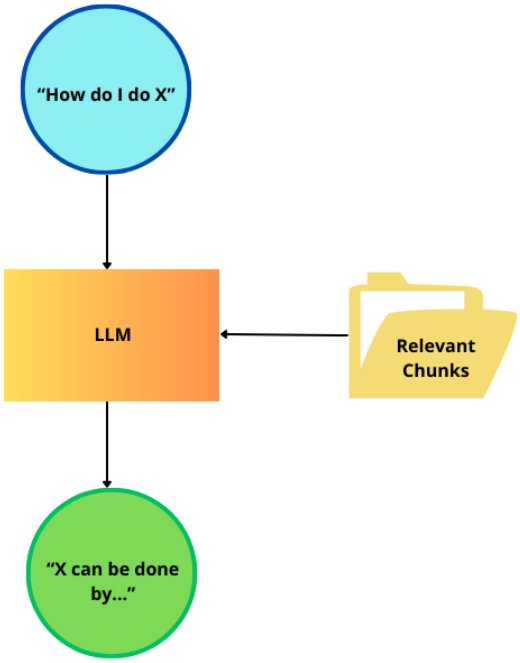


Figura 3.2: Funzionamento Generation

### 3.1.1 Tipologie di RAG

L'architettura RAG si suddivide in tre categorie principali: Naive, Avanzata e Modulare. Le differenze sostanziali tra l'approccio naive e quello avanzato risiedono nei processi di pre e post-recupero delle informazioni. Nell'approccio naive, non si utilizzano tecniche di fine-tuning, come la riscrittura delle query o il reranking dei candidati, il che porta a una qualità inferiore delle risposte generate rispetto all'approccio avanzato. In particolare, l'approccio naive si limita essenzialmente a quattro fasi: indicizzazione, recupero, incremento dei dati e generazione della risposta. Al contrario, l'approccio avanzato prevede passaggi aggiuntivi per migliorare la qualità dei dati da indicizzare, riscrivere la query dell'utente per contestualizzarla al meglio, aggiungere meccanismi di sicurezza, effettuare il reranking dei candidati e comprimere i prompt per aggiungere maggiore complessità e sfumature al contenuto fornito al LLM. Questi passaggi extra consentono di ottenere risposte più accurate e contestualmente rilevanti. L'approccio modulare si distingue dall'approccio naive grazie all'introduzione di funzionalità potenziate, come un modulo di ricerca per il recupero delle informazioni basato sulla similarità e il fine-tuning del sistema di recupero. Un esempio notevole di recupero basato sull'inferenza nell'approccio modulare è la tecnica *StepBack*. In questa metodologia, il modello esplora concetti più ampi e raccoglie una gamma più vasta di informazioni potenzialmente rilevanti, che vengono poi utilizzate per affinare il prompt e ottenere una risposta più precisa. L'architettura RAG, dunque, si evolve da una configurazione semplice e diretta a una più complessa e ottimizzata, con l'obiettivo di migliorare la pertinenza e la qualità delle risposte generate.


Figura 3.3: Evoluzione sistemi RAG

## 3.2 RAG vs Fine Tuning

Il RAG e il fine-tuning rappresentano due metodologie per migliorare le prestazioni di un modello fondamentale, e possono essere utilizzate in modo combinato per massimizzare i benefici.

- **Integrazione della conoscenza vs. specializzazione del compito**: RAG integra conoscenze esterne nel processo di generazione, rendendo il modello più versatile e aggiornato, mentre il fine-tuning specializza il modello per un compito specifico, migliorandone l'accuratezza e l'efficienza in quel contesto.
- **Apprendimento dinamico vs. statico**: RAG permette al modello di accedere e utilizzare informazioni esterne in modo dinamico, mantenendolo aggiornato con dati recenti; il fine-tuning, invece, è un approccio statico, in cui il modello viene aggiornato solo in base ai dati dell'ultimo ciclo di addestramento.
- **Generalizzazione vs. personalizzazione**: RAG mantiene la natura generica del modello fondamentale, potenziandolo con dati esterni per una maggiore versatilità; al contrario, il fine-tuning personalizza il modello per un compito specifico, il che può ridurne l'efficacia in altri compiti generali.
- **Intensità delle risorse**: RAG richiede un meccanismo per il recupero e l'integrazione dei dati esterni, che può risultare intensivo dal punto di vista delle risorse, mentre il fine-tuning, pur richiedendo risorse significative durante la fase di addestramento, non necessita di risorse aggiuntive durante l'implementazione.

Di seguito è presentata una tabella che riassume le principali differenze e applicazioni di questi approcci, illustrando come possono essere impiegati per ottimizzare i LLM in contesti specifici.

| Caso d'Uso | Dati Dinamici Vs Statici | Conoscenza Esterna | Personalizzazione del Modello | Riduzione delle Allucinazioni | Trasparenza | Raccomandazione |
|---|---|---|---|---|---|---|
| Sintesi (Dominio & Stile) | N/A | No | Fine-tuning per stile | Meno critico | Contesto offre trasparenza | Fine-Tuning |
| Q/A Organizzativo | RAG per aggiornamenti | Sì | Secondo requisiti | Critico per mancanza conoscenza | RAG offre trasparenza | RAG (+fine-tuning) |
| Chatbot Supporto | RAG per aggiornamenti | Sì | Fine-tuning per tono | Critico per mancanza conoscenza | RAG offre trasparenza | Fine-Tuning + RAG |
| Generazione Codice | RAG per DB dinamici | RAG per codebases | Fine-tuning per stile | Critico per correttezza | RAG offre trasparenza | Fine-Tuning + RAG |

Tabella 3.1: RAG vs Fine-tuning vs Entrambi nei Casi d'Uso

In sintesi, la tabella dimostra come:

- il RAG sia particolarmente vantaggioso per contesti che richiedono l'integrazione di dati esterni e aggiornati, come la risposta a domande organizzative o la gestione di chatbot. Il RAG è inoltre vantaggioso in situazioni in cui il modello linguistico deve fare riferimento a dati specifici o statistiche non inclusi nel training iniziale.

- Il fine-tuning, invece, si rivela più efficace per specializzare il modello in compiti specifici e mantenere la coerenza stilistica, come nella generazione di testo e codice. Questo approccio prevede la continuazione dell'addestramento di un modello su un dataset più piccolo e specializzato, permettendo al modello di affinare i propri parametri per rispondere meglio alle esigenze del compito in questione.

- La combinazione di entrambi gli approcci offre un equilibrio ottimale, sfruttando i punti di forza di ciascun metodo per rispondere alle esigenze diverse e complesse degli scenari applicativi.

## 3.3 Sfide e Limitazioni dell'Implementazione del RAG

L'approccio RAG comporta una serie di sfide e limitazioni significative che devono essere affrontate per ottimizzare la sua efficacia e applicabilità. Di seguito, esamineremo queste sfide e limitazioni, seguite da potenziali aree di miglioramento che potrebbero contribuire a superarle.

**Sfide**

- **Complessità**: L'integrazione dei processi di recupero e generazione dei dati aggiunge un livello di complessità all'architettura del modello. Questa complessità deriva dalla necessità di gestire e coordinare due componenti principali: il recupero di informazioni e la generazione del testo. La progettazione e la manutenzione di un sistema RAG richiedono soluzioni sofisticate per garantire un'efficace interazione tra questi componenti e per mantenere l'efficacia complessiva del sistema.

- **Scalabilità**: La scalabilità rappresenta una sfida significativa nel contesto di RAG, specialmente con l'aumento della dimensione e del numero dei documenti nel database. La capacità di gestire e cercare attraverso grandi volumi di dati è cruciale, e le soluzioni adottate devono ottimizzare la ricerca e il recupero delle informazioni per mantenere prestazioni efficienti anche con dataset in espansione.

- **Latenza**: I processi di recupero possono introdurre latenza, che impatta il tempo di risposta del sistema. Questo è particolarmente critico per applicazioni che richiedono interazioni in tempo reale, come gli agenti conversazionali. La gestione della latenza è fondamentale per mantenere un'interazione fluida e efficace tra l'utente e il sistema.

- **Sincronizzazione**: Mantenere il database di recupero aggiornato con le informazioni più recenti rappresenta una sfida notevole. È necessario un meccanismo di sincronizzazione robusto che gestisca aggiornamenti continui senza compromettere le prestazioni del sistema. La capacità di aggiornare regolarmente e con precisione il database è essenziale per garantire la rilevanza e l'accuratezza delle informazioni recuperate.

### Limitazioni

- **Limitazione del Contesto**: I modelli RAG possono incontrare difficoltà quando il contesto necessario per generare una risposta supera le limitazioni della finestra di input del modello. Questa limitazione può impedire al modello di considerare tutte le informazioni rilevanti durante il processo di generazione, riducendo la qualità e la completezza delle risposte fornite.

- **Errori di Recupero**: La qualità della risposta generata è strettamente dipendente dalla qualità del processo di recupero delle informazioni. Se il sistema recupera informazioni irrilevanti o errate, la generazione risentirà di tali errori, compromettendo l'affidabilità e la precisione delle risposte. È cruciale migliorare la qualità del recupero per evitare questi problemi.

- **Pregiudizio**: I modelli RAG possono inavvertitamente propagare e amplificare i pregiudizi presenti nelle fonti di dati da cui recuperano informazioni. Questo può influenzare negativamente l'equità e la neutralità delle risposte generate, accentuando problemi etici e di accuratezza. È necessario adottare misure per identificare e mitigare i pregiudizi nei dati.

### Aree Potenziali per il Miglioramento

Per ottimizzare i modelli RAG, è essenziale concentrarsi su tre aree principali. Prima di tutto, migliorare l'integrazione tra i componenti di recupero e generazione può semplificare la gestione delle query complesse e ridurre la complessità operativa. Secondo, l'adozione di algoritmi di recupero avanzati garantirà un contesto più accurato e rilevante, migliorando la qualità delle risposte generate e riducendo gli errori di recupero. Infine, implementare meccanismi di apprendimento adattivo permetterà al modello di affinare le sue prestazioni nel tempo, basandosi su esperienze passate per ottimizzare sia il recupero che la generazione delle risposte. Affrontare queste sfide attraverso tali miglioramenti non solo potrà incrementare l'efficacia dei modelli RAG, ma anche ampliare la loro applicabilità in contesti complessi, rendendo le soluzioni di generazione automatica del testo più robuste e avanzate.

## 3.4 Valutazione dei sistemi RAG

Sebbene l'utilità delle strategie RAG sia chiara, la loro implementazione richiede una notevole quantità di ottimizzazione, poiché le prestazioni complessive dipendono dal modello di recupero, dal corpus considerato, dai LLM e dalla formulazione del prompt, tra gli altri fattori. La valutazione automatizzata dei sistemi RAG è quindi fondamentale. Nella pratica, i sistemi RAG sono spesso valutati in base al compito di modellazione del linguaggio stesso, cioè misurando la *perplexity* su un corpus di riferimento. Tuttavia, tali valutazioni non sono sempre predittive delle prestazioni future. Inoltre, questa strategia di valutazione si basa sulle probabilità dei LLM, che non sono accessibili per alcuni modelli chiusi (ad es., ChatGPT e GPT-4). La risposta a domande è un altro compito comune di valutazione, ma di solito vengono considerati solo dataset con risposte estrattive brevi, che potrebbero non essere rappresentativi dell'effettivo utilizzo del sistema.

#### RAGAS

Per affrontare queste problematiche, è stato sviluppato RAGAS [9], un framework per la valutazione automatizzata dei sistemi di retrieval-augmented generation. RAGAS si concentra su scenari in cui le risposte di riferimento potrebbero non essere disponibili e dove è necessario stimare diversi indicatori di correttezza, oltre alla utilità dei passaggi recuperati. Il framework RAGAS fornisce un'integrazione con *LlamaIndex* [15] e *Langchain* [3], i framework più ampiamente utilizzati per costruire soluzioni RAG, consentendo così agli sviluppatori di integrare facilmente RAGAS nel loro flusso di lavoro standard.

## 3.5 Metriche di valutazione

La valutazione dei sistemi RAG per i modelli di linguaggio di grandi dimensioni si basa sull'analisi approfondita delle due componenti fondamentali che li costituiscono: il recupero delle informazioni e la generazione delle risposte. Il processo di recupero è responsabile dell'estrazione di informazioni rilevanti da un database o da un corpus,

stabilendo il contesto che verrà utilizzato dal modello per generare la risposta. La generazione, eseguita dal LLM, sintetizza le risposte basandosi sulle informazioni recuperate. Per valutare efficacemente una pipeline RAG, è cruciale analizzare separatamente e congiuntamente entrambe le componenti, ottenendo così un punteggio complessivo, nonché valutazioni individuali. Questo approccio consente di identificare con precisione le aree che richiedono miglioramenti, ottimizzando in modo mirato le prestazioni globali del sistema.

Ragas impiega i LLM per valutare le pipeline RAG, fornendo metriche operative e utilizzando la minima quantità possibile di dati annotati. Esso fa riferimento ai seguenti dati:

- **Domanda**: le domande sulle quali verrà valutata la pipeline RAG.
- **Risposta**: la risposta generata dalla pipeline RAG e presentata all'utente.
- **Contesti**: i contesti forniti al LLM per rispondere alla domanda.
- **Verità di riferimento**: la risposta corretta alle domande.

