**~~Sviluppo di Metriche Standardizzate per la Misurazione del Bias Geo-Culturale nei Modelli di Generazione di Immagini da Testo~~**

#### **~~Obiettivo:~~**

* ~~Questo progetto mira a sviluppare nuove metriche per identificare e misurare il bias geo-culturale nei modelli di generazione di immagini da testo, come DALL-E e MidJourney. L’obiettivo è creare un framework unificato che permetta una valutazione quantitativa e qualitativa del bias culturale, includendo aspetti come l’identità di genere non-binaria e le rappresentazioni culturali minoritarie.~~

#### **~~Domande di Ricerca:~~**

* ~~Come possono essere identificate e quantificate le rappresentazioni culturali distorte nei modelli generativi?~~
* ~~Quali sono le attuali lacune nelle metriche esistenti per la rilevazione del bias culturale nei modelli multimodali?~~
* ~~In che modo le nuove metriche proposte possono essere integrate nei processi di sviluppo dei modelli per migliorare la diversità e l’inclusività?~~

#### **~~Metodologia:~~**

* **~~Analisi Quantitativa:~~** ~~Revisione e confronto delle metriche esistenti per la rilevazione del bias, con un focus sulla quantificazione delle rappresentazioni distorte attraverso analisi statistica su un dataset ampio e diversificato.~~
* **~~Analisi Qualitativa:~~** ~~Studio di casi per esplorare come le immagini generate riflettano o distorcano aspetti culturali, utilizzando l’analisi critica dei contenuti visivi.~~
* **~~Sviluppo di nuove metriche:~~** ~~Progettazione di metriche specifiche per quantificare il bias geo-culturale, integrate con un'analisi qualitativa delle implicazioni culturali e sociali.~~
* **~~Test delle metriche:~~** ~~Applicazione delle metriche su dataset diversificati e analisi qualitativa delle discrepanze emerse per una valutazione comprensiva.~~

#### **~~Esempio di Applicazione:~~**

* ~~Generare immagini per la descrizione "una famiglia a cena" e valutare quantitativamente la diversità delle culture rappresentate rispetto alla distribuzione demografica globale. Parallelamente, condurre un'analisi qualitativa sull’aderenza o deviazione dalle aspettative culturali.~~

#### **~~Risultati Attesi:~~**

* ~~Creazione di un insieme di metriche standardizzate che consentano una valutazione quantitativa e qualitativa più equa e accurata del bias culturale nei modelli di generazione di immagini, migliorando la rappresentatività globale.~~

**~~Sviluppo di Metodi Avanzati di Post-Processing per la Mitigazione del Bias nei Modelli di Generazione di Immagini da Testo~~**

#### **~~Obiettivo:~~**

* ~~L'obiettivo di questa tesi è progettare e implementare tecniche di post-processing innovative che riducano il bias nelle immagini generate da modelli di AI, combinando un'analisi quantitativa della riduzione del bias con una valutazione qualitativa dell'impatto visivo.~~

#### **~~Domande di Ricerca:~~**

* ~~Quali tecniche di post-processing sono più efficaci per rimuovere bias demografici nelle immagini generate?~~
* ~~Come possono essere applicate queste tecniche senza compromettere la qualità visiva delle immagini?~~
* ~~Quali sono i limiti attuali delle tecniche di debiasing pre- e post-elaborazione, e come possono essere superati con nuovi metodi?~~

#### **~~Metodologia:~~**

* **~~Analisi Quantitativa:~~** ~~Sviluppo di algoritmi che identificano e quantificano la presenza di bias nelle immagini, seguito da una valutazione quantitativa dell'efficacia delle tecniche di post-processing.~~
* **~~Analisi Qualitativa:~~** ~~Valutazione dell'impatto visivo e percezione delle immagini post-processate attraverso focus group e analisi del contenuto visivo per garantire che le correzioni non compromettano l'integrità estetica.~~
* **~~Comparazione dei risultati:~~** ~~Confronto quantitativo tra immagini pre- e post-elaborate, integrato da un'analisi qualitativa delle immagini per determinare l'aderenza alle aspettative culturali e sociali.~~

#### **~~Esempio di Applicazione:~~**

* ~~Correggere immagini che rappresentano "un leader politico" e che tendono a raffigurare sempre un uomo bianco. Valutare quantitativamente la diversità introdotta e qualitativamente l’accettabilità delle immagini corrette da un punto di vista sociale e culturale.~~

#### **~~Risultati Attesi:~~**

* ~~Un set di tecniche di post-processing che migliorino l’equità e la rappresentatività delle immagini generate, mantenendo la qualità visiva e riducendo i bias, confermato da analisi quantitative e qualitative.~~

**Indagine sulla Sycophancy Visiva nei Modelli Multimodali e Sviluppo di Tecniche per la sua Mitigazione**

#### **Obiettivo:**

* Questa ricerca si propone di studiare il fenomeno della sycophancy visiva nei modelli multimodali e sviluppare tecniche che riducano la prevalenza di questo comportamento, combinando un'analisi quantitativa della coerenza tra input testuali e output visivi con un'analisi qualitativa dell'impatto sugli utenti.

#### **Domande di Ricerca:**

* Quali sono le cause principali della sycophancy visiva nei modelli multimodali?
* Come influisce la sycophancy visiva sull'affidabilità dei modelli in applicazioni reali?
* Quali strategie possono essere implementate per ridurre la dipendenza eccessiva dalle informazioni visive in modelli multimodali?

#### **Metodologia:**

* **Analisi Quantitativa:** Creazione e analisi di dataset progettati per misurare la coerenza tra testo e immagine, valutando la frequenza e la gravità della sycophancy visiva.
* **Analisi Qualitativa:** Studio delle percezioni degli utenti e dell’impatto della sycophancy visiva attraverso interviste e focus group, per comprendere le implicazioni pratiche e sociali del fenomeno.
* **Sviluppo di tecniche di mitigazione:** Progettazione di algoritmi che introducono pesi o regole decisionali, seguiti da un'analisi qualitativa dei cambiamenti apportati rispetto alle aspettative degli utenti.

#### **Esempio di Applicazione:**

* Verificare se un modello genera una "mela verde" o se tende a riprodurre una mela rossa presente in un'immagine di input anche quando il testo richiede una mela verde. Valutare quantitativamente la coerenza dei risultati e qualitativamente l’accettabilità degli output corretti dagli utenti.

#### **Risultati Attesi:**

* Un modello migliorato che mantenga la coerenza tra le descrizioni testuali e le immagini generate, riducendo la prevalenza della sycophancy visiva e aumentando l’affidabilità nei contesti decisionali, come confermato da analisi quantitative e qualitative.

### ****Trasparenza, Spiegabilità e Gestione della Sycophancy Visiva nei Modelli Multimodali: Un Approccio Integrato****

#### **Obiettivo:**

Esaminare come la sycophancy visiva influisce sulle prestazioni e sulle decisioni dei modelli multimodali e sviluppare tecniche per migliorare la trasparenza e la spiegabilità di tali modelli, garantendo al contempo una riduzione del bias.

#### **Domande di Ricerca:**

1. **In che modo la sycophancy visiva influisce sulle prestazioni e sulle decisioni dei modelli multimodali in contesti visivi e testuali contrastanti?**
2. **Quali tecniche di debiasing sono più efficaci nel ridurre la sycophancy visiva senza compromettere la qualità e la diversità delle risposte generate?**
3. **Come percepiscono gli utenti finali le decisioni dei modelli multimodali in presenza di sycophancy visiva, e come questa influenza la loro fiducia nel sistema?**
4. **Quali sono le implicazioni etiche dell'uso di modelli multimodali affetti da sycophancy visiva, e come possono essere sviluppate linee guida per mitigarne gli effetti negativi?**

#### **Metodologia Quantitativa:**

* **Sycophancy Detection:** Creazione e applicazione di un sistema automatizzato per rilevare la sycophancy visiva nei modelli multimodali, utilizzando un dataset eterogeneo che combina immagini e contesti testuali con opinioni contrastanti. La sycophancy sarà misurata valutando la conformità delle risposte del modello alle immagini dominanti rispetto ai contesti testuali forniti.
* **Intervento di Debiasing:** Implementazione di tecniche di debiasing, come la riformulazione dei dati d'addestramento e la regolazione dei parametri del modello, per ridurre la sycophancy visiva. L'impatto di queste tecniche sarà valutato attraverso esperimenti quantitativi, utilizzando metriche come precisione, diversità e accuratezza delle risposte.

#### **Metodologia Qualitativa:**

* **Interviste con Utenti Finali:** Conduzione di interviste approfondite con utenti finali per esplorare come percepiscono e comprendono le decisioni prese dai modelli multimodali, specialmente in situazioni in cui è presente sycophancy visiva. Le interviste mireranno a raccogliere feedback sulla fiducia nel sistema e sulle percezioni delle risposte fornite in contesti con immagini dominanti.
* **Analisi Etica:** Esplorazione delle implicazioni etiche della sycophancy visiva nei modelli multimodali, attraverso discussioni con specialisti in etica e AI. L'analisi si concentrerà sull'adeguatezza delle soluzioni attuali e sulla proposta di linee guida etiche per l'implementazione di modelli multimodali che minimizzino gli effetti negativi della sycophancy visiva.

#### **Esempi di Applicazione:**

* **Generazione di Contenuti Educativi:** Applicazione delle tecniche di debiasing per ridurre la sycophancy visiva in sistemi che generano contenuti educativi, assicurando che le risposte non siano eccessivamente influenzate da immagini dominanti e che riflettano accuratamente il contesto testuale e le diversità culturali.
* **Sistemi di Raccomandazione Multimodali:** Utilizzo delle tecniche sviluppate per migliorare la qualità delle raccomandazioni in sistemi che combinano immagini e testo, come nelle piattaforme di e-commerce o streaming, riducendo la tendenza del sistema a favorire immagini dominanti a discapito delle preferenze testuali dell'utente.
* **Assistenti Virtuali Multimodali:** Implementazione delle soluzioni per assistenti virtuali che utilizzano input visivi e testuali, garantendo che le risposte non siano eccessivamente conformiste rispetto alle immagini, ma riflettano una comprensione equilibrata del contesto fornito dall'utente.

#### **Risultato Atteso:**

* **Quantitativo:** Sviluppo e validazione di tecniche che riducono efficacemente la sycophancy visiva nei modelli multimodali, senza sacrificare la performance globale del modello, con miglioramenti nelle metriche di precisione e diversità delle risposte.
* **Qualitativo:** Fornire raccomandazioni pratiche su come migliorare la trasparenza e la spiegabilità nei modelli multimodali, affrontando le problematiche etiche associate alla sycophancy visiva. Le raccomandazioni saranno basate su feedback degli utenti finali e discussioni etiche, contribuendo allo sviluppo di linee guida per un'implementazione responsabile.

**Implementazione di Tecniche di Metamorphic Testing per la Rilevazione Automatica del Bias nei Modelli di Generazione di Immagini da Testo**

#### Obiettivo:

* Questa tesi si propone di sviluppare strumenti automatizzati basati su metamorphic testing, capaci di rilevare automaticamente bias nei modelli di generazione di immagini da testo, combinando un'analisi quantitativa dell'efficacia del testing con una valutazione qualitativa dell'impatto dei bias rilevati.

#### Domande di Ricerca:

* Come può il metamorphic testing essere applicato per rilevare automaticamente bias nei modelli generativi?
* Quali bias sono più difficili da rilevare con i metodi tradizionali e come possono essere identificati automaticamente con metamorphic testing?
* Quali sono i vantaggi e le limitazioni dell’uso del metamorphic testing rispetto ai metodi manuali?

#### Metodologia:

* **Progettazione di scenari di metamorphic testing:** Creare scenari che alterano leggermente le descrizioni testuali per testare quantitativamente la coerenza e l'equità delle immagini generate dai modelli, affiancando un'analisi qualitativa dei casi di bias rilevati.
* **Sviluppo di un sistema automatizzato:** Implementare uno strumento che esegua questi test su larga scala, integrando un'analisi quantitativa dei risultati con una valutazione qualitativa delle implicazioni sociali e culturali dei bias identificati.
* **Validazione dei risultati:** Confrontare i risultati ottenuti con metodi tradizionali di testing e analizzare qualitativamente l’efficacia del metamorphic testing nella rilevazione di bias complessi e nascosti.

#### Esempio di Applicazione:

* Creare descrizioni come "un gruppo di amici" e alterare il contesto geografico o culturale per vedere se il modello genera immagini che rappresentano solo una specifica etnia o cultura, ignorando la diversità. Analizzare quantitativamente la frequenza di bias rilevati e qualitativamente l'impatto sociale delle rappresentazioni stereotipate.

#### Risultati Attesi:

* Sviluppo di strumenti di testing automatico che migliorino la robustezza e l’equità delle immagini generate, supportati da un'analisi quantitativa della riduzione dei bias e una valutazione qualitativa delle implicazioni etiche e sociali.

Il **Metamorphic Testing (MT)** è una tecnica di testing del software utilizzata per verificare la correttezza di un programma in assenza di un "oracle", ossia un meccanismo che consente di sapere con certezza quale dovrebbe essere l'output corretto per un determinato input. Questa situazione si presenta spesso in applicazioni complesse come l'apprendimento automatico, la simulazione scientifica e altre aree dove è difficile o impraticabile determinare l'output corretto a priori.

### ****Come Funziona il Metamorphic Testing?****

Il Metamorphic Testing si basa sul concetto di **proprietà metamorfiche**. Una proprietà metamorfiche è una relazione prevedibile tra input e output di un programma, che deve essere mantenuta anche quando l'input viene modificato in modi specifici.

#### **Passaggi Chiave del Metamorphic Testing:**

1. **Identificazione delle Proprietà Metamorfiche**:
   * Gli sviluppatori identificano proprietà del programma che devono mantenersi vere anche quando gli input vengono modificati. Queste proprietà descrivono come l'output dovrebbe cambiare in risposta a modifiche specifiche degli input.
2. **Selezione degli Input di Base**:
   * Si scelgono uno o più input di base per il test iniziale del programma.
3. **Generazione di Input Derivati**:
   * A partire dagli input di base, si generano nuovi input (chiamati input derivati) modificando gli input originali secondo le proprietà metamorfiche identificate. Ad esempio, se l'input originale è un numero, si potrebbe generare un input derivato prendendo il numero negativo.
4. **Esecuzione del Programma**:
   * Si esegue il programma sia con gli input di base che con quelli derivati e si raccolgono i rispettivi output.
5. **Verifica delle Proprietà Metamorfiche**:
   * Si confrontano gli output ottenuti per verificare se rispettano le proprietà metamorfiche. Se una proprietà non viene rispettata, ciò indica la presenza di un bug o di un problema nel programma.

### ****Esempi di Proprietà Metamorfiche:****

* **Funzione Matematica**: Per una funzione f(x) = x^2, una proprietà metamorfiche potrebbe essere che f(x) = f(-x). Se x = 2, l'output dovrebbe essere 4, e per x = -2, l'output dovrebbe essere ancora 4.
* **Algoritmo di Ordinamento**: Un algoritmo di ordinamento dovrebbe produrre lo stesso risultato ordinato indipendentemente dalla permutazione degli elementi input che mantengono gli stessi valori. Ad esempio, se si ordina [3, 1, 2] e [1, 3, 2], entrambi dovrebbero restituire [1, 2, 3].
* **Ricerca del Percorso Minimo**: In un grafo, se si aggiunge un peso costante a tutti gli archi, il percorso minimo tra due nodi non dovrebbe cambiare, anche se il costo totale cambia.

### ****Perché Usare il Metamorphic Testing?****

Il Metamorphic Testing è particolarmente utile in situazioni in cui:

* **Non c'è un Oracle Affidabile**: Non è possibile determinare a priori l'output corretto del programma per un dato input, come in molti algoritmi di apprendimento automatico.
* **Applicazioni Complesse**: Il software scientifico, come i simulatori fisici, spesso ha risultati complessi e non prevedibili con precisione.
* **Test di Regresso**: Il Metamorphic Testing può essere utilizzato per garantire che i cambiamenti nel codice non introducano nuovi bug o alterino le proprietà fondamentali del software.

### ****Vantaggi del Metamorphic Testing****

1. **Efficienza nel Rilevare Bug**: Permette di rilevare errori anche in assenza di un oracle completo, aumentando la copertura del testing.
2. **Facilmente Automatizzabile**: Può essere implementato in strumenti automatizzati che generano input derivati e verificano le proprietà automaticamente.
3. **Utile per Sistemi di Apprendimento Automatico**: È ideale per testare modelli di apprendimento automatico dove gli output corretti sono spesso sconosciuti o difficili da definire.

### ****Limitazioni del Metamorphic Testing****

* **Identificazione delle Proprietà Metamorfiche**: Trovare proprietà metamorfiche appropriate può essere complesso e richiede una buona comprensione del dominio del problema.
* **Non Individua Tutti i Bug**: Il Metamorphic Testing può identificare solo bug che violano le proprietà metamorfiche definite; altri tipi di bug potrebbero non essere rilevati.

In sintesi, il **Metamorphic Testing** è una tecnica potente e flessibile per verificare la correttezza del software in situazioni in cui non è pratico o possibile conoscere gli output corretti, migliorando così la robustezza e l'affidabilità dei sistemi software complessi.

**Editing Granulare della Conoscenza nei Modelli di Linguaggio per la Mitigazione del Bias**

#### Obiettivo:

* L'obiettivo di questa tesi è esplorare e sviluppare tecniche di "knowledge editing" che consentano di modificare conoscenze specifiche all'interno dei LLM per ridurre il bias, integrando un'analisi quantitativa dell'impatto delle modifiche sulla performance e una valutazione qualitativa dell'efficacia etica e sociale.

#### Domande di Ricerca:

* In che modo l'editing granulare della conoscenza può essere utilizzato per correggere i bias specifici all'interno dei LLM?
* Quali sono le sfide tecniche nell'implementazione di tali modifiche senza compromettere la performance del modello?
* Quali tipi di bias (ad esempio, genere, razza, contesto culturale) possono essere efficacemente mitigati tramite questo approccio?

#### Metodologia:

* **Analisi e Revisione:** Studiare quantitativamente gli approcci esistenti di knowledge editing e il metodo FAST, integrando un'analisi qualitativa per esplorare le implicazioni etiche e sociali delle modifiche.
* **Implementazione:** Sviluppare un framework che applichi l'editing granulare della conoscenza su specifiche parti del modello, valutando quantitativamente l'impatto sulle prestazioni e qualitativamente l'efficacia etica.
* **Valutazione:** Testare l'efficacia delle modifiche attraverso benchmark standardizzati per bias e performance, affiancando un'analisi qualitativa delle implicazioni sociali delle modifiche.

#### Esempio di Applicazione:

* Modificare la rappresentazione interna di concetti culturali o di genere all'interno di un LLM per ridurre i bias, mantenendo l'accuratezza complessiva in attività di NLP come la traduzione automatica o l'analisi del sentiment. Valutare quantitativamente l'impatto delle modifiche e qualitativamente l'accettabilità sociale delle rappresentazioni corrette.

#### Risultati Attesi:

* Un sistema di editing granulare della conoscenza che migliora l’equità dei LLM, confermato da un'analisi quantitativa delle performance e una valutazione qualitativa dell’impatto sociale ed etico.

Il metodo "Fairness Stamp" (FAST) è un approccio emergente utilizzato nel contesto dei modelli di linguaggio di grandi dimensioni (LLM) per ridurre i bias senza compromettere le prestazioni complessive del modello. Sebbene il termine possa sembrare nuovo, si inserisce in una più ampia tradizione di tecniche volte a migliorare l'equità e la correttezza delle risposte generate da questi modelli.

### Cosa Fa FAST?

Il metodo FAST mira a correggere specifici bias all'interno di un modello linguistico, come pregiudizi di genere, razziali o altri tipi di parzialità sociale, attraverso un processo mirato di "modifica della conoscenza" (knowledge editing). Questo approccio consente di identificare le aree problematiche nel modello e intervenire su di esse, mantenendo inalterata la capacità del modello di eseguire altre funzioni linguistiche.

### Come Funziona?

Il metodo FAST si basa su una serie di passaggi:

1. **Identificazione dei Bias**: Utilizza metriche avanzate e dataset specifici per individuare le aree del modello in cui i bias sono presenti. Questi possono includere pregiudizi impliciti che emergono durante la generazione del testo.
2. **Intervento Mirato**: Una volta identificati i bias, FAST interviene in modo mirato per "correggere" queste conoscenze, spesso senza necessità di riaddestrare l'intero modello. Ciò potrebbe includere la modifica di pesi specifici nel modello o l'integrazione di dati correttivi durante la fase di fine-tuning.
3. **Valutazione e Validazione**: Dopo l'applicazione delle correzioni, il modello viene rigorosamente testato per garantire che le modifiche abbiano effettivamente ridotto il bias senza compromettere le sue prestazioni generali.

### Vantaggi di FAST

Uno dei principali vantaggi di FAST è che consente un intervento preciso e limitato, riducendo la necessità di enormi risorse computazionali per riaddestrare i modelli. Inoltre, preserva le prestazioni del modello in altri contesti, assicurando che la riduzione del bias non si traduca in una perdita di qualità delle risposte o in un degrado delle capacità linguistiche generali.

Questa tecnica è parte di un movimento più ampio nella ricerca sull'equità nei LLM, che include l'uso di metriche composite per misurare i bias (come il Large Language Model Bias Index, LLMBI) e l'applicazione di strategie di selezione di dimostrazioni rappresentative per migliorare la correttezza durante l'apprendimento in contesto (in-context learning)​ ([ar5iv](https://ar5iv.org/abs/2309.00770))​ ([Hugging Face](https://huggingface.co/papers/2312.14769" \t "_blank)).

In sintesi, FAST rappresenta una promettente innovazione per affrontare i bias nei modelli di linguaggio, rendendoli più equi e affidabili senza sacrificare la loro efficacia complessiva.

**Rimozione del Bias nei Modelli di Linguaggio Basata sulla Teoria Bayesiana**

#### **Obiettivo:**

* Questo progetto esplora l'utilizzo della teoria bayesiana per l'identificazione e la rimozione del bias implicito nei LLM, integrando un'analisi quantitativa dell'efficacia del metodo BTBR e una valutazione qualitativa dell'impatto sui modelli e sugli utenti.

#### **Domande di Ricerca:**

* Come può la teoria bayesiana essere applicata per identificare bias impliciti nei LLM?
* Quali vantaggi offre il BTBR rispetto ad altri approcci tradizionali di mitigazione del bias?
* Come può questo approccio essere integrato nei processi di addestramento e aggiornamento dei LLM?

#### **Metodologia:**

* **Sviluppo Teorico:** Esplorare e formalizzare l'applicazione della teoria bayesiana alla rimozione del bias nei LLM, integrando un'analisi quantitativa della sua efficacia rispetto ai metodi tradizionali.
* **Implementazione:** Creare un prototipo di BTBR e valutare quantitativamente la sua capacità di identificare e correggere bias, affiancando un'analisi qualitativa delle implicazioni etiche e sociali delle correzioni.
* **Validazione:** Testare l'efficacia del BTBR su dataset di benchmark, confrontando quantitativamente i risultati con metodi tradizionali e qualitativamente l’impatto sociale delle correzioni.

#### **Esempio di Applicazione:**

* Utilizzare il BTBR per correggere bias impliciti nel modello che influenzano il trattamento di gruppi demografici specifici nelle previsioni di sentiment analysis o generazione di testo. Valutare quantitativamente l'efficacia della correzione e qualitativamente l'accettabilità sociale e culturale dei risultati.

#### **Risultati Attesi:**

* Un metodo basato sulla teoria bayesiana che migliora la precisione e l'equità del modello, supportato da un'analisi quantitativa dell'efficacia e una valutazione qualitativa dell’impatto sociale.

Il metodo "Bayesian-Theory based Bias Removal" (BTBR) è un approccio innovativo per l'identificazione e la rimozione dei bias impliciti nei modelli di linguaggio di grandi dimensioni (LLM). Questo metodo si basa sulla teoria bayesiana per sviluppare un quadro sistematico che affronta i bias che possono sfuggire ai metodi di valutazione tradizionali.

### Come Funziona BTBR?

1. **Identificazione del Bias Implicito**: BTBR utilizza il concetto di "screening del rapporto di verosimiglianza" per analizzare le voci dei dati pubblici che potrebbero contenere bias incorporati durante la fase di addestramento del modello. Questo screening consente di identificare le informazioni che sono potenzialmente distorte da pregiudizi impliciti.
2. **Costruzione di Tripli di Conoscenza**: Una volta identificati i dati problematici, BTBR costruisce automaticamente tripli di conoscenza (strutture di dati che rappresentano relazioni tra entità) per individuare e isolare le parti del modello influenzate dal bias.
3. **Rimozione del Bias**: Il modello viene quindi modificato utilizzando tecniche di "model editing", che consentono di espungere le informazioni di bias senza dover riaddestrare completamente il modello. Questo approccio permette di mantenere le prestazioni generali del modello, correggendo solo le parti che presentano distorsioni.

### Vantaggi del BTBR

* **Precisione e Efficienza**: BTBR consente di intervenire in modo mirato, riducendo il bias senza impattare negativamente sulle prestazioni generali del modello. Questo lo rende un approccio molto efficace per affrontare pregiudizi sottili e difficili da rilevare.
* **Applicazione Automatica**: Grazie alla costruzione automatica di tripli di conoscenza e al processo di editing del modello, BTBR può essere implementato senza necessitare di interventi manuali estesi.

Questo metodo rappresenta un significativo passo avanti nella mitigazione dei bias impliciti nei LLM, rendendoli più equi e allineati con standard etici e sociali più elevati.

[Paper di riferimento](https://arxiv.org/abs/2408.10608)​ ([ar5iv](https://ar5iv.org/html/2408.10608v1))​ ([ar5iv](https://ar5iv.org/pdf/2408.10608)).

### ****Rimozione del Bias nei LLM Attraverso la Teoria Bayesiana: Una Valutazione Empirica e Qualitativa della Trasparenza e Spiegabilità****

#### **Obiettivo:**

Valutare l'efficacia dell'applicazione della teoria bayesiana per la riduzione del bias nei modelli linguistici di grandi dimensioni (LLM), accompagnata da un'analisi qualitativa sulla trasparenza e spiegabilità delle decisioni del modello.

#### **Domande di Ricerca:**

1. **In che modo l'integrazione della teoria bayesiana nei LLM può ridurre il bias rispetto ai metodi tradizionali di debiasing?**
2. **Quali sono le differenze nelle prestazioni dei LLM quando viene applicato un approccio bayesiano rispetto ai metodi tradizionali, in termini di accuratezza, precisione e tasso di errore?**
3. **Come viene percepita la trasparenza e la spiegabilità delle decisioni bayesiane da parte di utenti e specialisti rispetto ai modelli tradizionali?**
4. **Quali sono le principali sfide legate alla spiegabilità dei modelli bayesiani e come possono essere affrontate per migliorare la fiducia e l'accettabilità delle decisioni?**

#### **Metodologia Quantitativa:**

* **Implementazione Bayesiana:** Implementazione di un framework bayesiano all'interno di modelli linguistici di grandi dimensioni (LLM), progettato per aggiornare continuamente le probabilità a posteriori al fine di ridurre il bias. Gli esperimenti saranno condotti utilizzando benchmark standardizzati per la misurazione del bias, come GLUE (General Language Understanding Evaluation) o SQuAD (Stanford Question Answering Dataset). Questi benchmark permetteranno di valutare l'efficacia del metodo bayesiano in diversi contesti applicativi.
* **Analisi Comparativa:** Confronto dei risultati ottenuti con quelli di metodi tradizionali di debiasing. Saranno utilizzate metriche come accuratezza, precisione, recall e tasso di errore per determinare l’efficacia complessiva del metodo bayesiano nella mitigazione del bias, analizzando come questo approccio influenzi la performance generale del modello rispetto ai metodi esistenti.