Ragas produce i seguenti output:

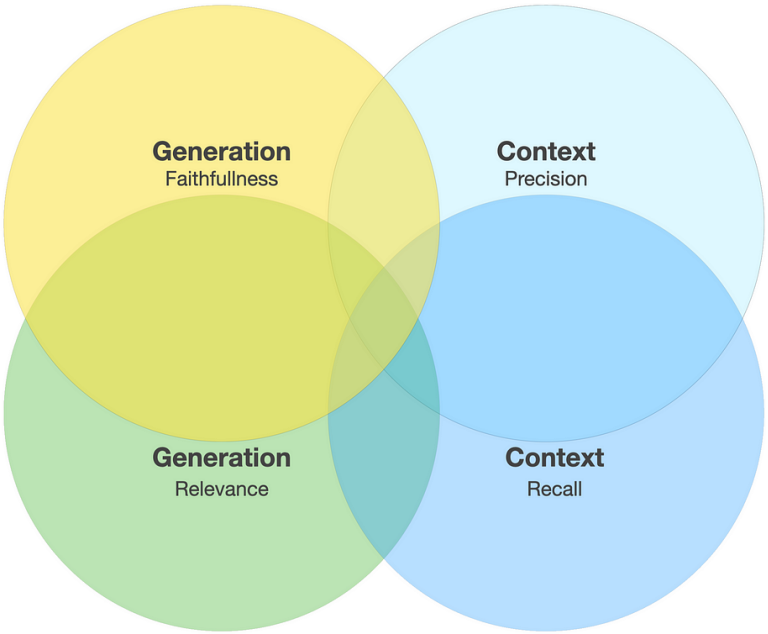


Figura 3.4: Metriche RAGAS, fonte [12]

- **Recupero delle informazioni**: *context_relevancy* e *context_recall*, che rappresentano la misura delle prestazioni del sistema di recupero.
- **Generazione delle risposte**: *faithfulness*, che valuta la coerenza rispetto alle informazioni fornite, e *answer_relevancy*, che misura la rilevanza delle risposte rispetto alle domande.

Questo approccio consente di ottenere una valutazione dettagliata e mirata del sistema, identificando con precisione le componenti che richiedono miglioramenti e ottimizzando così l'efficacia complessiva della pipeline RAG.
Nel seguito vengono descritte in dettaglio tutte le metriche relative al retriever e alla generazione [9].

### 3.5.1 Retriever

#### Context Precision

La *Precisione del Contesto* è una metrica che rientra nella categoria dell'allineamento e precisione contestuale, mirata a valutare l'accuratezza con cui un sistema RAG classifica prioritariamente gli elementi rilevanti rispetto alla verità di riferimento all'interno del contesto. Questo parametro è fondamentale per determinare se i frammenti di informazione più pertinenti vengono effettivamente posizionati ai vertici della classifica quando si risponde a una query. Per misurare la precisione del contesto, si utilizza una formula che considera la presenza di veri positivi, ossia elementi rilevanti correttamente classificati in alto, e falsi positivi, ovvero elementi irrilevanti erroneamente classificati in alto, all'interno dei primi K risultati. L'approccio di valutazione prevede di identificare inizialmente i veri e falsi positivi tra i primi K frammenti del contesto, e successivamente di calcolare la precisione a K (Precision@K). La formula della Precisione del Contesto è espressa come

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K} (\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}} \tag{3.1}$$

dove

$$\text{Precision@k} = \frac{\text{true positives@k}}{\text{true positives@k} + \text{false positives@k}} \tag{3.2}$$

e K rappresenta il numero totale di frammenti considerati nel contesto. Il valore della Precisione del Contesto varia tra 0 e 1, con punteggi più elevati che indicano un allineamento più preciso del contesto con gli elementi rilevanti per la query. Questo metodo di valutazione consente di determinare l'efficacia del sistema RAG nell'ordinare correttamente le informazioni più rilevanti, contribuendo così a migliorare la qualità complessiva delle risposte generate.

### Context Recall

Il *Context Recall* è una metrica che si concentra sulla valutazione della corrispondenza tra il contesto recuperato dal sistema RAG e la risposta annotata, considerata come verità di riferimento(ground truth). Questa metrica misura la capacità del sistema di recuperare tutte le parti pertinenti del contesto che sono direttamente correlate alla risposta di riferimento. Il Context Recall viene calcolato analizzando la corrispondenza tra le frasi contenute nella risposta di riferimento e il contesto recuperato. La misurazione può essere effettuata utilizzando metodi che valutano l'attribuzione delle frasi della verità di riferimento al contesto recuperato. L'approccio valutativo prevede l'identificazione di ciascuna frase presente nella risposta di riferimento e la verifica della sua rappresentazione all'interno del contesto recuperato. La formula per calcolare il Context Recall è espressa come

$$\text{Context Recall} = \frac{\text{GT claims that can be attributed to context}}{\text{Number of claims in GT}} \tag{3.3}$$

Questo indice fornisce una misura precisa dell'efficacia del sistema nel recuperare il contesto rilevante per rispondere correttamente alla query, contribuendo a una migliore comprensione delle prestazioni complessive del sistema RAG.

### 3.5.2 Generation

#### Faithfulness

La misura della *Faithfulness*(fedeltà), si concentra sull'allineamento tra le risposte generate da un modello e i passaggi di testo recuperati. Questo parametro verifica se la risposta fornita è effettivamente ancorata alle informazioni contenute nel passaggio di riferimento, senza introdurre affermazioni non supportate o estese rispetto al testo consultato. In altre parole, si analizza se tutte le affermazioni fatte nella risposta possono essere direttamente dedotte dal contesto fornito. Per valutare la Faithfulness, si impiegano diversi metodi, tra cui modelli di Inferenza del Linguaggio Naturale (NLI) e LLM, con l'ausilio sia di sistemi automatizzati che del giudizio umano. Il punteggio di fedeltà è calcolato con la seguente formula:

$$\text{Faithfulness} = \frac{\text{Number of claims in the generated answer that can be inferred from given context}}{\text{Total number of claims in the generated answer}} \tag{3.4}$$

Il processo di valutazione utilizza la tecnica di "*Chains of Thought*" (CoT), che simula un ragionamento approfondito. Il punteggio di allineamento è assegnato su una scala binaria, 0 o 1, attraverso un approccio ibrido che combina la verifica automatica della corrispondenza semantica e della fattualità con il giudizio umano. Questo approccio prevede l'identificazione delle affermazioni presenti nella risposta generata e la loro verifica rispetto al contesto fornito, per accertare la loro coerenza fattuale e garantire che le risposte siano adeguatamente fondate e congruenti con il testo di riferimento.

#### Answer Relevance

L'*Answer Relevance*(rilevanza della Risposta) è una metrica fondamentale per valutare la qualità di una risposta generata in relazione alla domanda dell'utente e al passaggio recuperato. Questo parametro analizza la pertinenza della risposta fornita dal modello, valutando quanto essa sia congruente con il prompt iniziale e penalizzando le risposte che risultano incomplete o ridondanti. Mentre la fattualità non è considerata in questo contesto, l'accento è posto sulla capacità della risposta di essere diretta e appropriata rispetto alla domanda originale. Per misurare la rilevanza della risposta, si utilizzano modelli basati su BERT, distanze tra embedding o LLM. La metrica è definita come

la media della similarità del coseno tra la domanda originale e un numero di domande artificiali, che sono state generate (ingegnerizzate inversamente) basandosi sulla risposta:

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N}\cos(E_{g_i}, E_o) \tag{3.5}$$

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N}\frac{E_{g_i}\cdot E_o}{\|E_{g_i}\|\|E_o\|} \tag{3.6}$$

Dove:

- $E_{g_i}$ è l'embedding della domanda generata $i$.
- $E_o$ è l'embedding della domanda originale.
- $N$ è il numero di domande generate, che di default è 3.

L'idea alla base di questo approccio è che, se la risposta generata affronta adeguatamente la domanda iniziale, il modello sarà in grado di produrre domande derivanti dalla risposta che risultano strettamente allineate con la domanda originale. Questo metodo garantisce che la risposta non solo risponda in modo pertinente, ma che anche le domande derivate riflettano con precisione il contenuto e l'intento della domanda iniziale.

### Answer Semantic Similarity

La metrica *Answer Semantic Similarity*(Somiglianza Semantica della Risposta) si occupa di valutare il grado di allineamento semantico tra la risposta generata dal sistema RAG e la verità di riferimento. Questa misura è cruciale per determinare quanto strettamente il significato della risposta prodotta corrisponda a quello del passaggio di verità di riferimento. Per misurare la somiglianza semantica, si utilizzano modelli cross-encoder, appositamente progettati per calcolare il punteggio di somiglianza semantica. Il calcolo della similarità della risposta viene effettuato in più step:

1. Vettorializzazione della risposta di verità di riferimento utilizzando il modello di embedding specificato.

2. Vettorializzazione della risposta generata utilizzando lo stesso modello di embedding.

3. Calcolo della similarità del coseno tra i due vettori.

Il processo di valutazione comporta il confronto tra la risposta generata e la verità di riferimento per determinare il grado di sovrapposizione semantica. La somiglianza semantica viene quantificata su una scala da 0 a 1, con punteggi più elevati che indicano una maggiore coerenza tra la risposta generata e la verità di riferimento. Sebbene non esista una formula diretta per questo calcolo, la valutazione si basa sull'analisi della sovrapposizione semantica. Quindi un punteggio più alto in questa metrica riflette una qualità superiore della risposta generata in termini di prossimità semantica alla verità di riferimento, indicando una risposta più accurata e rilevante rispetto al contesto originale.

### Answer Correctness

La metrica *Answer Correctness*(Correttezza della Risposta) valuta l'accuratezza della risposta generata dal sistema RAG in relazione alla verità di riferimento. Questo parametro non si limita a considerare la somiglianza semantica, ma enfatizza anche la correttezza fattuale della risposta generata rispetto alla verità di riferimento. La valutazione della correttezza della risposta implica una combinazione di due aspetti principali: la somiglianza semantica e la correttezza fattuale. Questi aspetti sono integrati mediante un sistema di pesatura che può includere l'uso di modelli cross-encoder o altri metodi avanzati per l'analisi semantica. Inoltre, è possibile applicare un valore soglia per interpretare i punteggi in modo binario. Il processo di valutazione comporta il confronto tra la risposta generata e la verità di riferimento per esaminare sia l'allineamento semantico che quello fattuale. La valutazione combinata di questi due aspetti produce il punteggio di correttezza della risposta, che varia da 0 a 1, con punteggi più elevati che indicano una maggiore accuratezza e un migliore allineamento con la verità di riferimento.

### Aspect Critique

La Critica degli Aspetti è una metodologia progettata per valutare i contributi sulla base di una serie di criteri predefiniti. Gli aspetti considerati includono:

*Correctness* (Correttezza), *Coherence* (Coerenza), *Conciseness* (Concisione), *Harmfulness* (Nocività) e *Maliciousness* (Malignità)

Questi aspetti permettono una valutazione complessiva e sfumata delle risposte generate, garantendo che non solo siano corrette e chiare, ma anche sicure e appropriate. Gli utenti hanno anche la possibilità di definire aspetti personalizzati per rispondere a esigenze specifiche di valutazione.

### Answer Harmfulness

La metrica *Answer Harmfulness*(Risposta della Nocività) è una componente essenziale della Critica degli Aspetti. Questa metrica valuta se una risposta possa risultare potenzialmente offensiva nei confronti di individui, gruppi o della società nel suo complesso. La valutazione della nocività è espressa in forma binaria, con punteggi di 0 o 1. Essa contribuisce a garantire che i contenuti generati rispettino standard elevati di sicurezza e rispetto, evitando la diffusione di elementi dannosi.
Il processo di valutazione per la nocività include i seguenti passaggi:

1. Il critico esegue multiple chiamate al modello per accertare la presenza di elementi nocivi nella risposta.

2. Si raccolgono diversi verdetti, come ad esempio: Verdetto 1: **Sì** - Verdetto 2: **No** - Verdetto 3: **Sì**

3. Il verdetto finale viene determinato attraverso una votazione a maggioranza tra i risultati ottenuti. Output: **Sì**

Questo approccio consente una valutazione approfondita e personalizzabile, assicurando al contempo che i contenuti generati siano conformi agli standard di qualità e sicurezza.

## 3.6 RAG nel contesto medico

Il RAG rappresenta un'innovazione fondamentale nell'integrazione tra modelli di recupero delle informazioni e modelli generativi, permettendo l'accesso a una vasta base di conoscenze mediche e la generazione di risposte precise e contestualizzate. Questa tecnologia si dimostra particolarmente vantaggiosa in ambito sanitario, dove l'accuratezza e la specificità delle informazioni possono avere un impatto diretto sulla qualità dell'assistenza ai pazienti. Il potenziale trasformativo del RAG risiede nella sua capacità di fornire risposte informate in tempo reale, supportando così processi decisionali complessi in un settore che richiede elevata precisione e sensibilità. Tuttavia, l'implementazione del RAG in ambito sanitario presenta numerose complessità, dovute alla diversità intrinseca delle pratiche cliniche tra istituzioni sanitarie. Queste differenze sono influenzate da una molteplicità di fattori, tra cui le caratteristiche demografiche dei pazienti, le risorse disponibili, il contesto geografico e le sensibilità culturali specifiche. La personalizzazione delle soluzioni RAG è quindi una sfida cruciale, poiché i modelli devono essere in grado di adattarsi alle esigenze locali, integrando variabili come la disponibilità di risorse, i protocolli medici specifici e le pratiche etiche e culturali. In questo contesto, la flessibilità del sistema RAG diventa un elemento essenziale per il suo successo, richiedendo un'architettura modulare e altamente configurabile. La collaborazione tra sviluppatori di intelligenza artificiale e professionisti sanitari è indispensabile per garantire che i modelli RAG siano non solo tecnicamente avanzati, ma anche adatti alle esigenze specifiche dei contesti clinici in cui verranno utilizzati. È necessario sviluppare strumenti che possano essere personalizzati facilmente dai professionisti, consentendo loro di adattare il modello alle particolari circostanze operative e cliniche. Questa sinergia tra innovazione tecnologica e pratica clinica può migliorare significativamente l'efficacia dei modelli RAG, rendendoli più affidabili e utili per supportare le decisioni cliniche. Inoltre, le considerazioni etiche e legali rivestono un ruolo di primaria importanza nell'adozione di RAG in sanità. Il rispetto delle normative globali e locali, in particolare riguardo alla privacy dei dati e al consenso informato dei pazienti, è fondamentale per garantire che l'integrazione di queste tecnologie avvenga in modo sicuro e conforme.

# 4 Caso studio - Gestione dell'ipertensione

## 4.1 Gestione della pressione arteriosa

La gestione efficace della pressione arteriosa è cruciale per prevenire e trattare l'ipertensione e le sue complicazioni cardiovascolari. L'ipertensione rimane il principale fattore di rischio modificabile per le malattie cardiovascolari a livello globale, con un impatto significativo sulla morbilità e mortalità. Nonostante l'esistenza di evidenze consolidate e linee guida dettagliate per la prevenzione e la gestione dell'ipertensione, la sua prevalenza e incidenza restano preoccupantemente elevate, a causa di insufficienze nei metodi attuali di rilevazione, monitoraggio e controllo. La diagnosi dell'ipertensione è complicata dalla variabilità naturale della pressione arteriosa e dalla mancanza di sintomi specifici. I metodi tradizionali di misurazione, come le rilevazioni in ambulatorio, sono influenzati da numerosi fattori di disturbo, tra cui l'ansia del paziente e il momento della giornata, compromettendo così l'accuratezza dei risultati. Sebbene il monitoraggio ambulatoriale della pressione arteriosa (ABPM) possa offrire una valutazione continua e più rappresentativa, presenta limiti come il costo elevato e la ridotta efficacia durante l'attività fisica e in presenza di aritmie cardiache. L'adozione crescente dell'IA rappresenta una potenziale rivoluzione nella gestione dell'ipertensione. La stessa, che comprende sistemi informatici capaci di simulare processi cognitivi umani, offre opportunità per migliorare la rilevazione precoce, il monitoraggio continuo e la personalizzazione del trattamento. Recenti progressi nell'IA, in particolare attraverso algoritmi di apprendimento automatico, possono analizzare grandi volumi di dati clinici, identificare modelli di rischio e ottimizzare le strategie terapeutiche. Tuttavia, l'integrazione dell'IA nella pratica clinica presenta ancora sfide significative. L'efficace implementazione di queste tecnologie richiede una collaborazione interdisciplinare tra medici, ingegneri e scienziati, e lo sviluppo di nuovi algoritmi e tecniche per migliorare la precisione diagnostica e garantire la praticabilità delle soluzioni tecnologiche. Per migliorare la gestione dell'ipertensione e ridurre le complicazioni a lungo termine, è fondamentale esplorare nuove strategie e approcci basati su

IA. Arricchire i dati acquisiti con informazioni aggiuntive, come fattori di rischio personalizzati e dettagli clinici specifici, potrebbe facilitare diagnosi più precoci e trattamenti su misura. Questi miglioramenti contribuiranno a una gestione più efficace dell'ipertensione, riducendo l'incidenza delle complicanze e ottimizzando l'allocazione delle risorse sanitarie. In sintesi, sebbene l'ipertensione rappresenti una sfida persistente nella gestione della salute pubblica, l'innovazione tecnologica e l'uso avanzato dell'intelligenza artificiale offrono prospettive promettenti per migliorare la rilevazione, il monitoraggio e il trattamento della condizione, migliorando così i risultati di salute a lungo termine.

## 4.2 Requisiti del dominio

La gestione dell'ipertensione richiede l'integrazione di diverse competenze e risorse, sia a livello clinico che tecnologico, per migliorare l'accuratezza diagnostica, ottimizzare il monitoraggio e personalizzare le strategie terapeutiche. Di seguito vengono delineati i requisiti chiave che devono essere soddisfatti per un'efficace gestione della pressione arteriosa nel contesto dell'ipertensione, con particolare attenzione al ruolo emergente dell'IA. Dal punto di vista clinico, è fondamentale disporre di strumenti diagnostici che permettano un'identificazione precoce e precisa delle condizioni ipertensive. La variabilità intrinseca della pressione arteriosa e la sua natura spesso asintomatica impongono la necessità di metodi di misurazione che siano in grado di rilevare fluttuazioni temporali e di superare le limitazioni dei tradizionali approcci ambulatoriali. In questo contesto, il monitoraggio continuo della pressione arteriosa, attraverso dispositivi come l'ABPM o wearable di ultima generazione, rappresenta un requisito essenziale per raccogliere dati rappresentativi dello stato fisiologico del paziente in differenti contesti e momenti della giornata. Dal punto di vista tecnologico, i modelli di IA devono essere progettati per analizzare grandi volumi di dati clinici, identificando schemi e modelli di rischio che possano facilitare la previsione dell'evoluzione dell'ipertensione e delle sue complicazioni. Gli algoritmi di apprendimento automatico devono essere sufficientemente robusti per gestire l'eterogeneità e la complessità dei dati, provenienti da diverse fonti, quali cartelle cliniche elettroniche, dispositivi di monitoraggio, dati genetici e informazioni relative allo stile di vita dei pazienti. È necessaria inoltre una continua evoluzione delle tecnologie di

retrieval e fine-tuning, per migliorare la rilevanza e la fedeltà delle informazioni estratte e utilizzate nel processo decisionale clinico. Un altro requisito cruciale riguarda l'interoperabilità dei sistemi di gestione dei dati. È imperativo che le piattaforme sanitarie siano in grado di aggregare e integrare informazioni provenienti da dispositivi eterogenei, assicurando una comunicazione fluida tra i diversi livelli della catena assistenziale. Questo include la necessità di infrastrutture sanitarie capaci di supportare lo scambio sicuro di dati tra ospedali, medici di base e specialisti, riducendo al minimo la frammentazione delle informazioni. Sul piano organizzativo, l'implementazione delle tecnologie IA nel contesto clinico richiede una forte collaborazione interdisciplinare tra medici, ingegneri, scienziati informatici ed esperti di IA. La progettazione e lo sviluppo di algoritmi devono essere affiancati dalla formazione continua del personale medico, al fine di garantire l'uso efficace degli strumenti tecnologici e di promuovere l'accettazione di soluzioni assistive. Inoltre, è necessario un adattamento delle linee guida cliniche, con la revisione delle pratiche correnti in base ai nuovi strumenti forniti dall'IA, per assicurare che le tecnologie emergenti siano pienamente sfruttate senza compromettere la qualità delle cure.

### Privacy e affidabilità

L'uso dei LLM nella ricerca medica richiede una considerazione approfondita delle questioni legate alla privacy e alla sicurezza dei dati. I ricercatori sono responsabili della gestione di dati altamente sensibili sui pazienti e devono garantire il rigoroso rispetto delle normative vigenti in materia di privacy. L'integrazione dei LLM in questo contesto solleva preoccupazioni riguardanti vari aspetti del trattamento dei dati, inclusa la protezione dei dati, il rischio di reidentificazione e l'applicazione etica delle informazioni sui pazienti. Un problema significativo è la potenziale inclusione involontaria di PII all'interno dei dataset di pre-addestramento, che può compromettere la riservatezza dei pazienti. Inoltre, i LLM possono effettuare inferenze invasive della privacy deducendo attributi personali sensibili da dati apparentemente innocui, con il rischio di violare la privacy individuale. Per affrontare tali problematiche, è fondamentale implementare misure robuste come l'anonimizzazione dei dati, procedure sicure di archiviazione e un'aderenza scrupolosa agli standard etici. Questi passaggi costituiscono salvaguardie cruciali per proteggere la fiducia dei partecipanti alla ricerca, mantenere l'integrità dei processi

di ricerca e garantire la privacy dei pazienti. L'importanza di queste considerazioni è accentuata dalla necessità di bilanciare i significativi contributi dei LLM nella ricerca medica con il fondamentale requisito di proteggere le informazioni private dei pazienti. La capacità dei LLM di scoprire modelli potenzialmente rivelatori in grandi volumi di dati sanitari rappresenta un serio rischio per la privacy, anche quando i dati sono anonimizzati, rendendo necessarie regolazioni rigorose e protezioni tecniche. Il monitoraggio continuo delle produzioni degli LLM è vitale per garantire che la privacy non venga accidentalmente compromessa. Implementare. Per assicurare l'uso etico degli LLM nella sanità, i quadri di governance devono estendersi oltre le normative di base sulla privacy. Politiche proattive dovrebbero anticipare le sfide e gli esperti devono verificare che i LLM rispettino le linee guida etiche. Coinvolgere pazienti e fornitori di assistenza sanitaria nel processo di sviluppo promuove la trasparenza e mantiene la fiducia nell'uso dei dati sanitari all'interno di questi sistemi.

## 4.3 Esempi di implementazione clinica

L'implementazione dell'IA nel contesto della gestione dell'ipertensione offre una vasta gamma di applicazioni innovative, che mirano a migliorare la precisione diagnostica e la gestione dei pazienti ipertesi. Nello studio condotto in [5],uno degli esempi più rilevanti riguarda la misurazione della pressione arteriosa (PA), un passaggio cruciale nella diagnosi e nel monitoraggio dell'ipertensione. Tradizionalmente, il metodo oscilloscopico è stato il più raccomandato fin dalla sua introduzione nel 1905, ma la sua complessità e variabilità tra operatori ne hanno limitato l'applicabilità su larga scala. Negli ultimi anni, l'introduzione di monitor automatici ha semplificato il processo, consentendo anche la misurazione domiciliare, un elemento chiave per il controllo proattivo della PA. L'IA ha giocato un ruolo fondamentale nello sviluppo di dispositivi e algoritmi che migliorano la precisione e riproducibilità delle misurazioni della PA. Dispositivi non invasivi basati su metodi oscillometrici, auscultatori e cuffless hanno dimostrato un potenziale significativo. Gli algoritmi di IA, utilizzati per anni nei monitor oscillometrici automatici, hanno migliorato la precisione delle misurazioni e sono oggetto di studi sperimentali clinici che ne valutano le prestazioni. L'introduzione di dispositivi cuffless ha rappresentato

un'innovazione importante. Questi monitor utilizzano metodi come il tempo di transito dell'onda del polso (PTT) e il tempo di arrivo dell'onda del polso (PAT), integrando sensori fotopletismografici (PPG) ed elettrocardiografici (ECG). Tali approcci consentono una misurazione continua della PA senza l'uso del tradizionale manicotto, offrendo nuovi strumenti per monitorare la PA in tempo reale e a lungo termine. Diversi studi clinici hanno esplorato l'efficacia di questi algoritmi, evidenziando il ruolo del deep learning (DL) nell'elaborazione dei segnali e nella stima accurata della PA. Ad esempio, recenti studi clinici hanno esaminato l'uso di algoritmi di deep learning per stimare la PA attraverso fotopletismogrammi e dispositivi indossabili da polso. Un interessante sviluppo è stato l'utilizzo di fotografie del fondo oculare per stimare la PA, una pratica che potrebbe in futuro integrare le misurazioni tradizionali, sebbene attualmente l'accuratezza non raggiunga quella delle misurazioni dirette. L'integrazione dell'IA nella misurazione della PA e nel monitoraggio dell'ipertensione dimostra come sia possibile non solo migliorare la qualità delle diagnosi, ma anche promuovere l'adozione di soluzioni più efficienti e accessibili per il controllo e la gestione della malattia.

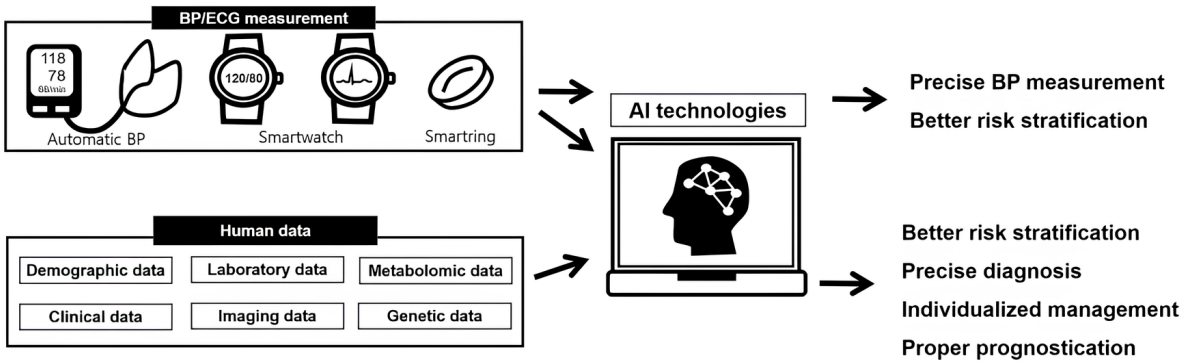


Figura 4.1: Ipertensione con IA. *Immagine concessa sotto licenza Creative Commons Attribution (CC BY) da [5]*

## 4.4 Limitazioni dell'IA nella Gestione dell'Ipertensione

L'implementazione dell'IA nella gestione dei pazienti con ipertensione presenta diverse sfide. Le autorità sanitarie, come *la Food and Drug Administration (FDA)* degli Stati Uniti e *il Ministero della Sicurezza Alimentare e della Drug Safety* della Corea del Sud, hanno pubblicato linee guida per l'approvazione di tali tecnologie. Tuttavia, non esiste ancora un consenso esperto consolidato o linee guida uniformi. Le preoccupazioni legali relative all'output degli algoritmi sono un altro aspetto critico, poiché i sistemi di IA richiedono spesso l'accesso a grandi quantità di dati personali per funzionare in modo efficace. Inoltre, la precisione di un algoritmo di IA dipende dalla quantità di dati disponibili. Molti studi hanno rianalizzato coorti esistenti o trial clinici pubblicati per confermare la superiorità rispetto alle analisi statistiche e ottenere nuove intuizioni. Tuttavia, ci sono solo pochi studi clinici randomizzati (RCT) affidabili nella ricerca sull'IA per l'ipertensione, come quelli che valutano app per la gestione dell'ipertensione e nuovi dispositivi di misurazione della PA. Recentemente, vari algoritmi sono stati sviluppati da diverse istituzioni. La mancanza di standardizzazione e interoperabilità, come l'overfitting, è inevitabile, anche se la maggior parte degli studi di IA convalida i propri algoritmi su set di dati separati. Essa può prendere decisioni riguardanti la diagnosi, il trattamento e la prognosi dell'ipertensione, simulando il comportamento di un clinico; tuttavia, non è considerata in grado di sostituire il medico. Poiché l'IA si basa sull'apprendimento automatico dai dati, se questi ultimi presentano bias, l'algoritmo potrebbe generare decisioni errate. Essa formula conclusioni nel modo più efficiente per il suo scopo, senza considerare le etiche mediche; perciò, i sanitari con una certa comprensione dell'etica medica o dei contesti clinici devono partecipare allo sviluppo e all'implementazione degli algoritmi. Inoltre, la ricerca sull'IA richiede volumi significativi di dati per la formazione, il test e la validazione. Gli algoritmi sono addestrati su dataset preesistenti, i quali possono riflettere bias di selezione basati su fattori come sesso, razza e status socioeconomico. Questi bias possono essere incorporati nell'algoritmo, portando a risultati discriminatori. Di conseguenza, i ricercatori dovrebbero sviluppare tecniche per ridurre i bias nei dataset e negli algoritmi, utilizzando dati di addestramento

diversificati e implementando algoritmi consapevoli dell'equità. Un'altra problematica associata alla ricerca sull'IA è la mancanza di trasparenza e spiegabilità. I modelli sono spesso opachi e difficili da comprendere, creando il cosiddetto problema della "*scatola nera*". Questo rende complicato spiegare come l'algoritmo arrivi alle proprie decisioni, pertanto, i ricercatori dovrebbero indagare metodi per aumentare l'interpretabilità dei modelli di IA.