#### **Metodologia Qualitativa:**

* **Studio di Caso:** Analisi di casi reali in cui i LLM basati sulla teoria bayesiana vengono utilizzati in contesti critici, come applicazioni mediche (ad esempio, supporto alla diagnosi) o legali (ad esempio, analisi di documenti). Lo studio di caso esplorerà le sfide legate alla trasparenza e alla spiegabilità delle decisioni del modello, valutando come queste caratteristiche influenzano la fiducia degli utenti e l’adozione del modello.
* **Focus Group e Interviste:** Raccolta di feedback da utenti e specialisti attraverso focus group e interviste, per valutare l’impatto della trasparenza e della spiegabilità sulle decisioni prese dai modelli bayesiani. Si esamineranno le differenze di percezione tra le decisioni basate su modelli bayesiani e quelle basate su metodi tradizionali, cercando di capire come la spiegabilità influenzi la fiducia e l'accettabilità delle decisioni in contesti applicativi reali.

#### **Esempi di Applicazione:**

* **Supporto alla Diagnosi Medica:** Implementazione del framework bayesiano in modelli utilizzati per supportare decisioni mediche, come la diagnosi di malattie basate su sintomi o immagini mediche. L'obiettivo sarà ridurre il bias diagnostico e migliorare la trasparenza delle decisioni, garantendo che i medici possano comprendere e fidarsi delle raccomandazioni del modello.
* **Analisi Legale Automatizzata:** Utilizzo del metodo bayesiano in applicazioni legali, come l'analisi di contratti o la previsione di esiti legali, per ridurre il bias e aumentare la trasparenza nelle decisioni automatizzate. La spiegabilità sarà cruciale per assicurare che gli avvocati e i giudici possano comprendere e fidarsi delle analisi fornite dal modello.
* **Servizi di Consulenza Finanziaria:** Applicazione del framework bayesiano in sistemi di consulenza finanziaria automatizzata, per garantire decisioni trasparenti e prive di bias, migliorando la fiducia dei clienti nelle raccomandazioni finanziarie generate dai modelli.

#### **Risultato Atteso:**

* **Quantitativo:** Dimostrare empiricamente che l'approccio bayesiano è più efficace nel ridurre il bias nei LLM rispetto ai metodi tradizionali, con risultati misurabili su benchmark standardizzati. Si prevede che il modello bayesiano non solo ridurrà il bias, ma manterrà anche, o migliorerà, l'accuratezza e la precisione del modello.
* **Qualitativo:** Identificazione delle principali sfide nella spiegabilità dei modelli bayesiani, con lo sviluppo di raccomandazioni per migliorare la trasparenza e l’accettabilità di questi modelli in applicazioni pratiche. L'obiettivo è evidenziare le differenze di percezione tra utenti e specialisti, fornendo una guida su come migliorare la comunicazione e la fiducia nei modelli basati sulla teoria bayesiana.

### ****Analisi del Bias e Trasparenza nei LLM Utilizzando la Teoria Bayesiana: Un Approccio Multi-Livello per la Riduzione del Bias in Scenari Complessi****

#### **Obiettivo:**

Valutare l'efficacia di un approccio multi-livello basato sulla teoria bayesiana per la riduzione del bias nei modelli linguistici di grandi dimensioni (LLM), esplorando come la trasparenza delle decisioni possa essere migliorata attraverso spiegazioni generate automaticamente.

#### **Domande di Ricerca:**

1. **Come può essere implementato un approccio multi-livello basato sulla teoria bayesiana nei LLM per monitorare e ridurre il bias in scenari complessi?**
2. **Quali sono le differenze in termini di riduzione del bias tra il modello bayesiano multi-livello e gli approcci tradizionali di debiasing nei LLM?**
3. **In che modo le spiegazioni generate automaticamente possono migliorare la trasparenza delle decisioni prese dai LLM e aumentare la fiducia degli utenti in contesti critici?**
4. **Quali sono le implicazioni pratiche e le sfide etiche nell'implementazione di modelli bayesiani multi-livello in applicazioni reali come il settore legale o medico?**

#### **Metodologia Quantitativa:**

* **Approccio Multi-Livello:** Implementazione di un modello bayesiano multi-livello nei LLM, progettato per monitorare e correggere i bias a diversi livelli di granularità (ad esempio, parola, frase, contesto intero). Il modello sarà testato su dataset complessi e sensibili, come testi legali e medici, dove i bias possono avere conseguenze significative.
* **Valutazione Empirica:** Confronto dell'efficacia del modello bayesiano multi-livello con approcci tradizionali di debiasing utilizzando metriche come il bilanciamento di classe, la precisione, la recall, e l'equità inter-classe. Questo confronto avverrà su scenari complessi, per determinare la capacità del modello di ridurre bias insidiosi e di migliorare la rappresentatività dei dati.

#### **Metodologia Qualitativa:**

* **Spiegazioni Generate Automaticamente:** Implementazione di un sistema di spiegazioni generate automaticamente per ogni decisione del modello. Queste spiegazioni saranno valutate attraverso focus group con esperti in campi critici (es. diritto, medicina) per analizzare la loro efficacia, chiarezza e capacità di migliorare la comprensione delle decisioni del modello.
* **Studio di Caso:** Conduzione di uno studio di caso in cui il modello bayesiano multi-livello viene applicato a decisioni critiche (ad esempio, consulenza legale automatizzata o supporto decisionale medico). Lo studio esaminerà come la trasparenza e le spiegazioni influenzano la fiducia degli utenti e la loro capacità di prendere decisioni informate basate sugli output del modello.

#### **Esempi di Applicazione:**

* **Consulenza Legale Automatizzata:** Applicazione del modello bayesiano multi-livello per analizzare documenti legali e offrire consulenza, riducendo il bias che potrebbe influenzare le interpretazioni legali, e generando spiegazioni trasparenti per le decisioni prese dal modello, migliorando la fiducia e la comprensione da parte dei professionisti legali.
* **Supporto Decisionale in Medicina:** Utilizzo del modello bayesiano multi-livello per analizzare cartelle cliniche e suggerire diagnosi o trattamenti, con un focus sulla riduzione del bias che potrebbe influenzare le raccomandazioni mediche, e fornendo spiegazioni chiare per ogni suggerimento, aiutando i medici a prendere decisioni più informate.
* **Gestione delle Risorse Umane:** Implementazione del modello per supportare decisioni in ambito HR, come selezione del personale o promozioni, garantendo che le decisioni siano prive di bias relativi a genere, etnia o altre caratteristiche protette, e offrendo spiegazioni trasparenti che migliorino la percezione di equità tra i dipendenti.

#### **Risultato Atteso:**

* **Quantitativo:** Fornire prove empiriche che dimostrino la superiorità del modello bayesiano multi-livello nella riduzione del bias rispetto agli approcci tradizionali, in particolare in scenari complessi come quelli legali e medici, con miglioramenti misurabili nelle metriche di accuratezza e equità.
* **Qualitativo:** Ottenere insight dettagliati su come le spiegazioni generate automaticamente possano essere utilizzate per migliorare la trasparenza e la fiducia nei modelli basati sulla teoria bayesiana, con raccomandazioni pratiche per l'implementazione di questi modelli in contesti reali e sensibili.

**~~Integrazione di Considerazioni Etiche e Sociali nei Modelli di Linguaggio di Grandi Dimensioni~~**

#### **~~Obiettivo:~~**

* ~~Questa tesi esplora come integrare considerazioni etiche e sociali nello sviluppo e nell'utilizzo dei LLM, combinando un'analisi quantitativa delle metriche di bias con una valutazione qualitativa delle implicazioni etiche e culturali.~~

#### **~~Domande di Ricerca:~~**

* ~~Quali sono le implicazioni etiche e sociali dei bias nei LLM, e come possono essere quantificate?~~
* ~~Come possono essere integrate considerazioni etiche e sociali nei processi di sviluppo dei LLM?~~
* ~~In che modo i framework di rilevazione del bias possono essere adattati per tener conto delle differenze culturali e regionali?~~

#### **~~Metodologia:~~**

* **~~Revisione Etica e Sociale:~~** ~~Esaminare quantitativamente l'impatto sociale dei bias nei LLM e le attuali metodologie di mitigazione, identificando qualitativamente le lacune nell'integrazione delle considerazioni etiche.~~
* **~~Sviluppo di Framework:~~** ~~Creare framework che combinano metriche quantitative di rilevazione del bias con valutazioni qualitative etiche e sociali, adattati a contesti culturali specifici.~~
* **~~Applicazione e Test:~~** ~~Applicare questi framework a casi di studio reali, valutando quantitativamente l'impatto del bias su vari gruppi demografici e qualitativamente le implicazioni etiche e sociali.~~

#### **~~Esempio di Applicazione:~~**

* ~~Valutare l'effetto delle risposte di un LLM su gruppi demografici specifici, considerando quantitativamente le differenze culturali e regionali, e sviluppare qualitativamente raccomandazioni per migliorare l'equità e la rappresentatività del modello.~~

#### **~~Risultati Attesi:~~**

* ~~Framework che permettono di integrare considerazioni etiche e sociali nello sviluppo dei LLM, contribuendo a modelli di linguaggio più equi e responsabili, confermato da un'analisi quantitativa e qualitativa.~~

### ****~~Analisi del Bias Contestuale nei Modelli Multimodali attraverso Dataset Cross-Domain e Studio degli Effetti della Sycophancy Visiva~~****

#### **~~Obiettivo:~~**

~~Esaminare come il bias contestuale si manifesta nei modelli multimodali utilizzando un dataset cross-domain per quantificare il bias, combinando questa analisi con uno studio qualitativo sull'effetto della sycophancy visiva (ossia la tendenza dei modelli a generare risposte che riflettono il contesto visivo in modo eccessivamente conformista o lusinghiero).~~

#### **~~Domande di Ricerca:~~**

1. **~~In che modo il bias contestuale si manifesta nei modelli multimodali quando vengono esposti a contesti culturali, sociali e semantici differenti?~~**
2. **~~Quali sono le metriche più efficaci per quantificare il bias contestuale nei modelli multimodali, e come possono queste metriche essere utilizzate per identificare e mitigare tale bias?~~**
3. **~~In che modo la sycophancy visiva influenza le risposte dei modelli multimodali, e quali sono le conseguenze di questa tendenza sulla qualità e accuratezza delle risposte?~~**
4. **~~Quali strategie possono essere sviluppate per ridurre la sycophancy visiva e migliorare la robustezza dei modelli multimodali in contesti variabili?~~**

#### **~~Metodologia Quantitativa:~~**

* **~~Dataset Cross-Domain:~~** ~~Creazione e utilizzo di un dataset che include immagini e testi provenienti da vari contesti culturali, sociali e semantici. Il dataset sarà progettato per esporre il modello a una gamma diversificata di input, al fine di valutare come il bias contestuale influisce sulle risposte generate. Il bias sarà misurato utilizzando metriche quantitative, come la divergenza Kullback-Leibler, per analizzare le deviazioni nelle risposte del modello in relazione ai diversi contesti.~~
* **~~Esperimenti Controllati:~~** ~~Confronto delle risposte del modello su input visivi simili ma con contesti testuali differenti per quantificare la presenza e l’entità del bias contestuale. Saranno misurate anche le variazioni di accuratezza e coerenza del modello in questi scenari, valutando come le risposte variano in funzione del contesto testuale associato all'immagine.~~

#### **~~Metodologia Qualitativa:~~**

* **~~Analisi delle Risposte del Modello:~~** ~~Condurre interviste semi-strutturate con esperti di intelligenza artificiale per discutere i risultati quantitativi ottenuti. Gli esperti offriranno insight critici su come i modelli multimodali interpretano e reagiscono al contesto visivo rispetto al contesto testuale, con un focus particolare sull'effetto della sycophancy visiva. Questo permetterà di identificare i meccanismi attraverso i quali il bias visivo si sviluppa e si manifesta nelle risposte del modello.~~
* **~~Focus Group:~~** ~~Coinvolgere utenti finali e specialisti per analizzare qualitativamente la percezione delle risposte del modello, soprattutto in scenari con forte presenza di bias visivo. I focus group serviranno a comprendere come il bias influisce sull'esperienza dell'utente e sulla percezione di affidabilità del modello, offrendo una prospettiva umana su come il bias e la sycophancy visiva possono essere percepiti e potenzialmente mitigati.~~

#### **~~Esempi di Applicazione:~~**

* **~~Assistenti Virtuali Multimodali:~~** ~~Valutazione del bias contestuale e della sycophancy visiva in assistenti virtuali che utilizzano input visivi e testuali per fornire raccomandazioni o risposte. L'obiettivo sarà migliorare la capacità dell'assistente di fornire risposte accurate e culturalmente appropriate, indipendentemente dal contesto visivo dominante.~~
* **~~Sistemi di Generazione di Contenuti Multimodali:~~** ~~Studio dell'impatto del bias contestuale e della sycophancy visiva in sistemi utilizzati per generare contenuti educativi o di intrattenimento, come video o articoli illustrati. L'obiettivo sarà ridurre la tendenza del modello a conformarsi eccessivamente al contesto visivo a discapito della coerenza testuale e della diversità culturale.~~
* **~~Piattaforme di Social Media:~~** ~~Analisi di come il bias contestuale e la sycophancy visiva influenzano la moderazione dei contenuti o la personalizzazione delle raccomandazioni su piattaforme social, con l'obiettivo di sviluppare modelli più equi e meno influenzati dai contesti visivi dominanti.~~