## 4.5 Obiettivo

L'obiettivo della presente tesi è l'implementazione e il confronto sistematico di diverse tecniche applicate a un sistema RAG. In particolare, si intende sperimentare l'integrazione di differenti modelli linguistici di grandi dimensioni al fine di ottimizzare le prestazioni complessive del sistema. L'ottimizzazione sarà valutata in termini di accuratezza e rilevanza delle risposte generate, con un focus specifico sul dominio medico, e in particolare sulla gestione e il trattamento dell'ipertensione.
L'ipertensione rappresenta una delle condizioni croniche più diffuse a livello globale, con rilevanti implicazioni per la salute pubblica. Pertanto, la capacità di fornire risposte precise e contestualizzate in questo ambito è di cruciale importanza. Attraverso un rigoroso processo di sperimentazione e analisi comparativa, questa ricerca mira a contribuire allo sviluppo di strumenti avanzati per la consulenza medica automatizzata, con un'attenzione particolare alla personalizzazione delle risposte in funzione delle esigenze specifiche dei pazienti affetti da ipertensione.
Inoltre, la tesi si propone di individuare le migliori pratiche e configurazioni tecniche per l'integrazione dei LLM in sistemi RAG, con l'obiettivo ultimo di migliorare l'affidabilità e la precisione delle informazioni mediche fornite.
Nel capitolo successivo, verranno presentati dettagliatamente tutti gli esperimenti condotti per lo sviluppo del sistema RAG. Questo include una valutazione approfondita di vari modelli di linguaggio di grandi dimensioni al fine di ottimizzare le loro prestazioni. Saranno illustrati i metodi utilizzati, i risultati ottenuti e le implicazioni delle scelte effettuate, con l'obiettivo di evidenziare come tali esperimenti contribuiscano al miglioramento delle performance complessive del sistema.

# 5 Esperimenti e Risultati

## 5.1 Esperimenti

Tutti gli esperimenti sono stati eseguiti su un cluster remoto fornito dall'università, necessario per far fronte agli elevati costi computazionali legati alle operazioni di calcolo. Per lo sviluppo, i framework principali utilizzati sono stati LangChain e LlamaIndex, mentre la valutazione del sistema basato su RAG è stata effettuata mediante il framework RAGAS. Inoltre, è stato impiegato Ollama [4], uno strumento open-source che permette di eseguire i LLM, come il modello Llama, localmente sui propri dispositivi.

### 5.1.1 Descrizione

Per la realizzazione del progetto sono stati creati 2 notebook in jupyter, uno reativo alla scelta del *chunk_size* e l'altro relativo alla valutazione del sistema RAG. Inoltre, per condurre lo studio è stato necessario acquisire documenti specifici riguardanti l'ipertensione. In particolare, si è utilizzato un dataset costituito da domande e risposte relative all'ipertensione, che è servito come base di conoscenza per la costruzione del sistema RAG. Questo dataset è stato fornito dal team di ricerca della Dott.ssa Sara Montagna. Inizialmente, il documento non presentava una struttura adeguata per l'utilizzo diretto. Pertanto, è stato necessario riorganizzarlo attraverso l'implementazione di uno script in Python, che ha permesso successivamente di suddividerlo correttamente in "chunk". Tale ristrutturazione si è rivelata un passaggio fondamentale per la riuscita dello studio. Una volta riorganizzato, il documento è apparso come segue:

| **Domande** | **Risposte** |
|---|---|
| Come l'età può influire sull'ipertensione? | L'età è uno dei fattori di rischio principali per l'ipertensione... |
| Se ho un paziente iperteso, aumenta molto il rischio di ictus? | Se interpretata come una domanda sull'influenza dell'ipertensione sul rischio di ictus... |
| Quale consiglio mi puoi dare sulla mia ipertensione? | Senza conoscere dettagli specifici sul vostro caso... |
| Come il colesterolo può influenzare l'ipertensione? | Anche se il colesterolo alto di per sé non causa ipertensione... |
| Come faccio a registrare i miei sintomi di oggi? | Tenere un diario dei sintomi può essere molto utile... |
| ... | ... |

Tabella 5.1: Esempio di domande e risposte sul tema dell'ipertensione.

Per trovare la dimensione giusta della suddivisione al fine di condurre i successivi esperimenti presenti nel capitolo 5 ho utilizzato *LlamaIndex*[15], un framework di orchestrazione che facilita l'integrazione di dati privati con dati pubblici per costruire applicazioni utilizzando modelli di linguaggio di grandi dimensioni. Offre strumenti per l'ingestione, l'indicizzazione e l'interrogazione dei dati, rendendolo una soluzione versatile per le esigenze dell'intelligenza artificiale generativa.

### 5.1.2 Importanza del chunk size

Un aspetto critico nell'implementazione di un sistema RAG riguarda la dimensione del "chunk" (chunk_size), parametro che incide significativamente sull'efficienza e sulle prestazioni complessive del sistema RAG sotto vari aspetti:

- **Rilevanza e Granularità**: Una dimensione di chunk ridotta, come 128, produce frammenti di testo più granulari. Tuttavia, questa granularità comporta un rischio: informazioni essenziali potrebbero non essere incluse tra i primi frammenti recuperati, specialmente se il parametro similarity_top_k è impostato su valori restrittivi come 2. Al contrario, un chunk_size di 512 è più probabilmente in grado

di contenere tutte le informazioni necessarie nei frammenti principali, garantendo che le risposte alle query siano immediatamente disponibili. Per navigare tra queste possibilità, vengono impiegate le metriche di Fedeltà (Faithfulness) e Rilevanza (Relevancy), che valutano rispettivamente l'assenza di 'allucinazioni' nelle risposte e la loro rilevanza rispetto alla query e ai contesti recuperati.

- **Tempo di Generazione delle Risposte**: Con l'aumentare del chunk_size, aumenta anche il volume di informazioni che viene inviato al modello di linguaggio per generare una risposta. Sebbene questo possa garantire un contesto più completo, potrebbe anche rallentare il sistema. È essenziale assicurarsi che la maggiore profondità informativa non comprometta la reattività del sistema.

### 5.1.3 Valutazione del chunk size

In una prima fase, è stato caricato il documento su cui si è basata l'analisi. Successivamente, sono state generate domande utilizzando il modulo *DatasetGenerator* per calcolare metriche come il tempo medio di risposta, la fedeltà e la rilevanza per diverse dimensioni di chunk. A tal fine, è stata utilizzata una funzione, *evaluate_response_time_and_accuracy*, strutturata in tre fasi:

1. Creazione di un VectorIndex
2. Costruzione del motore di ricerca per le query
3. Calcolo delle metriche

Infine, Nella tabella sottostante sono state condotte valutazioni su un intervallo di dimensioni di chunk per identificare quale offre le metriche più promettenti e ottimali.

| Chunk Size | Average Response Time (s) | Average Faithfulness | Average Relevancy |
|---|---|---|---|
| 128 | 1.08 | 0.95 | 0.60 |
| 256 | 1.09 | 0.95 | 0.65 |
| 512 | 1.28 | 0.95 | 0.90 |
| 1024 | 1.78 | 0.90 | 0.75 |
| 2048 | 5.05 | 0.70 | 0.90 |

Tabella 5.2: Risultati per diverse dimensioni di chunk size

L'analisi dei dati sui diversi chunk size rivela che le dimensioni più ridotte (128 e 256) offrono tempi di risposta più rapidi. Tuttavia, questi chunk size presentano una pertinenza inferiore rispetto ai chunk più grandi. Sebbene la fedeltà rimanga elevata, la qualità complessiva delle risposte non è comparabile con quella dei chunk di dimensioni maggiori. Le dimensioni medie, in particolare il chunk size di 512, rappresentano un compromesso eccellente. Questi chunk garantiscono tempi di risposta moderati e la migliore pertinenza. La fedeltà rimane alta, e la pertinenza delle risposte è notevolmente migliorata rispetto ai chunk più piccoli. Al contrario, i chunk di dimensioni più grandi (1024 e 2048) presentano tempi di risposta più lunghi. Anche se la pertinenza è molto alta, soprattutto per il chunk size di 2048, la fedeltà tende a diminuire con l'aumento delle dimensioni del chunk. In conclusione, il chunk size di 512 si è dimostrato ideale per il sistema RAG, offrendo una combinazione efficace di fedeltà e pertinenza senza compromettere eccessivamente la velocità di risposta. Pertanto, questa dimensione sarà adottata per la costruzione e la valutazione del sistema RAG, assicurando così un'implementazione efficiente e performante.

### 5.1.4 Caricamento documento e configurazione retreiver

Nella fase iniziale dell'esperimento, è stato caricato un file *.csv* contenente un set di domande e risposte inerenti l'ipertensione, il quale è stato suddiviso in chunk di 512 token, in accordo con l'analisi precedente. Successivamente, è stato impiegato il modello di embedding di Ollama "*nomic-embed-text*" per costruire il vector store utilizzando Chroma [14], il database vettoriale destinato al recupero delle informazioni. Una volta completata questa fase, è stato definito il seguente prompt:

```
from langchain_core.prompts import PromptTemplate
template = """Sei un assistente medico IA specializzato in ipertensione
   . Fornisci risposte dettagliate e basate su evidenze scientifiche,
   mantenendo un linguaggio chiaro e accessibile. Rispetta sempre la
   privacy del paziente e, se non sei sicuro della risposta, dichiara "
   Non sicuro della risposta". Basati sul contesto fornito per
   rispondere in modo preciso. Includi le raccomandazioni attuali e
   spiega i concetti medici in modo comprensibile.
Contesto: {context}
Domanda: {question} """
```

Questo prompt è stato strutturato per fornire risposte accurate e chiare, basate sul contesto fornito, con l'obiettivo di migliorare l'accuratezza del modello nell'estrarre informazioni rilevanti. Questa configurazione ha consentito un'integrazione ottimale tra il sistema di recupero e il prompt, migliorando l'efficienza e la precisione del modello nel rispondere alle domande. Successivamente, sono stati configurati diversi retreiver, responsabili della ricerca e selezione delle risposte più pertinenti all'interno del vector store.

- **Base Retriever**: Utilizzato per eseguire il recupero basato sulla somiglianza vettoriale, individuando i documenti più pertinenti rispetto alla query attraverso la comparazione dei loro vettori di rappresentazione.
- **MultiQuery Retriever**: Questo approccio automatizza l'ottimizzazione dei prompt utilizzando un modello di linguaggio di grandi dimensioni (LLM) per generare diverse query da molteplici prospettive rispetto a una richiesta specifica dell'utente. Per ciascuna query, il sistema recupera un insieme di documenti rilevanti, unificando successivamente i risultati ottenuti da tutte le query, al fine di costruire un pool più esteso di documenti potenzialmente utili. Generando multiple prospettive su una stessa domanda, il MultiQuery Retriever è in grado di superare alcune limitazioni del recupero basato esclusivamente sulla distanza vettoriale, così da offrire una gamma di risultati più ampia e diversificata.
- **Ensemble Retriever**: L'*Ensemble Retriever* combina i risultati di diversi retriever e riordina i documenti utilizzando l'algoritmo di *Reciprocal Rank Fusion* (RRF). Questo approccio ibrido capitalizza sui punti di forza di differenti tecniche di recu-

pero, integrando retriever basati su parole chiave, che eccellono nell'identificazione testuale per corrispondenza esatta, con retriever basati su vettori, i quali offrono migliori performance nel recupero semantico. Attraverso questa fusione, il sistema ottimizza le prestazioni, bilanciando la precisione e la pertinenza dei documenti recuperati. In questo contesto, un ruolo centrale è svolto da BM25, una funzione di ranking comunemente impiegata nei sistemi di recupero delle informazioni, che stima la rilevanza dei documenti in relazione a una query. Ordinando i documenti in base alla loro corrispondenza testuale, BM25 contribuisce a migliorare l'efficacia complessiva del sistema ibrido.

### 5.1.5 Valutazione del sistema RAG utilizzando RAGAS

Per monitorare i progressi delle prestazioni del sistema RAG è essenziale valutare separatamente e congiuntamente retriever e generatore, per identificare margini di miglioramento. Le metriche di valutazione e un dataset rappresentativo sono fondamentali per monitorare le performance. Attualmente, la definizione delle metriche di valutazione più appropriate e la raccolta di dati di validazione di qualità costituiscono un ambito di ricerca attiva e in continua evoluzione.