#### **~~Risultato Atteso:~~**

* **~~Quantitativo:~~** ~~Identificazione dei contesti in cui il bias contestuale è più pronunciato nei modelli multimodali, con suggerimenti quantitativi per la sua mitigazione. Questo potrebbe includere l'adattamento dei pesi nei modelli o l'introduzione di regole contestuali che permettano al modello di gestire meglio la variabilità dei contesti.~~
* **~~Qualitativo:~~** ~~Fornire una comprensione approfondita di come e perché i modelli multimodali sviluppano sycophancy visiva, con linee guida pratiche per migliorare la robustezza e l'affidabilità del modello in contesti variabili. Le raccomandazioni includeranno strategie per ridurre la tendenza dei modelli a generare risposte conformiste, migliorando così la diversità e l'accuratezza delle risposte.~~

### ****~~Trasparenza e Bias nei Modelli Multimodali: Un'Indagine Empirica e Qualitativa sui Limiti della Spiegabilità~~****

#### **~~Obiettivo:~~**

~~Esplorare la relazione tra bias e trasparenza nei modelli multimodali, valutando quantitativamente la presenza di bias e qualitativamente la capacità degli utenti di comprendere e fidarsi delle decisioni del modello.~~

#### **~~Domande di Ricerca:~~**

1. **~~In che modo la trasparenza influisce sulla capacità degli utenti di identificare e comprendere i bias nei modelli multimodali?~~**
2. **~~Quali tecniche di spiegabilità sono più efficaci nel rivelare bias nei modelli multimodali, e come queste influenzano la performance del modello?~~**
3. **~~Come variano la percezione della fiducia e la comprensione degli utenti in base ai diversi livelli di trasparenza offerti dai modelli multimodali?~~**
4. **~~Quali sono le implicazioni etiche e sociali di diversi livelli di trasparenza nei modelli multimodali, e come possono essere minimizzati i rischi associati a una trasparenza insufficiente?~~**

#### **~~Metodologia Quantitativa:~~**

* **~~Esperimenti di Trasparenza:~~** ~~Utilizzo di tecniche di spiegabilità, come LIME (Local Interpretable Model-agnostic Explanations) o SHAP (SHapley Additive exPlanations), per valutare come diversi livelli di trasparenza influiscono sulla capacità degli utenti di identificare il bias nei modelli multimodali. Gli esperimenti saranno condotti utilizzando un set di benchmark che include scenari noti per contenere potenziali bias.~~
* **~~Misurazione del Bias:~~** ~~Quantificazione del bias presente nei modelli multimodali in scenari con differenti livelli di trasparenza. Si utilizzeranno metriche di performance come F1-score, precisione, tassi di errore e altre metriche di equità per analizzare come la trasparenza influisca sulla variazione del comportamento del modello, e se l’aumento della trasparenza porti a una mitigazione del bias.~~

#### **~~Metodologia Qualitativa:~~**

* **~~Interviste e Focus Group:~~** ~~Coinvolgimento di utenti e specialisti attraverso interviste e focus group per esplorare le loro reazioni alle spiegazioni fornite dai modelli multimodali. Verranno analizzati come la trasparenza influisce sulla percezione della fiducia, sulla capacità di rilevare bias e sulle differenze di comprensione tra diverse categorie di utenti, come esperti e non esperti.~~
* **~~Studio Etico:~~** ~~Analisi delle implicazioni etiche della trasparenza nei modelli multimodali, considerando le aspettative degli utenti, i rischi associati a una trasparenza insufficiente o eccessiva, e le potenziali conseguenze sociali e legali. Saranno discusse diverse strategie di spiegabilità, con un focus su come bilanciare la trasparenza con la necessità di proteggere la privacy e prevenire l'uso improprio delle informazioni.~~

#### **~~Esempi di Applicazione:~~**

* **~~Sistemi di Diagnosi Medica:~~** ~~Applicazione delle tecniche di spiegabilità in modelli multimodali utilizzati per diagnosi mediche, valutando come la trasparenza influisce sulla fiducia dei medici e dei pazienti nelle decisioni del modello, e come queste decisioni possono essere influenzate dai bias.~~
* **~~Piattaforme di Raccomandazione Multimodale:~~** ~~Implementazione delle tecniche di spiegabilità su piattaforme di raccomandazione che utilizzano input multimodali (ad esempio, combinazioni di testo, immagini e dati utente) per comprendere come la trasparenza delle raccomandazioni possa influire sulla fiducia e sull'accettazione da parte degli utenti.~~
* **~~Sistemi di Sorveglianza e Sicurezza:~~** ~~Studio delle implicazioni etiche della trasparenza in sistemi di sorveglianza multimodale, esplorando come la spiegabilità possa aiutare a identificare e mitigare i bias, migliorando la fiducia del pubblico senza compromettere la sicurezza.~~

#### **~~Risultato Atteso:~~**

* **~~Quantitativo:~~** ~~Fornire prove empiriche su come la trasparenza influisca sulla rilevazione e mitigazione del bias nei modelli multimodali, identificando le tecniche di spiegabilità più efficaci. I risultati aiuteranno a comprendere meglio l’equilibrio tra trasparenza e performance, con implicazioni dirette per la progettazione di modelli più equi.~~
* **~~Qualitativo:~~** ~~Offrire una comprensione approfondita dei limiti della trasparenza nei modelli multimodali, con raccomandazioni pratiche per migliorare la spiegabilità e l’affidabilità percepita dai diversi tipi di utenti. Saranno proposte linee guida per minimizzare i rischi etici e sociali legati alla trasparenza, bilanciando la necessità di spiegabilità con altre considerazioni come la privacy e la sicurezza.~~

### ****~~Analisi dei Bias di Genere nei Modelli Linguistici: Un Approccio Comparativo~~****

#### **~~Obiettivo:~~**

~~Esaminare come i modelli linguistici generativi (LLM) gestiscono il bias di genere e confrontare l'efficacia di diverse tecniche di mitigazione, integrando analisi quantitative e qualitative per una comprensione più completa del fenomeno.~~

#### **~~Domande di Ricerca:~~**

1. **~~In che modo i modelli linguistici generativi (LLM) manifestano bias di genere nelle risposte generate?~~**
2. **~~Quali tecniche di mitigazione del bias di genere sono più efficaci nel ridurre la parzialità nei modelli linguistici?~~**
3. **~~Come viene percepito il bias di genere dagli utenti di diverse estrazioni sociali e culturali, e in che modo tali percezioni influenzano la fiducia e l'accettabilità del modello?~~**
4. **~~Quali sono i punti critici in cui emergono maggiormente i bias di genere, e come possono essere mitigati per migliorare la neutralità e l'inclusività dei modelli?~~**

#### **~~Metodologia Quantitativa:~~**

* **~~Identificazione del Bias di Genere:~~** ~~Utilizzo di un dataset testuale che include frasi e contesti culturalmente neutri (ad esempio, descrizioni di ruoli professionali o situazioni quotidiane) per osservare come i modelli linguistici generano risposte in base a contesti di genere. Verranno utilizzate frasi ambigue rispetto al genere per testare come i modelli assegnano ruoli o attributi.~~
* **~~Misurazione del Bias:~~** ~~Applicazione di metriche come la probabilità condizionale di associare determinati ruoli professionali, aggettivi o azioni a generi specifici. Verranno eseguite analisi statistiche per quantificare la presenza e l'intensità del bias, confrontando i risultati tra diversi modelli e tecniche di mitigazione (ad esempio, ri-addestramento con dati bilanciati, uso di filtri di post-processing, ecc.).~~

#### **~~Metodologia Qualitativa:~~**

* **~~Interviste con Utenti:~~** ~~Conduzione di interviste semi-strutturate con un gruppo di utenti di diverse estrazioni sociali, culturali e di genere. Le interviste mireranno a esplorare come gli utenti percepiscono il bias di genere nelle risposte generate dai modelli linguistici e a identificare differenze significative nella percezione del bias tra vari gruppi.~~
* **~~Analisi delle Risposte del Modello:~~** ~~Esecuzione di un'analisi qualitativa delle risposte generate dai modelli, concentrandosi su come la percezione del bias varia in diversi contesti e tra diversi gruppi di utenti. Verranno esplorate le differenze culturali e sociali che influenzano la sensibilità al bias di genere, identificando aree in cui i modelli falliscono nel mantenere la neutralità.~~

#### **~~Esempi di Applicazione:~~**

* **~~Chatbot per il Supporto Clienti:~~** ~~Valutazione dell'efficacia delle tecniche di debiasing in chatbot utilizzati per il supporto clienti, dove il bias di genere può influenzare negativamente l'esperienza utente e la fiducia nel sistema.~~
* **~~Sistemi di Reclutamento Automatico:~~** ~~Esplorazione dell'impatto del bias di genere in sistemi di raccomandazione o selezione automatica di candidati per ruoli professionali, dove il bias può portare a decisioni inique o discriminatorie.~~
* **~~Generazione di Contenuti Educativi:~~** ~~Analisi di come i modelli linguistici utilizzati per generare contenuti educativi (ad esempio, materiali didattici o spiegazioni automatizzate) possono introdurre bias di genere e come le tecniche di debiasing possono migliorare la neutralità e l'inclusività di tali contenuti.~~

#### **~~Risultato Atteso:~~**

* **~~Quantitativo:~~** ~~Identificazione dei punti critici in cui emergono bias di genere nei modelli linguistici, con dati empirici che mostrano l'efficacia delle diverse tecniche di mitigazione nel ridurre tale bias. Il confronto tra tecniche fornirà una chiara indicazione di quali approcci sono più efficaci in contesti specifici.~~
* **~~Qualitativo:~~** ~~Fornire insight sulle percezioni degli utenti riguardo al bias di genere, con raccomandazioni pratiche su come migliorare la neutralità e l'inclusività dei modelli linguistici, basate su una comprensione approfondita delle differenze culturali e sociali. Le raccomandazioni aiuteranno a sviluppare modelli più equi e inclusivi, capaci di operare in contesti globali diversificati.~~

### ****~~Impatto del Pre-processing dei Dati su Modelli Multimodali: Uno Studio Sperimentale~~****

#### **~~Obiettivo:~~**

~~Esaminare come diverse tecniche di pre-processing dei dati influenzano le prestazioni dei modelli multimodali, con un focus specifico su accuratezza, robustezza e generalizzazione.~~

#### **~~Domande di Ricerca:~~**

1. **~~Quali tecniche di pre-processing dei dati hanno l'impatto più significativo sull'accuratezza e robustezza dei modelli multimodali?~~**
2. **~~Come influiscono le tecniche di pre-processing sulla capacità di generalizzazione dei modelli multimodali quando vengono testati su dati non visti?~~**
3. **~~Quali sono le migliori pratiche di pre-processing per ottimizzare le prestazioni dei modelli multimodali in contesti applicativi reali?~~**
4. **~~In che modo gli esperti di machine learning interpretano e valutano i cambiamenti di performance dei modelli multimodali in funzione delle diverse tecniche di pre-processing applicate?~~**

#### **~~Metodologia Quantitativa:~~**

* **~~Implementazione delle Tecniche di Pre-processing:~~** ~~Applicazione di una serie di tecniche di pre-processing su un dataset multimodale. Le tecniche includeranno normalizzazione delle immagini, rimozione del rumore, augmentazione dei dati, ridimensionamento e altre pratiche comuni. Ogni tecnica sarà applicata separatamente e in combinazione per creare diverse versioni del dataset.~~
* **~~Valutazione delle Prestazioni:~~** ~~Utilizzo di metriche standard come accuratezza, precisione, richiamo, e F1-score per valutare l'impatto di ciascuna tecnica di pre-processing sulle prestazioni del modello. Verranno condotti test di robustezza su dati non visti per misurare la capacità del modello di generalizzare al di fuori del dataset di addestramento, valutando la resistenza a variazioni nei dati di input.~~

#### **~~Metodologia Qualitativa:~~**

* **~~Feedback degli Esperti:~~** ~~Conduzione di interviste con esperti di machine learning e data science per discutere i risultati quantitativi ottenuti. Le interviste si concentreranno sull'interpretazione dei risultati, le implicazioni delle diverse tecniche di pre-processing e su come queste possono essere ottimizzate per specifici casi d'uso.~~
* **~~Analisi dei Risultati:~~** ~~Esecuzione di un'analisi qualitativa approfondita dei cambiamenti osservati nelle prestazioni del modello, con particolare attenzione ai casi in cui il pre-processing ha avuto un impatto significativo, sia positivo che negativo. L'analisi si focalizzerà su come il pre-processing ha influenzato l'accuratezza del modello, la sua robustezza, e la capacità di generalizzazione, con lo scopo di individuare best practices.~~

#### **~~Esempi di Applicazione:~~**

* **~~Riconoscimento di Immagini Mediche:~~** ~~Applicazione di tecniche di pre-processing su dataset di immagini mediche (ad esempio, radiografie o risonanze magnetiche) per valutare come la qualità del pre-processing influenzi la capacità del modello di identificare correttamente patologie in immagini con diverse caratteristiche di rumore o qualità.~~
* **~~Analisi di Documenti Multimodali:~~** ~~Studio su come le tecniche di pre-processing influenzano le prestazioni dei modelli utilizzati per analizzare documenti multimodali che combinano testo e immagini, come moduli compilati a mano o documenti scansionati, per migliorare l'accuratezza del riconoscimento del testo e delle immagini.~~
* **~~Piattaforme di Social Media:~~** ~~Esplorazione dell'impatto del pre-processing su modelli multimodali utilizzati per analizzare contenuti di social media (combinazione di testo e immagini) per migliorare la moderazione dei contenuti, l'analisi dei sentimenti o la personalizzazione delle raccomandazioni.~~