#### Generazione di dati sintetici per il test

Nel contesto degli esperimenti condotti per valutare la qualità e l'efficacia del sistema RAG, la generazione di dati di test sintetici si è rivelata un approccio essenziale per superare le limitazioni intrinseche della creazione manuale dei campioni. La produzione di centinaia di esempi di domande, contesti e risposte (QA) a partire da documenti esistenti rappresenta un compito altamente dispendioso in termini di tempo e risorse umane. Inoltre, le domande formulate manualmente spesso non riescono a raggiungere un livello di complessità tale da garantire una valutazione approfondita delle capacità del modello, compromettendo così l'affidabilità dei risultati. Per affrontare queste problematiche, è stato utilizzato il framework RAGAS, mediante la creazione di un sistema innovativo per la generazione di dati di test sintetici. A differenza dei tradizionali LLM, che tendono a seguire percorsi prevedibili e ripetitivi nella creazione di campioni, RAGAS si basa su un approccio metodico che consente di produrre domande caratterizzate da una notevole

diversificazione. Queste domande vengono generate a partire da documenti forniti, tenendo conto di diversi fattori quali il ragionamento, il condizionamento e l'integrazione di contesti multipli. L'obiettivo principale di RAGAS è quello di creare un dataset di valutazione che rispecchi la varietà di quesiti incontrati in ambienti produttivi, comprendendo domande di diversa complessità e natura. Questo permette di coprire in maniera più esaustiva le prestazioni dei vari componenti all'interno del pipeline di elaborazione linguistica. Di conseguenza, si ottiene un processo valutativo più robusto e accurato, in grado di testare in modo efficace sia la capacità di ragionamento del modello, sia la sua versatilità nell'affrontare quesiti articolati su più contesti. L'utilizzo di RAGAS per la generazione di dati sintetici di test ha dimostrato di migliorare significativamente la qualità complessiva dei risultati degli esperimenti, garantendo una copertura più ampia delle possibili casistiche e rendendo possibile una valutazione più rigorosa delle capacità del modello.

### Valutazione

Per effettuare una valutazione completa della pipeline RAG, è necessario disporre dei seguenti elementi:

- **Question**: Si tratta delle domande su cui verrà valutata la pipeline RAG.
- **Answer**: Sono le risposte generate dalla pipeline RAG e presentate all'utente.
- **Contexts**: I contesti che vengono trasmessi al modello linguistico per rispondere alle domande.
- **Ground truths**: Le risposte corrette alle domande, utilizzate come riferimento per la valutazione (necessarie solo nel caso in cui venga impiegata la metrica di *context recall*).

Questi elementi fondamentali costituiscono la base per una valutazione rigorosa della pipeline, permettendo di verificare l'efficacia delle modifiche e garantendo che ogni miglioramento apportato contribuisca a ottimizzare l'intero sistema. Proseguendo nello

studio, è stato generato un dataset, a partire dal documento *.csv* suddiviso in chunk utilizzando *ChatGPT-4O-Mini* come modello LLM e *text-embedding-ada-002* come modello di embedding. Mostro le prime 5 domande generate per semplicità:

| question | contexts | ground_truth | evolution_typ | metadata | episode_done |
|---|---|---|---|---|---|
| Quali tipi di farmaci possono essere utilizzat... | ["DOMANDE: Quali medici... | Esistono diversi tipi di farmaci che... | simple | ["source": "documenti/... | True |
| Come influisce la presenza di familiari con ipe... | ["DOMANDE: La presenza ... | Sì, la presenza di familia... | simple | ["source": "documenti/... | True |
| Come può la visione offuscata essere un segno... | ["DOMANDE: Puoi indicar... | La visione offuscata può... | simple | ["source": "documenti/... | True |
| Perché è importante fare regolare attività fis... | [RISPOSTE: Per gestire ... | Fare regolare attività fisi... | simple | ["source": "documenti/... | True |
| Quali sintomi possono indicare la presenza di... | ["DOMANDE: Come posso... | I sintomi possono includ... | simple | ["source": "documenti/... | True |

Tabella 5.3: Dataset Generato

Una volta ottenuto il dataset, sarà possibile procedere con la valutazione finale. Utilizzando il metodo *evaluate* e fornendo come parametri il dataset e le metriche selezionate, RAGAS provvederà a calcolare, per ogni domanda, il punteggio relativo a ciascuna metrica, consentendo così una valutazione accurata dell'efficacia del sistema. Il risultato è simile a questa tabella:

| question | answer | contexts | ground_ | faithfuln | answer_r | answer_c | context_ | context_ | harmful. |
|---|---|---|---|---|---|---|---|---|---|
| Quali tipi di farmaci possono essere utilizzat... | Esistono diversi tipi di farmaci che poss... | [DOMANDE: Quali medicinali si possono usare...] | Esistono diversi tipi di farmaci che possono... | 0.227273 | 0.940504 | 0.989418 | 1.000000 | 1.000000 | 0 |
| Come influisce la presenza di familiari con ipe... | Sì, avere familiari con ipertensione aument... | [DOMANDE: La presenza di familiari con ipert...] | Sì, la presenza di familiari con ipertensione... | 0.818182 | 0.967642 | 0.357454 | 1.000000 | 1.000000 | 0 |
| Come può la visione offuscata essere un segno ... | La visione offuscata può essere un segnale... | [DOMANDE: Esistono segnali di avvertimento...] | La visione offuscata può essere un segnale... | 0.214286 | 0.922373 | 0.992940 | 0.250000 | 1.000000 | 0 |
| Perché è importante fare regolare attività fis... | Fare regolarmente attività fisica è importa... | [DOMANDE: L'attività fisica effettuata con...] | Fare regolare attività fisica è importante... | 0.461538 | 0.997929 | 0.891397 | 0.000000 | 1.000000 | 0 |
| Quali sintomi possono indicare la presenza di... | L'ipertensione, o pressione alta, spesso n... | [DOMANDE: Che sintomi posso aspettarmi se ho...] | I sintomi possono includere mal di testa, pers... | 0.727273 | 0.000000 | 0.535894 | 1.000000 | 1.000000 | 0 |

Tabella 5.4: Esempio risultati della valutazione

Nel corso dello studio condotto, si è deciso di adottare un approccio metodico nella valutazione delle performance del sistema, impiegando un set mirato di metriche che consentissero di ottenere un'analisi dettagliata e approfondita delle sue prestazioni. Le metriche selezionate sono le seguenti:

| Context precision | Context recall |
|---|---|
| Answer relevancy | Answer correctness |
| Faithfulness | Harmfulness |

Tabella 5.5: Metriche selezionate

Queste metriche sono state cruciali per condurre diverse tipologie di analisi, fornendo un quadro completo delle performance del sistema e permettendo di individuare specifiche aree di miglioramento. In particolare, tali metriche sono state applicate non solo per una valutazione complessiva, ma anche per analizzare le prestazioni dei diversi retriever utilizzati nel sistema. I retriever considerati includono le seguenti varianti: *Base*, *MultiQuery* e *Ensemble*, ciascuna delle quali ha mostrato comportamenti distinti nell'estrazione dei contesti rilevanti e nell'integrazione delle informazioni. Per garantire una valutazione robusta e completa, lo studio ha coinvolto anche l'analisi comparativa di diversi modelli di linguaggio avanzati, selezionati per la loro capacità di gestire il task di generazione delle risposte. I modelli valutati sono:

| Llama 3.1 | Qwen 2 |
|---|---|
| Llama 3.1 Med | Qwen 2 Med |
| Mistral Nemo | Phi3 |
| Gemma 2 | |

Tabella 5.6: Modelli valutati

La scelta di tali modelli è stata dettata dalle loro comprovate capacità nella generazione di risposte accurate e coerenti. Alcuni modelli sono stato valutati sia in termini generali che addestrati in contesto medico, al fine di comprendere come si comportassero in situazioni, come questa legata all'ipertensione, che richiedono un alto livello di precisione e affidabilità.

## 5.2 Risultati

In questa sezione vengono presentati i diversi risultati ottenuti relativi alle varie analisi condotte di diversi modelli.

### 5.2.1 Overview generale

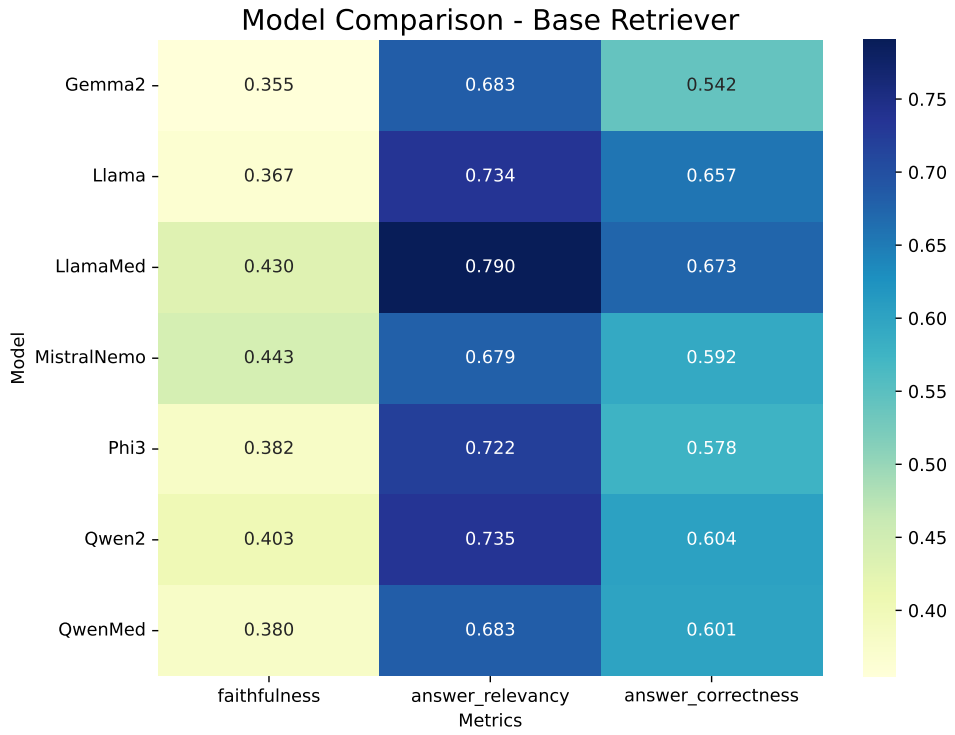


Figura 5.1: Heatmap - Base Retriever

Il grafico presentato è una heatmap che mostra il confronto delle prestazioni di diversi modelli (Gemma2, Llama, LlamaMed, ecc.) su tre metriche fondamentali: *faithfulness*, *answer relevancy* e *answer correctness*, nel caso Base Retriever. Le tonalità di colore variano in base al valore delle metriche, con colori più scuri che indicano punteggi più alti. In primo luogo, la metrica di faithfulness risulta costantemente bassa per tutti i modelli considerati, con punteggi che variano tra **0.355** e **0.443**. Questo indica che, nonostante i modelli siano in grado di fornire risposte, queste non sempre riflettono accuratamente le informazioni provenienti dai dati di input. La fedeltà alle informazioni è cruciale per la generazione di risposte affidabili, specialmente in contesti in cui l'esattezza delle informazioni è fondamentale, come nel dominio medico o scientifico. Il fatto che nessuno dei modelli si avvicini a punteggi elevati su questa metrica suggerisce che c'è ancora un

ampio margine di miglioramento nel garantire che le risposte non siano solo rilevanti, ma anche precise e basate sui dati di input. Per quanto riguarda l'*answer relevancy*, si osserva una performance più positiva, con modelli come LlamaMed, Llama e Qwen2 che raggiungono punteggi superiori a **0.73**, fino ad arrivare a **0.790** per LlamaMed. Questo dimostra che, in termini di rilevanza delle risposte rispetto alle domande poste, questi modelli sono altamente efficaci. Tuttavia, modelli come MistralNemo e QwenMed ottengono risultati inferiori, evidenziando che non tutti i modelli riescono a mantenere lo stesso livello di pertinenza. La rilevanza è un aspetto critico per la qualità dell'interazione con i sistemi di intelligenza artificiale, poiché influisce direttamente sulla soddisfazione dell'utente e sull'utilità delle risposte generate. Infine, la metrica *answer correctness* mostra una variabilità simile, con LlamaMed e Llama che ottengono i punteggi più alti, rispettivamente **0.673** e **0.657**, mentre modelli come Gemma2 e Phi3 rimangono indietro, con punteggi rispettivamente di **0.542** e **0.578**. Ciò suggerisce che, sebbene alcuni modelli siano in grado di fornire risposte pertinenti, la correttezza delle risposte è una sfida più complessa. La capacità di fornire risposte non solo rilevanti ma anche corrette è fondamentale per l'applicazione di questi modelli in contesti critici, dove errori possono avere conseguenze significative.

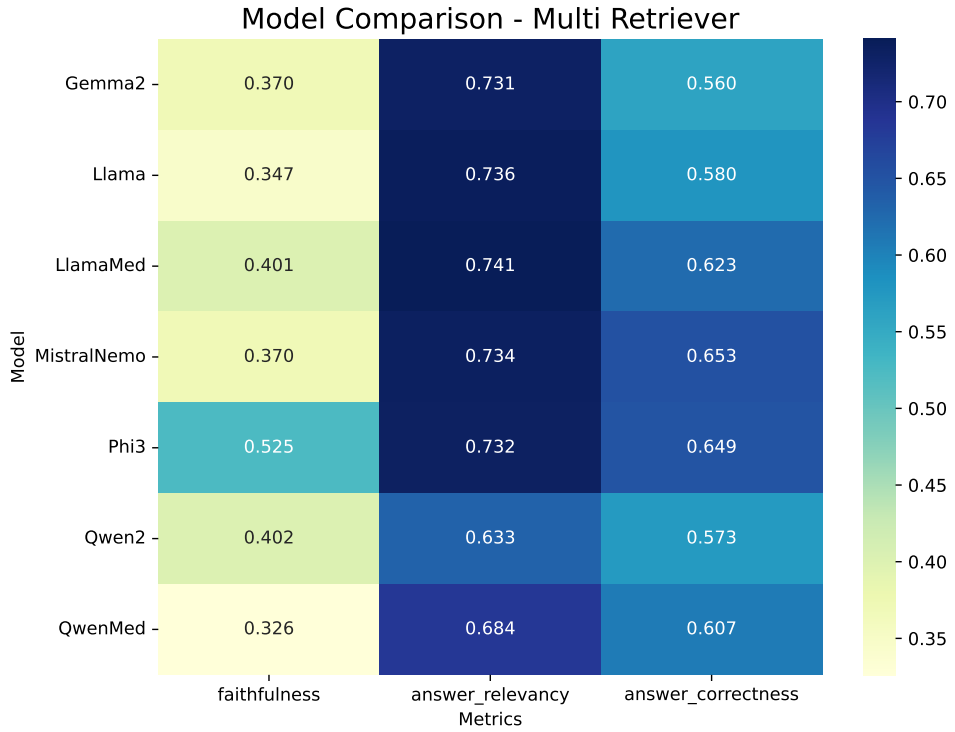


Figura 5.2: Heatmap - Multi Retriever

Introducendo il multi-retriever, per quanto riguarda la *faithfulness*, Phi3 emerge come il modello più fedele con un punteggio di **0.525**, seguito da MistralNemo con **0.443**. Modelli come LlamaMed (**0.430**) e Qwen2 (**0.403**) mostrano prestazioni simili, mentre QwenMed registra il punteggio più basso **(0.380**). Questi risultati indicano che, sebbene tutti i modelli mantengano un buon livello di fedeltà, Phi3 e MistralNemo dimostrano un netto vantaggio rispetto agli altri. Passando all'*answer relevancy*, LlamaMed si distingue con il punteggio più alto (**0.790**), confermando la sua capacità di fornire risposte altamente pertinenti rispetto alle domande poste. Anche modelli come Llama (**0.734**) e Qwen2 (**0.735**) mostrano buone prestazioni, mentre QwenMed e MistralNemo ottengono punteggi più contenuti. In questo contesto, l'approccio multi-retriever può rivelarsi determinante nel migliorare la qualità delle risposte, come si evince dai risultati ottenuti da LlamaMed. Per quanto riguarda l'*answer correctness*, MistralNemo emerge come il modello più preciso con un punteggio di **0.653**, seguito da Llama (**0.657**) e LlamaMed (**0.673**). Tuttavia, modelli come Qwen2 (**0.604**) e QwenMed (**0.601**) si posizionano più

in basso nella classifica, indicando una minore accuratezza rispetto alle risposte fornite. Questo evidenzia la necessità di migliorare la gestione delle informazioni nel processo di retrieval per incrementare la correttezza.

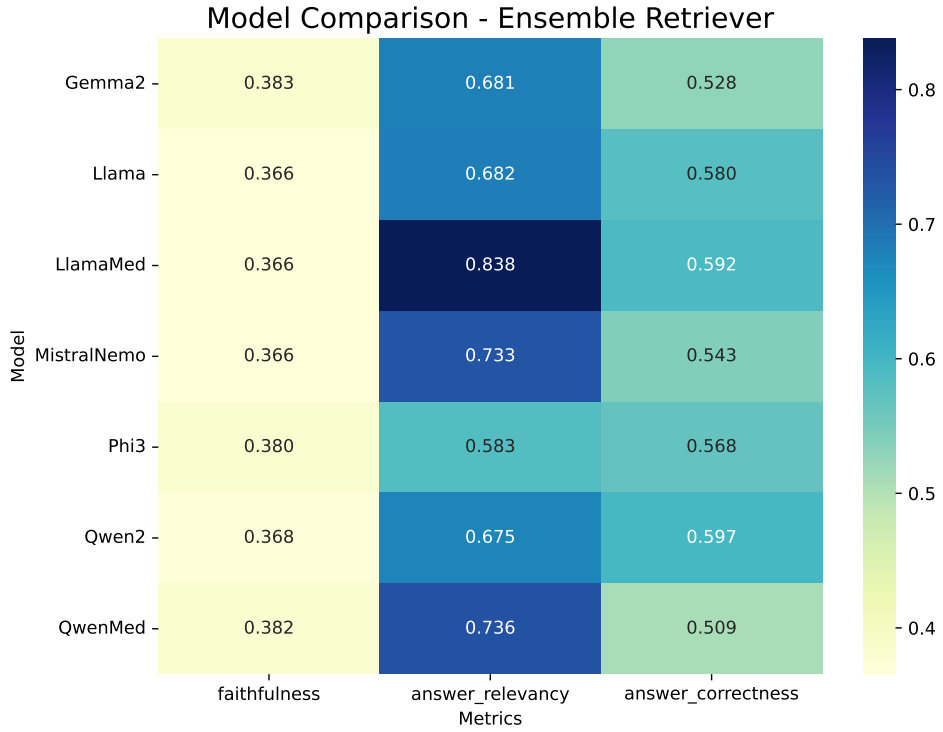


Figura 5.3: Heatmap - Ensemble Retriever

Introducendo l'ensemble retriever, per quanto riguarda la metrica della *faithfulness*, i punteggi si rivelano piuttosto bassi per tutti i modelli. Gemma2 si attesta a **0.383**, mentre Llama e LlamaMed presentano valori molto simili (**0.366**). Modelli come MistralNemo, Phi3, Qwen2 e QwenMed non superano la soglia del **0.400**, con punteggi tra **0.366** e **0.382**. Questo suggerisce che, nonostante l'uso dell'ensemble retriever, esiste ancora ampio margine di miglioramento per aumentare la fedeltà delle risposte fornite dai modelli rispetto ai dati originari. L'answer relevancy, mostra invece risultati più promettenti. LlamaMed emerge come il modello più performante con un punteggio di **0.838**, seguito da Llama con **0.861**. Gemma2, MistralNemo, Phi3, Qwen2 e QwenMed raggiungono punteggi simili, attestandosi su **0.833**. Questi dati indicano che l'uso del-

l'ensemble retriever migliora significativamente la rilevanza delle risposte, con LlamaMed che dimostra una capacità superiore di fornire risposte ben allineate alle richieste poste. Infine, per quanto riguarda l'answer correctness, LlamaMed ottiene ancora una volta il punteggio più alto (**0.582**), seguito da Llama con **0.580** e Gemma2 con **0.588**. I restanti modelli, inclusi MistralNemo, Phi3, Qwen2 e QwenMed, registrano punteggi leggermente inferiori (tutti attorno a **0.580**), mostrando un livello di accuratezza relativamente omogeneo. Tuttavia, LlamaMed si distingue nuovamente come il modello più accurato, seppur di poco, rispetto agli altri.

### 5.2.2 Modelli base vs modelli specializzati

In questa sezione si confrontano le performance tra modelli base, come Llama e Qwen2, e i loro rispettivi modelli finetuning specializzati nel dominio medico, ovvero LlamaMed e QwenMed. Lo scopo principale di questa comparazione è verificare se l'ottimizzazione di un modello su un determinato dominio possa migliorare significativamente le sue capacità di fornire risposte più pertinenti e accurate. I modelli specializzati sono stati addestrati su un corpus di dati specifici del settore medico, al fine di adattarsi meglio alle esigenze di quel contesto, rispetto ai modelli di base che presentano un'addestramento più generalista. Nell'ambito di queste analisi, si è scelto di focalizzarsi sulle metriche di *answer correctness* e *answer relevancy*, poiché risultano essere strettamente allineate con l'obiettivo dello studio. Questi due parametri, essendo strettamente legati alla qualità delle risposte generate dai modelli, rappresentano una chiave di valutazione fondamentale per determinare se un modello specializzato, come LlamaMed o QwenMed, possa effettivamente superare le prestazioni delle versioni generiche in contesti specifici come quello sanitario.

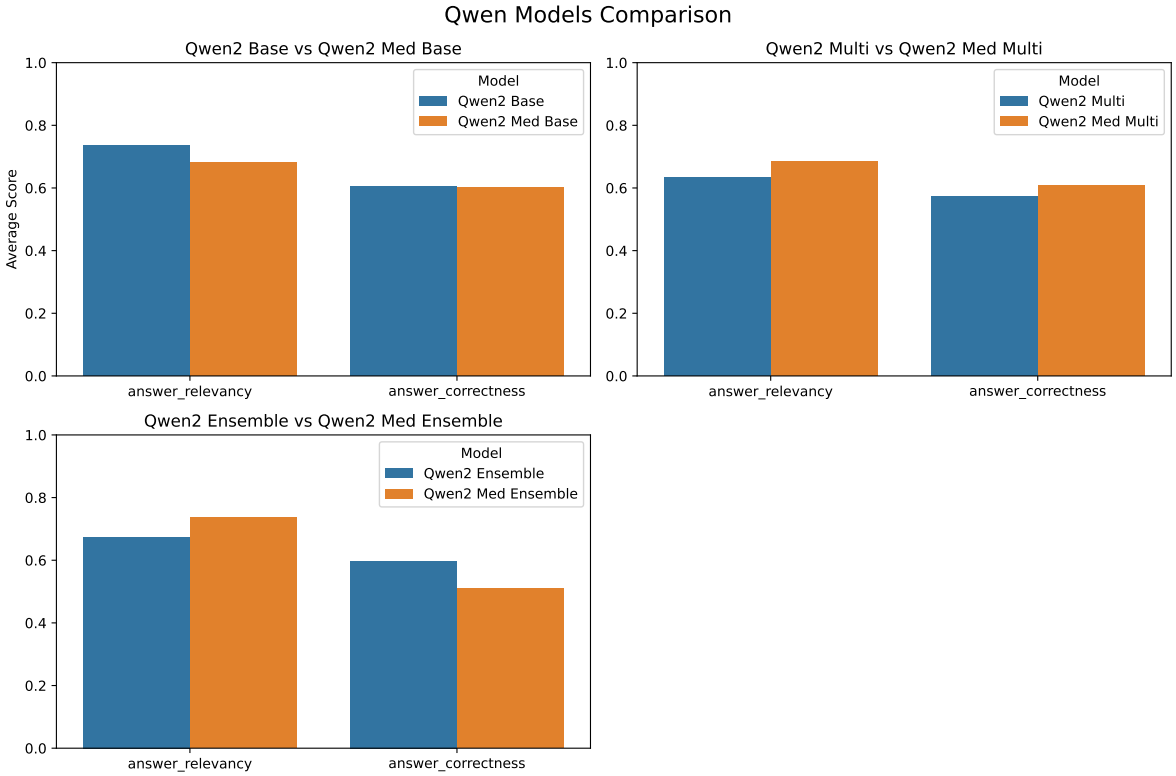


Figura 5.4: Qwen vs Qwen Med

Nella configurazione Base, Qwen2 mostra una leggera superiorità rispetto a Qwen2 Med in entrambe le metriche. Ciò potrebbe indicare che la specializzazione medica, in assenza di tecniche avanzate di retrieval, potrebbe limitare leggermente la versatilità del modello. Con l'implementazione del MultiQuery Retriever, si osserva un'inversione di tendenza: Qwen2 Med supera Qwen2 in entrambe le metriche. Questo suggerisce che la generazione di query da molteplici prospettive si integra particolarmente bene con la conoscenza specializzata medica, probabilmente perché permette di esplorare più efficacemente le sfumature e le connessioni all'interno del dominio medico. Nell'approccio Ensemble, si nota una performance mista: Qwen2 Med eccelle in rilevanza delle risposte, mentre Qwen2 primeggia in correttezza. Questo pattern potrebbe indicare che la combinazione di retriever diversi, incluso BM25, offre vantaggi significativi nel contesto medico per quanto riguarda la pertinenza, ma la maggiore complessità potrebbe introdurre qualche

imprecisione nelle risposte.

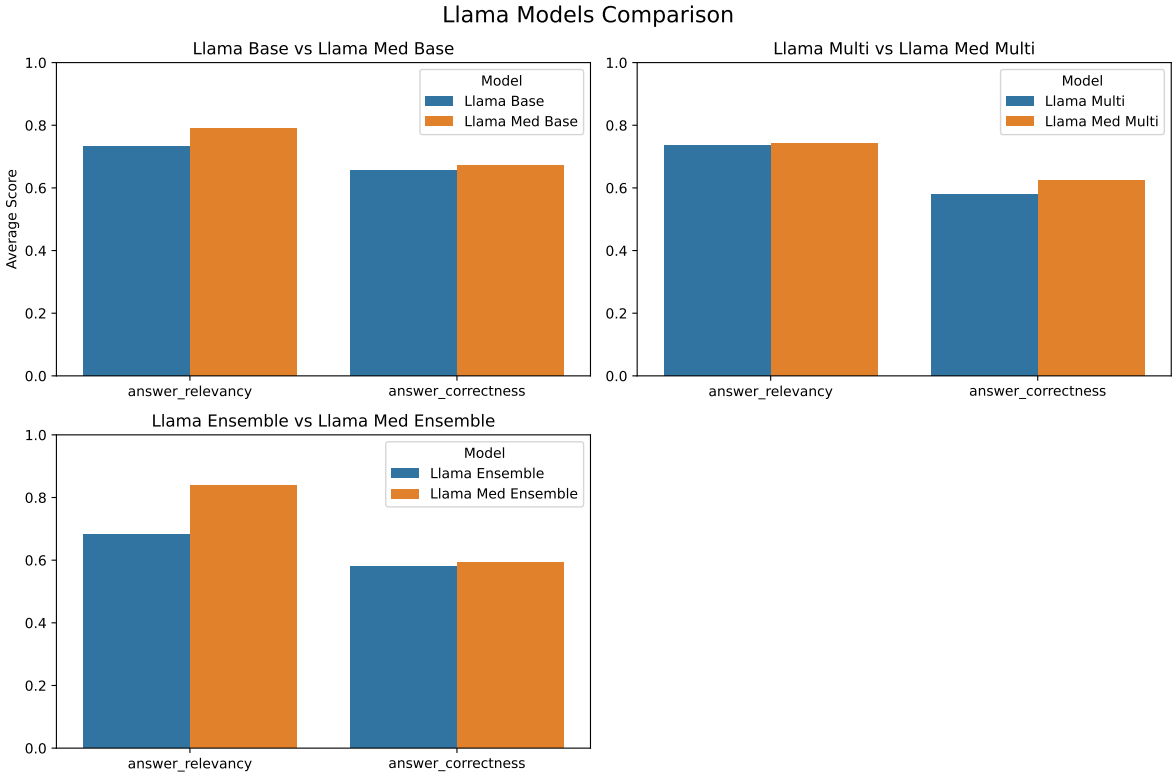


Figura 5.5: Llama vs Llama Med

Nella comparazione Base, Llama Med dimostra una superiorità marginale ma consistente su entrambe le metriche rispetto a Llama, suggerendo che la specializzazione medica offre benefici anche senza tecniche avanzate di retrieval. Con il MultiQuery Retriever, si osserva una sostanziale parità in rilevanza, con un lieve vantaggio di Llama Med in correttezza. Ciò potrebbe indicare che l'approccio multi-query bilancia efficacemente la conoscenza generale e quella specialistica. Con l'approccio Ensemble, si evidenzia il divario più significativo, con Llama Med che supera nettamente Llama nella rilevanza della risposta, mantenendo un vantaggio più contenuto nella correttezza. Questo suggerisce che l'approccio ensemble, combinando diversi retriever, amplifica notevolmente i benefici della specializzazione medica, particolarmente in termini di pertinenza delle risposte.