#### **~~Risultato Atteso:~~**

* **~~Quantitativo:~~** ~~Fornire dati empirici che dimostrano l'effetto delle tecniche di pre-processing sui modelli multimodali, identificando le tecniche che migliorano significativamente accuratezza, robustezza e capacità di generalizzazione, e quelle che possono causare overfitting o degradare le prestazioni.~~
* **~~Qualitativo:~~** ~~Offrire una guida pratica basata su insight raccolti dagli esperti e dai risultati qualitativi per ottimizzare l'uso delle tecniche di pre-processing nei modelli multimodali, aiutando a migliorare le prestazioni del modello in applicazioni reali.~~

### ****~~Studio sull'Effetto delle Dimensioni del Dataset nell'Addestramento dei Modelli Linguistici~~****

#### **~~Obiettivo:~~**

~~Analizzare l'effetto delle dimensioni del dataset sull'accuratezza e la generalizzazione dei modelli linguistici, con un focus su modelli pre-addestrati di piccole e medie dimensioni, al fine di ottimizzare le risorse di training.~~

#### **~~Domande di Ricerca:~~**

1. **~~In che modo le dimensioni del dataset influenzano l'accuratezza e la capacità di generalizzazione dei modelli linguistici pre-addestrati?~~**
2. **~~Qual è la dimensione minima del dataset necessaria per raggiungere prestazioni ottimali, bilanciando l'accuratezza con l'efficienza delle risorse?~~**
3. **~~Come variano le interpretazioni semantiche e il comportamento del modello in funzione delle dimensioni del dataset utilizzato per l'addestramento?~~**
4. **~~Quali sono le percezioni degli utenti riguardo l'affidabilità e l'accuratezza delle risposte generate da modelli addestrati con dataset di dimensioni diverse?~~**

#### **~~Metodologia Quantitativa:~~**

* **~~Dimensionamento del Dataset:~~** ~~Utilizzo di versioni scalate di un dataset esistente (ad esempio, 10%, 50%, 100%) per addestrare un modello linguistico. Ogni versione del dataset verrà utilizzata per addestrare il modello, e le risposte saranno poi testate su un set di dati di test standardizzato per valutare le prestazioni.~~
* **~~Misurazione delle Prestazioni:~~** ~~Confronto delle prestazioni del modello addestrato su dataset di dimensioni diverse, utilizzando metriche come l'accuratezza, la capacità di generalizzazione (misurata su dati non visti) e l'overfitting (differenza tra le prestazioni su training e test set). Questo confronto aiuterà a identificare il punto di rendimento decrescente in cui aumentare la dimensione del dataset non porta a miglioramenti significativi nelle prestazioni del modello.~~

#### **~~Metodologia Qualitativa:~~**

* **~~Analisi del Comportamento del Modello:~~** ~~Analisi qualitativa delle risposte generate dai modelli addestrati con dataset di diverse dimensioni. L'obiettivo è esplorare come le variazioni nella dimensione del dataset influenzano le interpretazioni semantiche e il comportamento del modello, identificando eventuali tendenze o cambiamenti significativi nelle risposte.~~
* **~~Feedback degli Utenti:~~** ~~Raccolta di feedback da parte degli utenti che interagiscono con i modelli addestrati su dataset di dimensioni diverse. Attraverso questionari o interviste, gli utenti esprimeranno le loro percezioni riguardo all'accuratezza, affidabilità e utilità delle risposte generate, permettendo di correlare le dimensioni del dataset con l'esperienza utente.~~

#### **~~Esempi di Applicazione:~~**

* **~~Supporto alla Traduzione Automatica:~~** ~~Analisi di come la dimensione del dataset influenzi la qualità delle traduzioni automatiche in un sistema multilingue, determinando il dataset ottimale per bilanciare accuratezza e risorse computazionali.~~
* **~~Chatbot Aziendali:~~** ~~Esplorazione dell'effetto delle dimensioni del dataset su chatbot aziendali, utilizzati per fornire supporto ai clienti o assistenza interna, con l'obiettivo di ottimizzare l'addestramento per risorse limitate mantenendo alti livelli di accuratezza e soddisfazione dell'utente.~~
* **~~Sistemi di Raccomandazione:~~** ~~Studio sull'impatto delle dimensioni del dataset nei sistemi di raccomandazione, valutando come la dimensione del dataset influenzi la qualità delle raccomandazioni personalizzate, specialmente in contesti con un'ampia varietà di preferenze utente.~~

#### **~~Risultato Atteso:~~**

* **~~Quantitativo:~~** ~~Fornire dati empirici che dimostrano come le dimensioni del dataset influiscono sulle prestazioni dei modelli linguistici, identificando la dimensione ottimale per specifici casi d'uso, con particolare attenzione all'equilibrio tra accuratezza, generalizzazione e risorse di training.~~
* **~~Qualitativo:~~** ~~Fornire insight sulle variazioni nel comportamento del modello e nelle percezioni degli utenti in relazione alle dimensioni del dataset, con raccomandazioni pratiche su come ottimizzare la dimensione del dataset per migliorare la user experience e l'efficienza del modello.~~

### ****~~Spiegabilità nei Sistemi di Raccomandazione: Un Approccio Comparativo~~****

#### **~~Obiettivo:~~**

~~Confrontare diverse tecniche di spiegabilità nei sistemi di raccomandazione e valutarne l'impatto sulla fiducia, accettabilità e adozione da parte degli utenti.~~

#### **~~Domande di Ricerca:~~**

1. **~~Quali tecniche di spiegabilità sono più efficaci nel migliorare la fiducia degli utenti nei sistemi di raccomandazione?~~**
2. **~~In che misura la spiegabilità influisce sul tasso di accettazione delle raccomandazioni e sull'engagement degli utenti?~~**
3. **~~Come variano le percezioni degli utenti rispetto alla trasparenza e all'usabilità dei sistemi di raccomandazione in base al tipo di spiegazione fornita?~~**
4. **~~Quali sono le sfide principali nell'implementazione di tecniche di spiegabilità nei sistemi di raccomandazione, e come possono essere affrontate per massimizzare l'adozione da parte degli utenti?~~**

#### **~~Metodologia Quantitativa:~~**

* **~~Implementazione delle Tecniche di Spiegabilità:~~** ~~Integrazione di diverse tecniche di spiegabilità in un sistema di raccomandazione esistente. Le tecniche possono includere spiegazioni basate su feature (ad esempio, perché un prodotto è stato raccomandato in base a certe caratteristiche), spiegazioni controfattuali (ad esempio, "se avessi scelto X invece di Y, ti sarebbe stato raccomandato Z"), e altre tecniche di spiegazione comuni nei sistemi di raccomandazione.~~
* **~~Valutazione dell'Impatto:~~** ~~Misurazione quantitativa dell'impatto di ciascuna tecnica di spiegabilità sull'accuratezza delle raccomandazioni e sulla soddisfazione degli utenti. Verranno utilizzate metriche come il tasso di accettazione delle raccomandazioni, l'engagement degli utenti (ad esempio, tempo speso sulla piattaforma, interazioni con raccomandazioni), e il churn (abbandono del sistema). Queste metriche aiuteranno a determinare quali tecniche di spiegabilità hanno il maggiore impatto positivo sulle prestazioni del sistema.~~

#### **~~Metodologia Qualitativa:~~**

* **~~Interviste con Utenti:~~** ~~Conduzione di interviste semi-strutturate con utenti del sistema per raccogliere feedback su come percepiscono le diverse spiegazioni fornite. Le interviste mireranno a esplorare come le spiegazioni influenzano la fiducia, la percezione di trasparenza e la disposizione ad accettare le raccomandazioni.~~
* **~~Analisi delle Percezioni:~~** ~~Esecuzione di un'analisi qualitativa delle percezioni degli utenti, esplorando come la spiegabilità delle raccomandazioni influisca sulla loro fiducia nel sistema e sulla facilità d'uso. Saranno raccolte anche le opinioni degli utenti su come le spiegazioni potrebbero essere migliorate per aumentare ulteriormente la loro soddisfazione e adozione.~~

#### **~~Esempi di Applicazione:~~**

* **~~E-commerce:~~** ~~Implementazione di tecniche di spiegabilità in un sistema di raccomandazione di un sito di e-commerce per valutare come le spiegazioni influenzano la fiducia degli utenti nelle raccomandazioni di prodotti e l'effetto che ciò ha sul tasso di acquisto e sulla fedeltà al sito.~~
* **~~Piattaforme di Streaming:~~** ~~Applicazione delle tecniche di spiegabilità in una piattaforma di streaming video o musicale per vedere come le spiegazioni delle raccomandazioni influiscono sull'engagement degli utenti, ad esempio se sono più propensi a esplorare nuovi contenuti suggeriti dal sistema.~~
* **~~Servizi di Consulenza Personalizzata:~~** ~~Utilizzo di spiegazioni in sistemi di raccomandazione impiegati per consulenze personalizzate (ad esempio, consigli finanziari o di salute) per valutare come la trasparenza influisca sulla fiducia dell'utente nel seguire i consigli forniti.~~

#### **~~Risultato Atteso:~~**

* **~~Quantitativo:~~** ~~Identificazione delle tecniche di spiegabilità che migliorano maggiormente il tasso di accettazione delle raccomandazioni, l'engagement e la retention degli utenti, dimostrando l'importanza della trasparenza nei sistemi di raccomandazione.~~
* **~~Qualitativo:~~** ~~Fornire insight su come le spiegazioni influenzano la percezione degli utenti in termini di fiducia, trasparenza e usabilità, con raccomandazioni pratiche su come implementare le tecniche di spiegabilità più efficaci in scenari reali per ottimizzare l'esperienza utente.~~

### ****~~Analisi del Ruolo del Contesto nelle Interazioni Multimodali: Uno Studio di Caso con Assistenti Virtuali~~****

#### **~~Obiettivo:~~**

~~Indagare come il contesto influenzi l'efficacia delle interazioni tra utenti e assistenti virtuali multimodali, al fine di migliorare la progettazione e l’adattabilità degli assistenti in contesti variabili.~~

#### **~~Domande di Ricerca:~~**

1. **~~In che modo il contesto specifico (es. casa, lavoro, mobilità) influisce sulla qualità e precisione delle risposte fornite dagli assistenti virtuali multimodali?~~**
2. **~~Quali contesti risultano più critici per l'efficacia degli assistenti virtuali, e quali adattamenti possono migliorare le performance in questi contesti?~~**
3. **~~Come varia la soddisfazione dell'utente e la percezione dell'efficacia dell'assistente virtuale in base al contesto di utilizzo?~~**
4. **~~Quali elementi contestuali devono essere considerati prioritari nella progettazione di assistenti virtuali multimodali per ottimizzare l'esperienza utente in scenari diversificati?~~**

#### **~~Metodologia Quantitativa:~~**

* **~~Definizione del Contesto:~~** ~~Selezione di diversi contesti di utilizzo (es. casa, lavoro, mobilità) in cui gli utenti interagiscono con assistenti virtuali. In ogni contesto, verranno definiti scenari tipici di utilizzo per valutare come l'assistente risponde in ciascuno di essi.~~
* **~~Valutazione delle Risposte:~~** ~~Misurazione delle prestazioni dell'assistente virtuale in ciascun contesto, utilizzando metriche quantitative come la precisione delle risposte, il tempo di completamento delle attività, e la capacità dell'assistente di adattarsi alle esigenze contestuali. Verranno analizzate le differenze di performance per identificare i contesti più sfidanti.~~

#### **~~Metodologia Qualitativa:~~**

* **~~Osservazione Diretta:~~** ~~Conduzione di sessioni di utilizzo durante le quali gli utenti interagiscono con l'assistente virtuale in diversi contesti. Durante queste sessioni, verranno annotati i comportamenti degli utenti, le reazioni e le interazioni specifiche con l'assistente, per comprendere meglio come il contesto influenzi l'interazione.~~
* **~~Feedback degli Utenti:~~** ~~Raccogliere feedback attraverso questionari strutturati o interviste semi-strutturate con gli utenti, esplorando come il contesto influisca sulla loro soddisfazione e percezione dell'efficacia dell'assistente. Questo feedback aiuterà a identificare le aspettative e le esigenze specifiche degli utenti in vari contesti.~~

#### **~~Esempi di Applicazione:~~**

* **~~Assistenza Domestica:~~** ~~Valutare come l'assistente virtuale risponde a richieste domestiche (es. controllo della domotica, gestione della spesa) in un ambiente casalingo, e analizzare come le sue prestazioni possono essere migliorate per adattarsi meglio al contesto di vita quotidiana.~~
* **~~Supporto al Lavoro:~~** ~~Esaminare l'efficacia dell'assistente virtuale in un contesto di lavoro, ad esempio durante la gestione di riunioni, l'organizzazione di attività o la risposta a richieste professionali, con un focus su come l'assistente possa ottimizzare la produttività in un ambiente lavorativo.~~
* **~~Mobilità e Assistenti Virtuali:~~** ~~Studiare l'interazione con l'assistente virtuale durante situazioni di mobilità (es. guida, utilizzo di mezzi pubblici), valutando come l'assistente può adattarsi a contesti dinamici e rispondere efficacemente a richieste che variano in base al cambiamento di ambiente.~~

#### **~~Risultato Atteso:~~**

* **~~Quantitativo:~~** ~~Fornire una mappa delle performance dell'assistente virtuale in vari contesti, identificando i contesti in cui l'assistente è più o meno efficace, e suggerendo adattamenti specifici per migliorare le sue prestazioni.~~
* **~~Qualitativo:~~** ~~Comprensione dettagliata di come il contesto influenzi l'efficacia delle interazioni multimodali, con raccomandazioni pratiche su come progettare assistenti virtuali più adattabili e sensibili al contesto. Le raccomandazioni saranno basate su osservazioni dirette e feedback degli utenti, mirate a migliorare la user experience in diverse situazioni d’uso.~~