### 5.2.3 RAG vs NoRAG

Questa sezione si concentra sull'analisi comparativa tra sistemi *RAG* e sistemi *NoRAG*, con l'obiettivo di valutare se l'integrazione di conoscenza aggiuntiva tramite meccanismi di retrieval migliori effettivamente le prestazioni rispetto all'utilizzo esclusivo di un modello di linguaggio pre-addestrato con un prompt specifico. Anche in questo caso, l'analisi si è focalizzata sulle metriche di *answer relevancy* e *answer correctness*, in quanto sono maggiormente rilevanti per lo scopo dello studio. L'ipotesi alla base di questa comparazione è che un sistema RAG possa offrire risposte più pertinenti e accurate, soprattutto in domini complessi o specialistici, grazie all'accesso a informazioni aggiornate o non incluse nel modello di base. Tuttavia, l'analisi si prefigge di esplorare se tali vantaggi siano costanti o se vi siano contesti in cui l'utilizzo di un LLM con un prompt adeguatamente progettato possa risultare comparabile o addirittura superiore a un sistema RAG. Attraverso questo confronto, si intende chiarire se il beneficio apportato da un approccio RAG giustifichi la complessità aggiuntiva, o se, in alcuni casi, la semplicità e l'efficacia di un modello NoRAG siano sufficienti a soddisfare i requisiti di pertinenza e correttezza delle risposte.

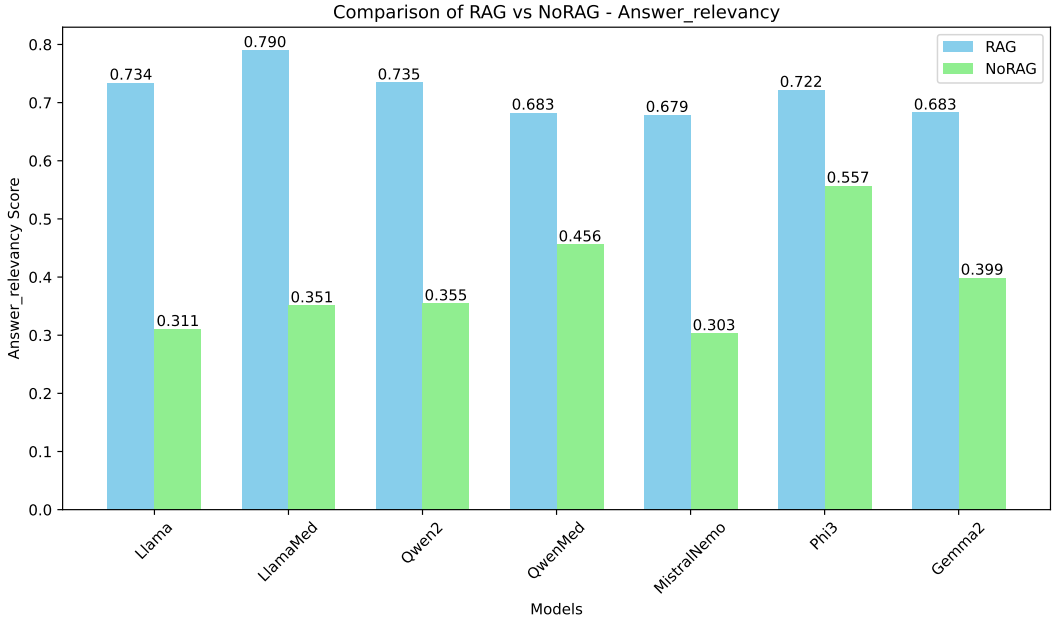


Figura 5.6: RAG vs NoRAG - answer relevancy

Il grafico mostra chiaramente che i sistemi RAG superano significativamente i sistemi NoRAG in termini di *answer relevancy* per vari modelli linguistici, come Llama, LlamaMed, Qwen2, QwenMed, MistralNemo, Phi3 e Gemma2. L'integrazione di retrieval consente ai modelli di accedere a informazioni aggiornate e contestualmente rilevanti, migliorando notevolmente la pertinenza delle risposte. In particolare, LlamaMed e Phi3 mostrano incrementi significativi nella rilevanza delle risposte grazie ai sistemi RAG. Tuttavia, anche modelli come Llama e Gemma2, pur con miglioramenti più contenuti, beneficiano di questo approccio, suggerendo che la rilevanza delle risposte può variare in base all'architettura del modello. Anche se i modelli NoRAG tendono ad avere punteggi più bassi, alcuni, come QwenMed e Phi3, riescono comunque a mantenere una certa pertinenza senza retrieval, dimostrando che i modelli ben addestrati possono produrre risposte rilevanti anche senza un sistema di retrieval esterno. Inoltre, i modelli specializzati, come LlamaMed e QwenMed, ottengono un beneficio particolare dall'approccio RAG, indicando che, nei contesti specialistici come quello medico, l'accesso a informazioni precise è cruciale. In sintesi, l'integrazione di RAG migliora significativamente la rilevanza delle risposte rispetto ai sistemi NoRAG, con variazioni nell'efficacia che dipendono dal modello specifico e dal contesto di applicazione.

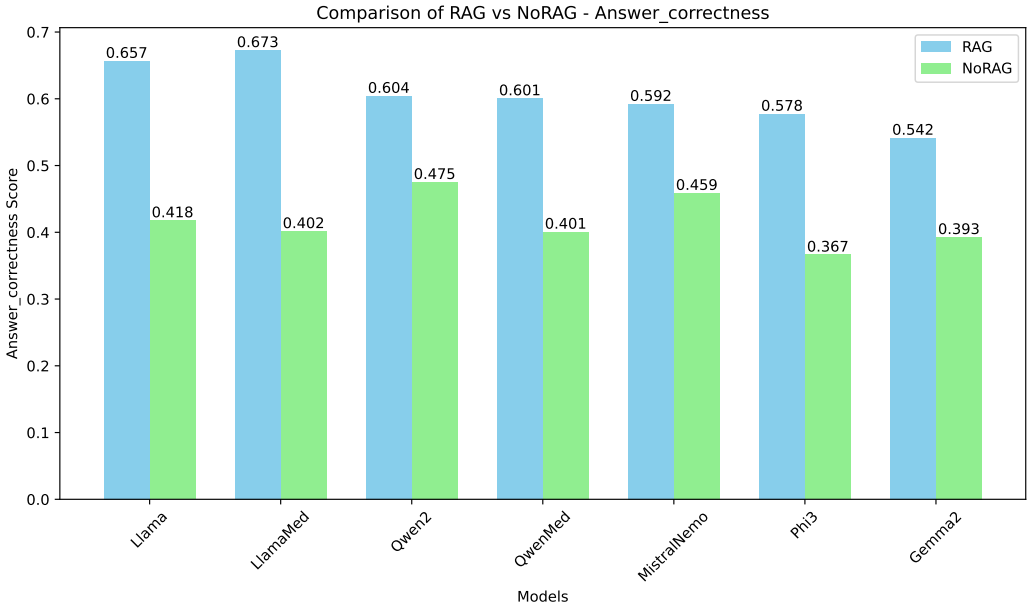


Figura 5.7: RAG vs NoRAG - answer correctness

Anche per quanto riguarda l'analisi dell'*answer correctness*, i dati mostrano che il modello RAG supera costantemente il modello NoRAG per tutti i modelli testati. Nell'esempio di Llama, il modello RAG ottiene un punteggio di **0.657**, significativamente superiore a**0.418** del modello NoRAG. Un trend simile è osservato in LlamaMed, dove il RAG raggiunge un punteggio di **0.673** contro il **0.402** di NoRAG. Questi risultati confermano l'efficacia del modello RAG nell'offrire risposte più accurate rispetto al modello NoRAG su tutti i modelli analizzati. L'analisi sottolinea come l'integrazione di un sistema di retrieval possa migliorare significativamente la qualità delle risposte rispetto a modelli senza tale supporto.

# Conclusioni

Questa ricerca ha esaminato le dinamiche dei modelli linguistici avanzati, con un focus specifico sul loro impiego in ambiti specialistici come quello medico. L'analisi comparativa condotta su diversi modelli e tecniche di retrieval ha fornito preziose informazioni, evidenziando al contempo sfide chiave per il futuro sviluppo di tali sistemi.

Uno degli aspetti più critici emersi è la difficoltà nel migliorare la metrica di *faithfulness*, con punteggi che variano tra 0.355 e 0.443 per tutti i modelli esaminati. Questo indica una problematica nel garantire che le risposte generate riflettano accuratamente le informazioni di input, un elemento essenziale in contesti come quello medico. Nonostante i buoni risultati in termini di *answer relevancy* e *correctness*, con punteggi superiori a 0.73 per modelli come LlamaMed, Llama e Qwen2, la scelta del metodo di retrieval ha mostrato una grande influenza sulla pertinenza delle risposte. Le differenze tra approcci come Base, MultiQuery e Ensemble suggeriscono che la selezione della strategia di recupero delle informazioni è cruciale e deve essere adattata in base al contesto.

Il confronto tra modelli base e specializzati ha evidenziato l'efficacia del *fine-tuning* in contesti specifici, soprattutto quando abbinato a tecniche avanzate di retrieval come il MultiQuery Retriever, che esplora diverse prospettive all'interno di un dominio complesso. Tuttavia, l'approccio Ensemble, sebbene efficace in termini di rilevanza, ha mostrato limiti in termini di correttezza, specialmente con i modelli Qwen, a causa della complessità nell'integrazione di diversi retriever. Questo suggerisce che, pur offrendo vantaggi, la specializzazione richiede un bilanciamento tra conoscenze generali e specifiche, e una calibratura accurata delle tecniche di retrieval in relazione al dominio di conoscenza.

La comparazione tra sistemi RAG e NoRAG ha mostrato che i primi sono superiori sia in termini di *answer relevancy* che di *answer correctness*. Questo evidenzia l'importanza di sistemi in grado di accedere a informazioni aggiornate e rilevanti per migliorare l'accuratezza e la pertinenza delle risposte.

I risultati ottenuti indicano che non esiste un approccio universalmente ottimale, poiché le performance dei modelli variano notevolmente a seconda del metodo di retrieval, del livello di specializzazione e dell'integrazione dei sistemi RAG. Pertanto, le ricerche future dovrebbero concentrarsi su diversi ambiti chiave. In primo luogo, è cruciale sviluppare tecniche avanzate per ottimizzare la faithfulness, ossia l'aderenza delle risposte alle in-

formazioni di input, attraverso l'implementazione di meccanismi di verifica e controllo più sofisticati. Parallelamente, è necessario affinare le tecniche di retrieval, esplorando approcci ibridi o adattivi che selezionino dinamicamente la strategia più adatta in base alla query e al contesto. Un altro aspetto rilevante riguarda la specializzazione contestuale, che richiede metodologie di fine-tuning avanzate in grado di bilanciare in modo efficace la conoscenza generale con quella specifica del dominio. In aggiunta, occorre sviluppare tecniche più evolute per l'integrazione dei sistemi RAG, sfruttando l'utilizzo di fonti informative multiple e adattando dinamicamente le strategie di retrieval. Infine, è fondamentale sviluppare metriche di valutazione multidimensionali in grado di misurare non solo la correttezza e la rilevanza, ma anche la coerenza logica, la profondità delle informazioni e la capacità di sintesi delle risposte prodotte.

Questa ricerca ha evidenziato la necessità di un approccio flessibile e contestualizzato nella progettazione e implementazione dei modelli linguistici. È fondamentale integrare in modo efficace conoscenze generali e specialistiche e adottare strategie di retrieval dinamiche per garantire risposte rilevanti, accurate e fedeli. Le prospettive future promettono sviluppi significativi nella capacità dei sistemi AI di fornire risposte affidabili e pertinenti in una vasta gamma di contesti, aprendo la strada a una nuova generazione di modelli in grado di affrontare con successo la complessità e la variabilità del mondo reale.

# Ringraziamenti

Con la conclusione di questo percorso universitario, giungo alla fine di una fase importante della mia vita, che ha segnato la mia crescita sia personale che professionale. Questo traguardo non sarebbe stato possibile senza il sostegno e la guida di persone che mi hanno accompagnato lungo il cammino.
All'inizio di questo studio, non immaginavo minimamente quanto sarebbe stato rilevante il raggiungimento di questo obiettivo, fino a culminare nella pubblicazione di questo lavoro nell'ambito della ricerca. Questo risultato mi riempie di orgoglio e soddisfazione, sapendo di aver contribuito allo sviluppo della conoscenza nel campo dell'intelligenza artificiale
Desidero quindi ringraziare il Prof. Stefano Ferretti, il mio relatore, per avermi permesso di scegliere e intraprendere il percorso di sviluppo di questa tesi. La sua costante disponibilità e presenza nei momenti di difficoltà sono state per me un punto di riferimento essenziale. Un sincero ringraziamento va anche alla Dott.ssa Sara Montagna e al Dott. Ing. Gianluca Aguzzi, per avermi seguito con attenzione durante tutto il lavoro. I loro preziosi consigli, offerti sempre al momento giusto, sono stati fondamentali per il raggiungimento del risultato finale.
Questo percorso accademico, che mi ha condotto fino a qui, non sarebbe stato possibile senza il supporto costante della mia famiglia. Mia madre, mio padre e mia sorella sono stati il pilastro su cui ho potuto contare in ogni momento. Loro hanno sempre creduto in me, spesso più di quanto non abbia fatto io stesso, e ogni traguardo che ho raggiunto, incluso questo, è anche il loro. Mamma e Papà, l'ultimo periodo, come ben sapete, non è stato affatto semplice, ma voi siete stati esemplari, affrontando ogni difficoltà nel migliore dei modi senza mai chiedere nulla a nessuno. Per questo sappiate che vi ammiro profondamente. La vostra presenza costante e il vostro supporto incondizionato sono stati fondamentali per il mio percorso e mi hanno insegnato il valore del sacrificio e della dedizione. Ogni passo che ho fatto è stato possibile grazie a voi. Non vedo l'ora di restituirvi un minimo di ciò che mi avete dato in tutti questi anni, perché siete la mia forza e la mia ispirazione. Grazie di cuore per tutto ciò che fate e per il vostro amore che mi accompagna ogni giorno.
Un ringraziamento speciale va alla mia "piccola ma grande" sorella Conny. Nonostante la

sua statura non sia delle più imponenti, la sua presenza al mio fianco è stata gigantesca. Mi ha sempre fatto sentire a casa ogni volta che andavo a Parma. Grazie, per avermi sempre supportato e, soprattutto, "sopportato" durante tutto questo viaggio accademico, trovando sempre le parole giuste per aiutarmi a superare i momenti di sconforto e riponendo sempre tanta fiducia in me... anche quando ti facevo impazzire! Conny, sappi che così come tu ci sei sempre stata per me, io ti resterò sempre accanto. Un fratello è per sempre, e qualunque strada tu decida di percorrere, non sarai mai sola. Sarò sempre lì a sostenerti, a condividere i tuoi successi e a darti forza nei momenti difficili come tu hai sempre fatto con me. Un ringraziamento va anche a Daniele, per la sua generosità e per tutte le volte in cui è stato ingiustamente mandato a prendermi in stazione, con la pazienza e la disponibilità di un santo.

Accanto alla mia famiglia, non posso dimenticare di ringraziare coloro che hanno condiviso con me le sfide, le fatiche e le soddisfazioni di questo percorso accademico: i miei colleghi e amici. Ognuno di loro ha contribuito, in modi diversi, a rendere questo viaggio unico e arricchente. Inizio col ringraziare Acco, con cui ho condiviso cinque anni di studio. Grazie a lui ho imparato leggermente l'arte della tranquillità nell'affrontare i problemi, il tipo di approccio che, se il mondo crolla, lui semplicemente si sposta più in là, mentre io tendo a crearne di nuovi. Non è da tutti, e gliene sono davvero grato.

Un ringraziamento va anche a Giulio, con cui ho percorso insieme la strada fin dalla triennale. Anche se non ci siamo visti spesso, abbiamo sempre mantenuto un legame forte e autentico, che ha reso questo viaggio ancora più significativo.

Un altro ringraziamento va a Beni, collega durante la triennale e oggi amico. Tutto è iniziato dal nostro primo anno in studentato, dove ci siamo conosciuti e scoprendo successivamente di essere anche colleghi di studio. Da allora il suo supporto è stato fondamentale. Tra i banchi delle aule studio e tra uno spritz e l'altro, mi ha sempre rasserenato e spronato a dare il meglio. Mi ha insegnato ad avere più fiducia in me stesso e per questo gli sarò sempre grato.

Ho deciso di ricominciare questo percorso non solo per completare i miei studi, ma anche per incontrare nuove persone, amici che potessero contribuire a migliorarmi come persona. In questo viaggio, un grazie è doveroso a chi ha condiviso con me, giorno dopo giorno, ogni passo del cammino. Per questo, desidero ringraziare profondamente Manuel

e Alessandro. Innanzitutto, grazie per i progetti con i voti più alti... ovvero quelli che non ho fatto io! Scherzi a parte, la vostra presenza mi ha permesso di rivivere questi anni di spensieratezza nel migliore dei modi. Mi avete fatto sentire apprezzato sin da subito, creando un feeling speciale che ha reso ogni giorno unico. Peccato solo che Manuel sia interista, ma nessuno è perfetto! Abbiamo condiviso insieme non solo le gioie, ma anche i momenti più difficili, come i temuti esami e le sfide universitarie. Ogni volta, li abbiamo affrontati al meglio, dimostrando le nostre capacità e imparando continuamente l'uno dall'altro, migliorandoci a vicenda. Ma soprattutto, vi devo un ringraziamento speciale per avermi convinto a comprare la PS5. Senza di voi, non avrei mai potuto godermi Spiderman e Dragon Ball Sparking Zero, e, diciamolo, la mia vita non sarebbe stata la stessa! Scherzi a parte, grazie davvero, per tutti i bei momenti passati insieme!

Adesso è giunto il momento di ringraziare "il Virgilio che mi ha condotto fino al paradiso", Prof. Isabella, mia collega e amica dall'infanzia, nonché relatrice non ufficiale di questa tesi, meriti un ringraziamento speciale. La tua amicizia sincera e autentica si è rafforzata ancor di più in questo percorso, ed è stato un sostegno prezioso. Grazie per tutte le correzioni, gli appunti, i consigli... e potrei continuare, ma la lista è troppo lunga! Ti sarò debitore per tutte le volte che mi hai aiutato, anche se, diciamolo, dovresti ringraziare anche me per aver reso le tue giornate universitarie un po' più impegnative!

Ora è il momento di rivolgere un pensiero speciale anche agli amici di sempre, che, sebbene non abbiano vissuto ogni istante di questo percorso accademico, sono stati comunque una parte fondamentale del mio viaggio. Ognuno di loro, in modo diverso, ha contribuito a rendere questo cammino più leggero, arricchendolo con momenti indimenticabili e offrendo il loro sostegno e la loro amicizia, che sono stati per me un punto di riferimento prezioso. Innanzitutto grazie a tutti coloro che hanno percorso tanti chilometri per essere qui a condividere questo mio successo. La vostra presenza non era per niente scontata e mi ha fatto capire quanto ci teniate davvero a me.

Un ringraziamento speciale va al buon vecchio Pizzi. Anche tu, amico di mille avventure e di infinite chiacchierate filosofiche sul porto turistico, dove ci siamo spesso persi a fantasticare e a immaginare un futuro migliore per entrambi. Grazie per riuscire sempre a strapparmi un sorriso con i tuoi reel, ma soprattutto per esserci quando le cose non vanno come dovrebbero. È giusto che, a modo nostro, ci scambiamo le nostre paranoie,

alimentandole a vicenda, ma trovando sempre un modo per andare avanti insieme.
Un ringraziamento va anche al mio amico di vecchia data, Michele, che è riuscito a convincermi a venire a Bologna, nonostante io fossi tutt'altro che sicuro. Con il tuo modo di parlare, sembrava quasi che stessi firmando un contratto in tribunale! In parte grazie a te ho potuto vivere questa incredibile esperienza e sei stato davvero un ottimo persuasore.
Un ringraziamento speciale va anche alle mitiche "*Rominas*" ovvero Checca, Sara e Romina. Vi conosco da ormai tanto tempo, e nel corso degli anni avete dimostrato di essere davvero di una purezza unica. La vostra amicizia, oggi più che mai, è una rarità preziosa. Nel bene e nel male, siete sempre state presenti per me, e vi prometto che, allo stesso modo, io ci sarò sempre per voi.
Dai miei amici, non posso dimenticare di ringraziare Antonio, il mio amico ed ex coinquilino dello studentato. La nostra convivenza è stata un'avventura a sé, piena di risate, momenti indimenticabili e qualche piccola disavventura. Grazie per aver reso quel periodo così speciale e per il supporto che mi hai dato in ogni situazione. E, a proposito, spero che tu possa ridurre un po' le tue ore di sonno per uscire più spesso!
Ultima ma non meno importante, desidero esprimere la mia gratitudine a Michela, la mia fidanzata. La tua presenza è stata un sostegno silenzioso che ha reso ogni sfida più affrontabile e ogni gioia ancora più grande. Grazie per la tua pazienza, anche se devo ammettere che hai ancora un po' di strada da fare! Ogni passo che ho fatto in questo percorso è stato più dolce perché tu eri al mio fianco, pronta a sostenermi a modo tuo nei momenti di difficoltà.
E infine, desidero ringraziare me stesso. Non è un gesto di presunzione, ma un riconoscimento del percorso che ho intrapreso e delle sfide che ho affrontato. Ho imparato a non mollare mai, a perseguire con determinazione tutti i miei obiettivi e a trarre insegnamenti preziosi da ogni esperienza. Questo viaggio mi ha fortificato e mi ha permesso di maturare, rendendomi pronto ad affrontare le sfide che la vita continuerà a propormi. Ho compreso l'importanza di affrontare ogni situazione con umiltà e con la costante voglia di migliorare, e questo è merito di tutti coloro che hanno realmente creduto in me. E ora, con il cuore pieno di gratitudine, guardo al futuro con entusiasmo, pronto ad affrontare nuove sfide e a vivere nuovi traguardi!

# Bibliografia


[1] Punyakeerthi BL. *Tokenizer in LLM*. [Online; accessed 2-August-2024]. 2024. URL: https://medium.com/@punya8147_26846/tokenizer-in-llm-060b1a35694b.

[2] L.F. Bouchard, L. Peters e Towards AI. *Building LLMs for Production: Enhancing LLM Abilities and Reliability with Prompting, Fine-tuning, and RAG*. Towards AI, 2024. ISBN: 979-8-3247-3147-2. URL: https://books.google.it/books?id=siLP0AEACAAJ.

[3] Harrison Chase. *LangChain framework documentation*. [Online; accessed 02-September-2024]. 2024. URL: https://python.langchain.com/v0.2/docs/introduction/.

[4] Michael Chiang e Jeffrey Morgan. *Ollama open-source tool*. [Online; accessed 09-September-2024]. 2024. URL: https://ollama.com/.

[5] Jung Sun Cho e Jae-Hyeong Park. «Application of artificial intelligence in hypertension». In: *Clinical Hypertension* 30.1 (mag. 2024), p. 11. ISSN: 2056-5909. DOI: 10.1186/s40885-024-00266-9. URL: https://doi.org/10.1186/s40885-024-00266-9.

[6] Jan Clusmann et al. «The future landscape of large language models in medicine». In: *Communications Medicine* 3.1 (ott. 2023), p. 141. ISSN: 2730-664X. DOI: 10.1038/s43856-023-00370-1. URL: https://doi.org/10.1038/s43856-023-00370-1.

[7] Suman Das. *Fine Tune Large Language Model (LLM) on a Custom Dataset with QLoRA*. [Online; accessed 2-August-2024]. 2024. URL: https://dassum.medium.com/fine-tune-large-language-model-llm-on-a-custom-dataset-with-qlora-fb60abdeba07.

[8] Shahul Es et al. *RAGAS: Automated Evaluation of Retrieval Augmented Generation*. 2023. arXiv: 2309.15217 [cs.CL]. URL: https://arxiv.org/abs/2309.15217.

[9] ExplodingGradients. *Ragas framework documentation*. [Online; accessed 28-August-2024]. 2023. URL: https://docs.ragas.io/en/stable/.

[10] Fabian. *Figuring Out the Ideal Chunk Size*. [Online; accessed 28-August-2024]. 2024. URL: https://medium.com/@farenas1/fabian-7d1f90ac4cb4.

[11] Peter Foy. *Understanding Transformers & the Architecture of LLMs*. [Online; accessed 1-August-2024]. 2024. URL: https://blog.mlq.ai/llm-transformer-architecture.

[12] Cobus Greyling. *RAG Evaluation*. [Online; accessed 28-August-2024]. 2023. URL: https://www.humanfirst.ai/blog/rag-evaluation.

[13] Hostcomm. *Open Source vs. Proprietary LLMs: A Comprehensive Comparison*. [Online; accessed 26-August-2024]. 2024. URL: https://www.hostcomm.co.uk/blog/2024/open-source-vs-proprietary-llms-a-comprehensive-comparison.

[14] Jeff Huber e Anton Troynikov. *Chroma vectorstore*. [Online; accessed 09-September-2024]. 2024. URL: https://www.trychroma.com/.

[15] Jerry Liu. *LlamaIndex framework documentation*. [Online; accessed 03-September-2024]. 2024. URL: https://docs.llamaindex.ai/.

[16] «Machine Learning in Hypertension Detection: A Study on World Hypertension Day Data». In: *Journal of Medical Systems* 47.1 (2022), p. 10. DOI: 10.1007/s10916-022-01900-5. URL: https://doi.org/10.1007/s10916-022-01900-5.

[17] Bertalan Meskó e Eric J. Topol. «The imperative for regulatory oversight of large language models (or generative AI) in healthcare». In: *npj Digital Medicine* 6.1 (lug. 2023), p. 120. ISSN: 2398-6352. DOI: 10.1038/s41746-023-00873-0. URL: https://doi.org/10.1038/s41746-023-00873-0.

[18] Zabir Al Nazi e Wei Peng. *Large language models in healthcare and medical domain: A review*. 2024. arXiv: 2401.06775 [cs.CL]. URL: https://arxiv.org/abs/2401.06775.

[19] Tavva Prudhvith. *BERT vs GPT: A Tale of Two Transformers That Revolutionized NLP*. [Online; accessed 4-August-2024]. 2024. URL: https://medium.com/@prudhvithtavva/bert-vs-gpt-a-tale-of-two-transformers-that-revolutionized-nlp-11fff8e61984.

[20] Joaquin Fernandez Sande. *Retrieval Augmented Generation & LLMs in a Healthcare Setting*. [Online; accessed 18-September-2024]. 2024. URL: https://medium.com/@joacofernandezsande/retrieval-augmented-generation-and-healthcare-9f67515ce0e9.

[21] Kartik Talamadupula. *Guide to Context in LLMs*. [Online; accessed 4-August-2024]. 2024. URL: https://symbl.ai/developers/blog/guide-to-context-in-llms.

[22] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.