### ****Mitigazione del Bias Contestuale e Spiegabilità nei Modelli Multimodali Attraverso l'Implementazione di Framework Bayesiani e Dataset Multi-Dominio****

#### **Obiettivo:**

Integrare approcci bayesiani nei ~~modelli multimodali per ridurre il bias contestuale e migliorare la trasparenza delle decisioni, utilizzando dataset cross-domain che rappresentano diversi contesti culturali e sociali.~~

#### **~~Domande di Ricerca:~~**

1. **~~In che modo un framework bayesiano può essere implementato nei modelli multimodali per aggiornare dinamicamente le probabilità in risposta a input contestuali diversi?~~**
2. **~~Quali metriche sono più efficaci per misurare la riduzione del bias contestuale nei modelli multimodali che utilizzano un framework bayesiano?~~**
3. **~~In che misura la trasparenza delle decisioni, influenzata da un framework bayesiano, può migliorare la comprensibilità e la fiducia degli utenti nei modelli multimodali?~~**
4. **~~Quali sono le sfide principali nell'integrazione di un framework bayesiano nei modelli multimodali, e come possono essere affrontate per garantire un equilibrio tra complessità del modello e trasparenza per l'utente finale?~~**

#### **~~Metodologia Quantitativa:~~**

* **~~Framework Bayesiano per Multimodalità:~~** ~~Implementazione di un framework bayesiano nei modelli multimodali per aggiornare le probabilità delle decisioni in modo dinamico, basato su input contestuali provenienti da diverse fonti multimodali (immagini e testo). Il framework sarà progettato per ridurre il bias contestuale che potrebbe emergere da questi input.~~
* **~~Misurazione del Bias:~~** ~~Valutazione quantitativa dell'efficacia del framework bayesiano nella riduzione del bias contestuale. Saranno utilizzate metriche come l'accuratezza predittiva e l'area sotto la curva ROC (AUC-ROC) per misurare la precisione e l'equità delle decisioni del modello su dataset multi-dominio che includono immagini e testo da vari contesti culturali e sociali.~~

#### **~~Metodologia Qualitativa:~~**

* **~~Focus Group con Esperti di AI:~~** ~~Conduzione di focus group con esperti in intelligenza artificiale per discutere la comprensibilità e l'efficacia delle decisioni generate dal modello. Gli esperti analizzeranno le spiegazioni fornite dal modello, valutando come il framework bayesiano influisca sulla trasparenza e sulla percezione della decisione finale.~~
* **~~Analisi delle Risposte del Modello:~~** ~~Esecuzione di un'analisi qualitativa delle risposte del modello, concentrandosi su come la trasparenza delle decisioni (fornita dalle spiegazioni bayesiane) influenzi la fiducia e l'interpretazione delle decisioni da parte degli utenti finali. Saranno analizzati casi studio specifici per valutare l'impatto del framework sulla comprensibilità delle decisioni in situazioni reali.~~

#### **~~Esempi di Applicazione:~~**

* **~~Diagnosi Medica Multimodale:~~** ~~Applicazione del framework bayesiano in modelli utilizzati per diagnosticare condizioni mediche, combinando immagini diagnostiche (come radiografie) e dati testuali (come rapporti clinici). Il framework aiuterà a ridurre il bias contestuale, migliorando l'accuratezza delle diagnosi e la trasparenza delle raccomandazioni fornite ai medici.~~
* **~~Sistemi di Sorveglianza Multimodali:~~** ~~Utilizzo del framework per migliorare l'equità e la precisione in sistemi di sorveglianza che analizzano simultaneamente flussi video e dati testuali (ad esempio, descrizioni di incidenti). Il framework ridurrà il rischio di bias contestuali nelle decisioni di sorveglianza, assicurando che le interpretazioni siano più accurate e trasparenti.~~
* **~~Assistenti Virtuali Multimodali:~~** ~~Implementazione del framework bayesiano in assistenti virtuali che utilizzano input multimodali (testo, immagini) per interagire con utenti provenienti da diverse culture. Il sistema sarà progettato per comprendere meglio i contesti culturali, riducendo il bias contestuale e migliorando la trasparenza e l'affidabilità delle risposte fornite.~~

#### **~~Risultato Atteso:~~**

* **~~Quantitativo:~~** ~~Dimostrazione empirica che l'integrazione di un framework bayesiano nei modelli multimodali riduce significativamente il bias contestuale, con un miglioramento nelle metriche di accuratezza predittiva e AUC-ROC, senza compromettere la performance generale del modello.~~
* **~~Qualitativo:~~** ~~Una comprensione approfondita di come la trasparenza e la spiegabilità possano essere migliorate nei modelli multimodali attraverso l'uso di framework bayesiani, con raccomandazioni pratiche su come bilanciare la complessità del modello con la comprensibilità per l'utente finale, basate su feedback degli esperti e analisi qualitative.~~

### ****Sviluppo di Rappresentazioni Multivettoriali Dinamiche per Migliorare la Diversità Culturale e Socioeconomica nei Modelli Vision-Language****

#### **Obiettivo:**

Esplorare e sviluppare un sistema di rappresentazioni multivettoriali che si adattano dinamicamente ai contesti culturali e socioeconomici nei modelli Vision-Language (VLMs), migliorando l'equità e la rappresentatività degli output generati.

#### **Domande di Ricerca:**

1. **Come possono essere progettate rappresentazioni multivettoriali dinamiche per adattarsi efficacemente a contesti culturali e socioeconomici variabili nei modelli Vision-Language?**
2. **Quali sono i principali benefici delle rappresentazioni multivettoriali rispetto agli embeddings tradizionali nel contesto della riduzione del bias culturale e socioeconomico?**
3. **In che modo le rappresentazioni multivettoriali dinamiche influenzano la percezione e l'interpretazione dei risultati generati dai modelli Vision-Language da parte degli utenti di diverse culture e classi socioeconomiche?**
4. **Quali metriche sono più efficaci per valutare l'equità e la rappresentatività culturale dei modelli che utilizzano rappresentazioni multivettoriali dinamiche?**

#### **Metodologia Quantitativa:**

* **Creazione di Embeddings Multivettoriali:** Sviluppo di un framework per generare rappresentazioni multivettoriali che variano in base ai contesti culturali e socioeconomici. Il framework utilizzerà tecniche di embedding contestuale avanzato, dove i vettori saranno modulati per riflettere le specificità culturali e socioeconomiche del contesto d'uso.
* **Valutazione del Modello:** Implementazione del modello Vision-Language con rappresentazioni multivettoriali e test su dataset multimodali diversificati come COCO e Visual Genome. Le prestazioni del modello saranno misurate attraverso metriche di fairness (ad esempio, demographic parity, equal opportunity), accuratezza, e rappresentatività culturale, per verificare l'efficacia delle nuove rappresentazioni.

#### **Metodologia Qualitativa:**

* **Analisi delle Rappresentazioni:** Conduzione di analisi qualitative per esaminare come le rappresentazioni multivettoriali dinamiche influenzano le risposte del modello in vari contesti culturali e socioeconomici. L'analisi si concentrerà sulla coerenza culturale e l'inclusività delle risposte generate.
* **Focus Group con Esperti di Cultura e Socioeconomia:** Coinvolgimento di esperti in studi culturali e socioeconomici per valutare l'efficacia e la rappresentatività delle nuove rappresentazioni multivettoriali. I focus group discuteranno le implicazioni delle risposte generate dai modelli e forniranno raccomandazioni su come ottimizzare ulteriormente il sistema.

#### **Esempi di Applicazione:**

* **Generazione di Contenuti Culturali Sensibili:** Utilizzo delle rappresentazioni multivettoriali per generare descrizioni di immagini e contenuti testuali in applicazioni culturali, come guide turistiche o materiali educativi, assicurando che i contenuti siano culturalmente rilevanti e rispettosi delle diverse tradizioni.
* **Miglioramento delle Interazioni Multimodali in Assistenti Virtuali:** Applicazione delle rappresentazioni multivettoriali nei modelli Vision-Language utilizzati in assistenti virtuali, per migliorare la qualità delle risposte in contesti culturali specifici, come l'assistenza sanitaria in paesi con diverse pratiche mediche tradizionali.
* **Personalizzazione delle Esperienze Utente in Piattaforme Globali:** Implementazione delle rappresentazioni multivettoriali in piattaforme globali come social media o e-commerce, per personalizzare le esperienze utente in base a contesti culturali e socioeconomici specifici, aumentando l'engagement e la soddisfazione dell'utente.

#### **Risultato Atteso:**

* **Quantitativo:** Sviluppo di un modello Vision-Language che riduce significativamente il bias culturale e socioeconomico attraverso l'uso di rappresentazioni multivettoriali dinamiche, con miglioramenti misurabili nelle metriche di fairness e rappresentatività.
* **Qualitativo:** Fornire insight su come le rappresentazioni dinamiche possono essere utilizzate per creare modelli più equi e culturalmente sensibili, con raccomandazioni per ulteriori miglioramenti basati su feedback di esperti e analisi qualitative.

### ****Rappresentazioni Multivettoriali Dinamiche nei Modelli Linguistici di Grandi Dimensioni (LLM): Un Approccio Culturale e Socioeconomico****

#### **Obiettivo:**

Esplorare l'integrazione di rappresentazioni multivettoriali dinamiche nei modelli linguistici di grandi dimensioni (LLM) per migliorare la diversità e l'equità nelle risposte generate in contesti culturali e socioeconomici diversi.

#### **Domande di Ricerca:**

1. **In che modo le rappresentazioni multivettoriali dinamiche possono essere implementate nei LLM per adattarsi efficacemente a contesti culturali e socioeconomici variabili?**
2. **Quali vantaggi offrono le rappresentazioni multivettoriali dinamiche rispetto agli embeddings contestuali tradizionali in termini di equità e diversità delle risposte?**
3. **Come influenzano i multivettori la percezione e l'interpretazione delle risposte generate dai LLM da parte di utenti provenienti da diversi background culturali e socioeconomici?**
4. **Quali sono le sfide principali nell'integrazione di rappresentazioni multivettoriali dinamiche nei LLM, e come possono essere superate per migliorare l'inclusività del modello?**

#### **Metodologia Quantitativa:**

* **Implementazione di Embeddings Multivettoriali:** Sviluppo di un sistema di rappresentazioni multivettoriali dinamiche che si adattino a contesti specifici (ad esempio, culturali, geografici, socioeconomici), utilizzando tecniche avanzate di embeddings contestuali. Questo sistema sarà integrato nei LLM per generare risposte più rappresentative e inclusive.
* **Valutazione delle Risposte Generate:** Misurazione quantitativa delle risposte generate dai LLM attraverso metriche di fairness (ad esempio, equità di genere, equilibrio culturale) e rappresentatività. Il modello sarà testato su dataset che riflettono una vasta gamma di variazioni culturali e socioeconomiche per valutare l'efficacia delle rappresentazioni multivettoriali.

#### **Metodologia Qualitativa:**

* **Analisi Qualitativa delle Risposte:** Conduzione di un'analisi qualitativa delle risposte generate dal modello, focalizzandosi su come i multivettori influenzano la percezione del contenuto in contesti culturali e socioeconomici diversi. L'analisi esplorerà l'aderenza del modello ai valori culturali e la sensibilità alle variabili socioeconomiche.
* **Interviste con Utenti Finali:** Raccogliere feedback attraverso interviste con utenti provenienti da diverse culture e classi socioeconomiche. Le interviste mireranno a valutare l'accettabilità, la fiducia e la percezione dell'equità nelle risposte fornite dal modello multivettoriale, identificando eventuali aree di miglioramento.

#### **Esempi di Applicazione:**

* **Personalizzazione di Contenuti Educativi:** Utilizzare le rappresentazioni multivettoriali per adattare i contenuti educativi generati dai LLM a specifici contesti culturali e socioeconomici, garantendo che il materiale sia rilevante e rispettoso delle diverse tradizioni e valori.
* **Generazione di Risposte in Assistenza Virtuale:** Applicare le rappresentazioni multivettoriali per migliorare la qualità delle risposte fornite da assistenti virtuali in contesti globali, assicurando che le risposte siano culturalmente appropriate e socioeconomicamente sensibili in diverse regioni del mondo.
* **Supporto Multilingue e Multiculturale:** Integrare multivettori nei LLM utilizzati per la traduzione automatica o per chatbot multilingue, garantendo che le risposte riflettano accuratamente le sfumature culturali e le sensibilità socioeconomiche delle lingue e delle culture target.

#### **Risultato Atteso:**

* **Quantitativo:** Miglioramento significativo della diversità e dell'equità nelle risposte dei LLM, dimostrato attraverso un adattamento dinamico e contestualizzato delle rappresentazioni multivettoriali a contesti culturali e socioeconomici specifici.
* **Qualitativo:** Insight approfonditi su come le rappresentazioni multivettoriali dinamiche possono essere utilizzate per sviluppare modelli linguistici più inclusivi, sensibili alle diverse culture e classi sociali, con raccomandazioni pratiche per l'implementazione di queste tecniche nei LLM esistenti.

### ****Utilizzo di Reti Neurali Avversarie per la Mitigazione del Bias Culturale nei Modelli Vision-Language Multimodali****

#### **Obiettivo:**

Progettare e implementare un sistema di reti neurali avversarie (Adversarial Networks) per identificare e mitigare il bias culturale e socioeconomico nei modelli Vision-Language, migliorando l'equità e la precisione dei risultati.

#### **Domande di Ricerca:**

1. **In che modo le reti neurali avversarie possono essere progettate e ottimizzate per rilevare e mitigare efficacemente il bias culturale e socioeconomico nei modelli Vision-Language?**
2. **Quali sono le metriche più appropriate per valutare l'efficacia delle reti avversarie nella riduzione del bias nei modelli Vision-Language, e come possono essere applicate per ottenere risultati affidabili?**
3. **In che modo le tecniche avversarie influenzano la comprensione e l'interpretazione delle risposte da parte degli utenti finali, e come possono essere ottimizzate per mantenere un equilibrio tra equità e accuratezza?**
4. **Quali sono le principali sfide etiche nell'implementazione di reti avversarie nei modelli multimodali, e come possono essere affrontate per garantire la trasparenza e la fiducia degli utenti?**

#### **Metodologia Quantitativa:**

* **Sviluppo di un Modello Avversario:** Creazione di un sistema avversario specifico per il post-processing dei modelli Vision-Language, mirato a identificare e correggere bias culturali e socioeconomici. Il sistema utilizzerà una rete avversaria che apprende a riconoscere e penalizzare risposte che mostrano segni di bias, inducendo il modello principale a generare output più equi.
* **Valutazione del Debiasing:** Conduzione di esperimenti utilizzando dataset annotati con bias, come FairFace e Diversity in Faces. Le performance del modello saranno valutate tramite metriche di fairness (ad esempio, equalized odds, demographic parity), precisione, e diversità delle risposte, per misurare l'efficacia del debiasing introdotto dalle reti avversarie.

#### **Metodologia Qualitativa:**

* **Interviste con Esperti di Fairness in AI:** Conduzione di interviste con esperti di fairness e machine learning per ottenere feedback dettagliato sull'impatto delle tecniche avversarie sul bias nei modelli multimodali, esplorando le sfide etiche associate e le percezioni sulla trasparenza e affidabilità del sistema.
* **Analisi delle Risposte Corrette:** Analisi qualitativa delle risposte generate dal modello post-processato per esaminare come le tecniche di debiasing influenzano l'interpretazione e la comprensione delle immagini e dei testi da parte degli utenti finali. L'analisi si concentrerà su come le correzioni influenzano la percezione della neutralità e dell'equità nelle risposte.

#### **Esempi di Applicazione:**

* **Mitigazione del Bias nei Sistemi di Raccomandazione Visiva:** Applicare le reti avversarie per ridurre il bias nei sistemi che raccomandano contenuti visivi (ad esempio, in piattaforme di social media o e-commerce), garantendo che le raccomandazioni riflettano una rappresentazione equa e diversificata delle culture e delle identità.
* **Correzione di Descrizioni in Sistemi di Generazione Immagine-Testo:** Utilizzare le reti avversarie per migliorare la neutralità delle descrizioni generate in applicazioni di immagine-testo (ad esempio, descrizioni di immagini per non vedenti), correggendo automaticamente descrizioni che potrebbero riflettere pregiudizi culturali o socioeconomici.
* **Assistenza Multimodale in Contesti Sensibili:** Implementare le tecniche di debiasing in assistenti virtuali multimodali utilizzati in contesti culturali sensibili, come nel supporto alla traduzione o nella fornitura di informazioni educative, per assicurarsi che le risposte siano culturalmente rispettose e inclusivi.

#### **Risultato Atteso:**

* **Quantitativo:** Riduzione significativa del bias culturale e socioeconomico nei modelli Vision-Language attraverso l'uso di reti avversarie, dimostrata attraverso risultati su benchmark specifici, con miglioramenti nelle metriche di fairness e diversità delle risposte.
* **Qualitativo:** Sviluppo di linee guida etiche per l'integrazione delle reti avversarie nei modelli multimodali, con raccomandazioni pratiche su come bilanciare l'efficacia del debiasing con la trasparenza e la fiducia degli utenti, basate su analisi qualitative e feedback degli esperti.

### ****Integrazione di Reti Neurali Avversarie per il Post-Processing nei Modelli Linguistici di Grandi Dimensioni (LLM): Un Approccio alla Mitigazione del Bias****

#### **Obiettivo:**

Esplorare l'uso delle reti neurali avversarie nel post-processing dei LLM per identificare e correggere bias culturali e socioeconomici, migliorando l'equità e la trasparenza delle risposte generate.

#### **Domande di Ricerca:**

1. **Come possono le reti neurali avversarie essere progettate per rilevare e mitigare efficacemente i bias culturali e socioeconomici nei LLM durante il post-processing?**
2. **Qual è l'impatto dell'integrazione di reti avversarie sulle prestazioni generali dei LLM, in termini di accuratezza e neutralità delle risposte?**
3. **In che modo l'utilizzo di spiegazioni generate automaticamente può migliorare la trasparenza delle decisioni prese dai modelli linguistici?**
4. **Quali sono le implicazioni etiche e sociali dell'uso di reti neurali avversarie nel post-processing dei LLM, e come possono essere affrontate per garantire un uso responsabile della tecnologia?**

#### **Metodologia Quantitativa:**

* **Sviluppo di Reti Avversarie:** Progettazione di reti avversarie specifiche per il post-processing dei LLM, con l'obiettivo di rilevare e correggere bias culturali e socioeconomici nelle risposte generate. Questo includerà la creazione di un generatore (LLM) e un discriminatore (rete avversaria) che lavorano in tandem per migliorare l'equità delle risposte.
* **Valutazione dell'Efficacia:** Conduzione di esperimenti su dataset annotati con bias (ad esempio, dataset contenenti bias razziali, di genere o socioeconomici). Le prestazioni del sistema saranno valutate utilizzando metriche di fairness, accuratezza e neutralità delle risposte, confrontando i risultati pre- e post-processamento.

#### **Metodologia Qualitativa:**

* **Analisi delle Spiegazioni Generate:** Studio delle spiegazioni generate automaticamente dalle reti avversarie durante il post-processing, con un focus su come queste spiegazioni possono migliorare la trasparenza e la comprensibilità delle decisioni del modello per gli utenti finali.
* **Interviste con Esperti di Etica in AI:** Conduzione di interviste con esperti per discutere le implicazioni etiche dell'uso di reti avversarie nei LLM. Queste interviste mireranno a sviluppare linee guida per l'integrazione responsabile delle reti avversarie, assicurando che la riduzione del bias non comprometta la fiducia degli utenti.

#### **Esempi di Applicazione:**

* **Correzione di Risposte Stereotipate:** Supponiamo che un LLM tenda a generare risposte che stereotipano determinate professioni in base al genere (ad esempio, "un'infermiera è gentile"). La rete avversaria potrebbe essere utilizzata per identificare queste risposte e riformularle in modo che siano neutre rispetto al genere, migliorando l'equità delle risposte senza sacrificare la qualità.
* **Mitigazione del Bias in Contesti Legali o Medici:** In applicazioni critiche come la consulenza legale o medica, dove il bias culturale può avere gravi conseguenze, le reti avversarie potrebbero essere utilizzate per garantire che le risposte generate dai LLM siano neutrali e rispettose delle diversità culturali, aumentando così la fiducia e l'affidabilità del sistema.
* **Generazione di Contenuti Educativi Inclusivi:** Nei sistemi utilizzati per generare contenuti educativi, le reti avversarie possono aiutare a garantire che il materiale prodotto non rifletta pregiudizi socioeconomici o culturali, promuovendo un'educazione più inclusiva e diversificata.

#### **Risultato Atteso:**

* **Quantitativo:** Dimostrazione che l'integrazione di reti avversarie nel post-processing dei LLM può ridurre efficacemente il bias culturale e socioeconomico senza compromettere la qualità delle risposte, come dimostrato dalle metriche di valutazione su diversi dataset.
* **Qualitativo:** Sviluppo di raccomandazioni pratiche e linee guida etiche per l'uso delle reti avversarie nei LLM, bilanciando l'efficacia del debiasing con la necessità di mantenere la fiducia e la trasparenza per gli utenti finali. Le interviste con esperti offriranno insight su come implementare queste tecnologie in modo responsabile e trasparente.

**RACCOMANDAZIONI:**

In queste aree tematiche lo stato dell’arte è da evolvere

### ****1. Focus sul Bias Contestuale****

**Descrizione:** Il bias contestuale si riferisce a distorsioni nei modelli linguistici che emergono in risposta a specifici contesti piuttosto che attraverso tendenze generali presenti nei dati. Questi bias possono essere sottili e dipendenti dalle circostanze o dall'ambiente in cui il modello viene utilizzato, e possono non essere rilevati dalle metriche tradizionali di valutazione del bias che spesso si concentrano su bias più evidenti o generalizzati (ad esempio, il genere o l'etnia).z

**Cosa comporta lo sviluppo di metodi per rilevare e mitigare il bias contestuale:**

* **Identificazione del Bias Contestuale:** Implica l'analisi delle risposte dei modelli linguistici in una varietà di contesti per individuare se e come queste risposte cambiano in base al contesto in modo pregiudizievole. Ad esempio, un modello potrebbe generare risposte diverse a una stessa domanda posta in contesti culturali differenti, privilegiando involontariamente un contesto rispetto a un altro.
* **Sviluppo di Metriche Specifiche:** Le metriche tradizionali potrebbero non catturare la complessità del bias contestuale. Pertanto, è necessario sviluppare nuove metriche che possano quantificare queste sottili variazioni e distorsioni. Queste metriche potrebbero includere analisi della coerenza delle risposte in contesti variabili, oppure metriche che confrontano le risposte del modello in situazioni simili ma culturalmente diverse.
* **Mitigazione del Bias Contestuale:** Una volta identificato, il bias contestuale deve essere mitigato. Questo può comportare l'adattamento dei modelli attraverso tecniche di fine-tuning mirate, o l'introduzione di meccanismi che consentano al modello di gestire meglio le variazioni di contesto, garantendo risposte più neutrali e appropriate indipendentemente dal contesto.

### Comprendere il Bias Contestuale e il Paradosso dell'Adattamento al Contesto

1. **Bias Contestuale**: Il bias contestuale si riferisce a distorsioni nei modelli linguistici che emergono in risposta a specifici contesti d'uso. Questi bias non sono necessariamente dovuti a tendenze generali nei dati di addestramento (come bias di genere o etnici evidenti), ma piuttosto a come i modelli interpretano e rispondono a informazioni specifiche del contesto in modo potenzialmente pregiudizievole o inappropriato.
2. **Il Paradosso della Consapevolezza del Contesto**: Il dilemma sta nel fatto che, mentre vogliamo che i modelli di intelligenza artificiale siano **context-aware** e adattivi, il fatto che siano influenzati dal contesto può introdurre bias o comportamenti indesiderati. Ad esempio, un sistema potrebbe imparare che in determinati contesti culturali alcune risposte sono preferite rispetto ad altre e iniziare a dare priorità a queste risposte, ignorando la necessità di neutralità o equità.

### Perché il Bias Contestuale è Considerato un Problema?

Il bias contestuale è problematico perché può portare a:

* **Risposte Inappropriate o Non Neutrali**: Anche se un modello è progettato per essere consapevole del contesto, può accidentalmente rafforzare stereotipi o pregiudizi culturali. Ad esempio, in un contesto culturale diverso, il modello potrebbe rispondere in un modo che sembra offensivo o insensibile, anche se non era questa l'intenzione.
* **Rafforzamento di Stereotipi**: Quando un modello linguistico utilizza il contesto per personalizzare le risposte, potrebbe rafforzare inconsapevolmente stereotipi culturali o sociali. Ad esempio, potrebbe associare certi comportamenti o tratti a specifiche culture in modo inappropriato o non accurato.
* **Impatto Negativo sull'Equità**: Il bias contestuale può influire negativamente sull'equità dei sistemi di AI, specialmente quando questi sono utilizzati in applicazioni critiche come l'educazione, l'assistenza sanitaria, o il recruiting. In questi casi, risposte inadeguate o distorte possono avere conseguenze gravi.

### Perché Mitigare il Bias Contestuale Invece di Rimuovere la Consapevolezza del Contesto?

Mitigare il bias contestuale non significa rimuovere la capacità dei modelli di essere context-aware, ma piuttosto:

* **Rendere i Modelli Più Equi e Rappresentativi**: L'obiettivo è assicurarsi che i modelli possano adattarsi al contesto senza introdurre pregiudizi. In altre parole, vogliamo che il modello sia in grado di riconoscere e rispettare le diversità culturali e socioeconomiche senza favorire involontariamente un contesto rispetto a un altro.
* **Migliorare l'Inclusività e la Sensibilità Culturale**: Mitigando il bias contestuale, si cerca di evitare situazioni in cui il modello offenda o escluda determinate culture o gruppi sociali. Questo è particolarmente importante per applicazioni globali che servono un pubblico eterogeneo.
* **Garantire Risposte Appropriate e Neutrali**: In molti casi, la neutralità o un approccio equilibrato è preferibile. Per esempio, in un contesto di assistenza sanitaria, un modello dovrebbe fornire informazioni basate su evidenze scientifiche piuttosto che influenzate da particolari pratiche culturali che potrebbero non essere universali o sicure.

### Come Mitigare il Bias Contestuale

* **Sviluppo di Metriche Specifiche**: Creare nuove metriche per rilevare e quantificare il bias contestuale che tengano conto delle sottili differenze che possono emergere in vari contesti. Questo aiuta a misurare il bias in modo più preciso.
* **Fine-Tuning e Adattamento del Modello**: Utilizzare tecniche di fine-tuning mirate per addestrare il modello su dati che siano più equilibrati e rappresentativi di diverse culture e contesti socioeconomici.
* **Controllo dell'Output del Modello**: Introdurre meccanismi che permettano al modello di gestire meglio le variazioni di contesto, garantendo che le risposte siano appropriate e neutrali.

### Correzione delle Componenti Context-Aware

Correggere direttamente le componenti **context-aware** potrebbe essere un'azione complementare, ma richiede una comprensione approfondita di come il contesto influisce sulle risposte del modello. Alcune strategie includono:

* **Analisi delle Componenti Contextuali**: Valutare come le componenti che gestiscono il contesto influenzano il processo decisionale del modello e apportare modifiche per ridurre l'impatto dei bias.
* **Inclusione di Diversità nei Dati di Addestramento**: Assicurarsi che i dati di addestramento riflettano una gamma sufficientemente ampia di contesti culturali e socioeconomici per addestrare il modello a rispondere in modo equo e rappresentativo.

### Conclusione

Il bias contestuale è considerato un difetto perché può portare a risultati inappropriati o pregiudizievoli, anche quando il sistema è progettato per essere personalizzato e consapevole del contesto. Mitigare il bias contestuale non significa eliminare la capacità di essere context-aware, ma piuttosto affinare questa capacità per garantire che i modelli possano adattarsi ai contesti in modo equo, rappresentativo e rispettoso delle diversità culturali e socioeconomiche.

### ****2. Dataset Cross-Domain****

**Descrizione:** Un dataset cross-domain è un insieme di dati che include contenuti provenienti da una vasta gamma di domini (ad esempio, diversi settori, culture, lingue, stili di scrittura, ecc.) e rappresenta la diversità linguistica e culturale globale. L'obiettivo è creare un dataset che rifletta fedelmente le variegate realtà del mondo reale, piuttosto che essere limitato a un unico dominio o cultura.

**Cosa comporta la creazione e l'utilizzo di dataset cross-domain:**

* **Raccolta di Dati Diversificati:** La creazione di un dataset cross-domain richiede la raccolta di dati da molteplici fonti che rappresentano diverse lingue, culture, discipline, e contesti sociali. Ad esempio, un dataset potrebbe includere testi scientifici, notizie, social media, e letteratura provenienti da diverse regioni del mondo.
* **Bilanciamento del Dataset:** È importante che il dataset sia bilanciato in modo che nessun dominio, lingua, o cultura sia sovrarappresentato. Questo riduce il rischio che il modello sviluppi bias dovuti alla predominanza di un particolare contesto nei dati di addestramento.
* **Addestramento e Valutazione:** Utilizzare questo dataset per addestrare e valutare i modelli linguistici di grandi dimensioni (LLM) permette di sviluppare modelli che sono più robusti e capaci di generalizzare meglio attraverso i diversi domini. Inoltre, consente di valutare come i modelli performano su una varietà di contesti, identificando aree in cui potrebbero esserci pregiudizi o lacune.

### Come i Dataset Cross-Domain Aiutano nella Mitigazione del Bias

1. **Rappresentazione Diversificata dei Dati**:
   * **Descrizione**: I dataset cross-domain includono contenuti provenienti da una vasta gamma di fonti, settori, lingue, culture, stili di scrittura e contesti sociali. Questo approccio aiuta a garantire che i modelli siano esposti a una più ampia varietà di dati durante l'addestramento.
   * **Contributo**: Questa diversità riduce il rischio che il modello sviluppi bias basati su un set limitato di dati o su una singola cultura o lingua dominante. Ad esempio, se un dataset contiene testi provenienti solo da una lingua o cultura specifica, il modello può sviluppare un bias verso quella lingua o cultura. Includendo dati da molteplici culture e lingue, il modello impara a essere più neutro e a generalizzare meglio.
2. **Bilanciamento dei Dati**:
   * **Descrizione**: È cruciale che i dataset cross-domain siano bilanciati per evitare la sovrarappresentazione di qualsiasi dominio, lingua, o cultura. Un dataset sbilanciato può portare a un modello che è più accurato o preferenziale verso un contesto rispetto ad altri.
   * **Contributo**: Bilanciare i dati aiuta a mitigare il bias che potrebbe emergere a causa della predominanza di una particolare cultura, lingua o stile di scrittura. Questo bilanciamento può essere fatto tramite tecniche come il campionamento ponderato, l'aggiunta di dati sottorappresentati o la rimozione di dati sovrarappresentati.
3. **Valutazione su Diversi Contesti**:
   * **Descrizione**: Utilizzando dataset cross-domain per la valutazione, i ricercatori possono testare i modelli in una varietà di contesti per capire come performano su dati che non appartengono a un singolo dominio.
   * **Contributo**: Questo tipo di valutazione aiuta a identificare pregiudizi specifici del modello che potrebbero non emergere durante l'addestramento su dati più limitati. Ad esempio, un modello potrebbe funzionare bene su testi in inglese ma meno bene su testi in lingue minoritarie o su stili di scrittura non occidentali. Un dataset cross-domain permette di valutare queste differenze di performance e di apportare correzioni per migliorare l'equità del modello.
4. **Miglioramento della Capacità di Generalizzazione del Modello**:
   * **Descrizione**: Un dataset cross-domain aiuta a migliorare la capacità del modello di generalizzare attraverso diversi domini e contesti. Invece di imparare da un singolo tipo di dati, il modello è esposto a vari tipi di contenuti che rappresentano diverse reali situazioni d'uso.
   * **Contributo**: Questa esposizione a una vasta gamma di dati aiuta il modello a sviluppare una comprensione più ampia e robusta del linguaggio e dei contesti culturali, riducendo la probabilità di bias quando è esposto a nuovi dati che non erano presenti nel dataset di addestramento.

### Utilità dei Dataset Cross-Domain per la Ricerca sul Bias

I dataset cross-domain sono estremamente utili per migliorare l'attuale stato dell'arte nella ricerca sulla mitigazione del bias per diversi motivi:

* **Creazione di Modelli Inclusivi**: Contribuiscono a creare modelli che sono più inclusivi e rappresentativi di diverse culture, lingue e contesti sociali. Questo è particolarmente importante per applicazioni globali di AI, come assistenti virtuali, traduzione automatica, e sistemi di raccomandazione.
* **Identificazione di Nuove Forme di Bias**: L'uso di dati diversificati permette di identificare forme di bias che potrebbero non essere visibili con dataset più limitati o omogenei. Ad esempio, il bias culturale che emerge solo quando il modello è esposto a testi provenienti da culture diverse.
* **Sviluppo di Metriche di Valutazione Più Raffinate**: I dataset cross-domain stimolano la creazione di nuove metriche di valutazione che sono più sensibili ai bias contestuali e culturali. Queste metriche possono valutare come il modello si comporta in una varietà di contesti, aiutando a identificare aree di miglioramento.
* **Promozione dell'Equità nella Tecnologia AI**: L'uso di dataset cross-domain è una delle pratiche migliori per promuovere l'equità nei modelli di AI. Aiuta a garantire che i modelli non siano solo accurati in un contesto limitato ma anche equi e utili per una vasta gamma di utenti globali.

### Sfide Aperte e Futuri Sviluppi

Nonostante i vantaggi, ci sono ancora alcune sfide associate all'uso di dataset cross-domain:

* **Raccolta e Curazione di Dati Diversificati**: La raccolta di dati che rappresentano adeguatamente tutte le culture, lingue, e domini è complessa e richiede uno sforzo significativo per evitare sovrarappresentazioni e bias non intenzionali.
* **Costi e Risorse**: La creazione e gestione di dataset cross-domain richiede risorse significative, inclusi tempo, denaro e potenza computazionale, specialmente quando si tratta di dati multilingua o provenienti da domini meno documentati.
* **Mantenimento della Qualità dei Dati**: Assicurare che i dati raccolti siano di alta qualità, accurati e non contengano errori o distorsioni richiede uno sforzo continuo di curazione e pulizia dei dati.

### ****3. Trasparenza e Spiegabilità****

**Descrizione:** Trasparenza e spiegabilità sono fondamentali per comprendere come un modello linguistico arriva a determinate conclusioni o output. La trasparenza riguarda la capacità di tracciare e comprendere le decisioni del modello, mentre la spiegabilità si riferisce alla capacità di comunicare queste decisioni in modo comprensibile agli utenti.

**Cosa comporta il miglioramento della trasparenza e spiegabilità:**

* **Strumenti di Spiegabilità:** Sviluppare strumenti che consentano agli utenti di capire meglio come un modello linguistico ha prodotto un certo output. Questi strumenti potrebbero includere tecniche come LIME (Local Interpretable Model-agnostic Explanations) o SHAP (SHapley Additive exPlanations), che spiegano l'importanza di diverse feature o input per un determinato output.
* **Tracciamento delle Origini degli Output:** Significa sviluppare capacità che permettano di risalire alle fonti o ai processi interni che hanno influenzato un output pregiudicato. Questo può aiutare a identificare specifiche parti del dataset o fasi del processo di addestramento che hanno contribuito al bias.
* **Feedback Utente e Interventi:** La spiegabilità non riguarda solo la comprensione interna degli sviluppatori, ma anche la capacità degli utenti finali di fidarsi e comprendere le decisioni del modello. Questo potrebbe includere interfacce utente che mostrano come il modello ha preso una decisione o che consentono agli utenti di intervenire e correggere potenziali pregiudizi.

### Problemi Aperti nel Bias delle AI

1. **Bias Invisibili o Non Riconosciuti**:
   * **Descrizione**: Molti bias nei modelli di AI sono sottili e difficili da rilevare con le tecniche attuali. Mentre alcuni bias evidenti (come quelli di genere o razza) sono stati ampiamente studiati, ci sono bias più sottili che emergono in contesti specifici e non sono facilmente identificabili dalle metriche tradizionali di valutazione del bias.
   * **Problema Aperto**: Sviluppare metodi e strumenti migliori per identificare bias che non sono immediatamente evidenti e che possono influire sulle decisioni del modello in modi più subdoli. Questi includono bias culturali, bias socioeconomici, o bias basati su preferenze linguistiche che possono non essere uniformemente distribuiti nei dati di addestramento.
2. **Spiegabilità nei Modelli Complessi e Multi-Modali**:
   * **Descrizione**: Modelli linguistici avanzati, come quelli basati su reti neurali profonde (ad esempio, Transformer come GPT, BERT) o modelli multi-modali (che combinano testo e immagini), sono estremamente complessi e spesso operano come "scatole nere", rendendo difficile spiegare le loro decisioni.
   * **Problema Aperto**: Creare strumenti di spiegabilità che siano efficaci anche per modelli complessi e multi-modali. Attualmente, strumenti come LIME o SHAP forniscono spiegazioni limitate che possono non essere sufficienti per comprendere a fondo come interazioni complesse tra feature portano a determinati output.
3. **Spiegazioni Utilizzabili per Mitigare il Bias**:
   * **Descrizione**: Anche quando i bias vengono identificati, non è sempre chiaro come utilizzare queste informazioni per migliorare il modello o mitigare il bias senza compromettere le prestazioni.
   * **Problema Aperto**: Sviluppare metodi che non solo spieghino il bias ma forniscano anche un percorso pratico e chiaro per la sua mitigazione. Ciò potrebbe includere tecniche di ri-addestramento mirato, aggiustamento dei pesi del modello, o la rimozione di dati problematici, mantenendo però l'integrità e l'accuratezza del modello.
4. **Equità Interpretativa**:
   * **Descrizione**: La spiegabilità può variare significativamente tra diversi gruppi di utenti. Quello che è comprensibile per un tecnico potrebbe non esserlo per un utente comune o per un individuo con un background culturale diverso.
   * **Problema Aperto**: Come garantire che le spiegazioni siano utili, comprensibili e applicabili a una vasta gamma di utenti con diverse competenze tecniche e background culturali. Questo include la creazione di interfacce utente e modelli di comunicazione che siano chiari per tutti gli utenti, indipendentemente dal loro livello di competenza.
5. **Rischio di Spiegazioni Ingannevoli o Superficiali**:
   * **Descrizione**: Esiste il rischio che le spiegazioni generate automaticamente siano ingannevoli o troppo semplificate, dando agli utenti un falso senso di sicurezza sul comportamento del modello.
   * **Problema Aperto**: Assicurarsi che le spiegazioni siano accurate e complete. Evitare spiegazioni che mascherano la complessità del modello o che omettono dettagli cruciali che potrebbero rivelare pregiudizi o limiti del modello.
6. **Bias Indotti dalla Spiegabilità**:
   * **Descrizione**: Ironia della sorte, gli stessi strumenti di spiegabilità possono introdurre bias se enfatizzano in modo eccessivo alcune feature o decisioni del modello rispetto ad altre.
   * **Problema Aperto**: Sviluppare strumenti di spiegabilità che siano imparziali e non introducano ulteriori bias nelle loro spiegazioni. Bilanciare la semplicità e l'accessibilità delle spiegazioni con la necessità di rappresentare accuratamente la complessità delle decisioni del modello.
7. **Tracciamento dell'Origine del Bias nei Pipeline Complessi**:
   * **Descrizione**: Nei sistemi di AI complessi, le decisioni finali del modello possono essere influenzate da molteplici fasi del pipeline di elaborazione dei dati (ad esempio, raccolta dei dati, preprocessamento, addestramento, ecc.). Questo rende difficile tracciare l'origine esatta di un bias.
   * **Problema Aperto**: Sviluppare strumenti e metodologie per tracciare come i bias emergono e si propagano attraverso diverse fasi del processo di sviluppo del modello, facilitando l'identificazione e la correzione del bias alla sua fonte.
8. **Trasparenza Versus Privacy e Sicurezza**:
   * **Descrizione**: Rendere i modelli più trasparenti può entrare in conflitto con la necessità di proteggere la privacy degli utenti e la sicurezza dei dati. Inoltre, la piena trasparenza del modello può rendere il sistema vulnerabile agli attacchi o all'uso improprio.
   * **Problema Aperto**: Bilanciare la necessità di trasparenza e spiegabilità con le considerazioni di privacy, sicurezza e protezione del modello. Trovare metodi per fornire spiegazioni sufficientemente dettagliate per la fiducia e la comprensione dell'utente senza esporre dati sensibili o vulnerabilità del modello.