Seconda Parte Progetto IA

LLM-Eval

Data	14/01/2025
Destinatario	Prof. V. Deufemia, Dott. G. Cimino
Presentato da	Arcangeli Giovanni, Ciancio Vittorio, Di Maio Marco

Sommario

1. Introduzione

1.1 Contesto del progetto

L'intelligenza artificiale (IA) sta rapidamente trasformando il modo in cui interagiamo con la tecnologia. I chatbot, in particolare, stanno diventando sempre più diffusi in diversi settori, come il servizio clienti, l'assistenza sanitaria e l'e-commerce. Questi agenti conversazionali, basati su modelli di linguaggio avanzati, sono in grado di simulare conversazioni umane e fornire informazioni, assistenza e supporto agli utenti.

Tuttavia, la crescente adozione dei chatbot solleva importanti questioni sulla loro efficacia e sulla qualità dell'interazione uomo-macchina. È fondamentale disporre di metodi affidabili per valutare le prestazioni dei chatbot e garantire che siano in grado di soddisfare le esigenze degli utenti. La valutazione automatica dei dialoghi tra bot e umani rappresenta quindi una sfida cruciale per lo sviluppo e il miglioramento di queste tecnologie.

1.2 Obiettivi

Questo progetto si propone di affrontare la sfida della valutazione automatica dei dialoghi uomo-macchina, con i seguenti obiettivi specifici:

- **Comprendere la metrica LLM-EVAL** proposta nell'articolo "LLM-EVAL: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models".
- **Implementare il framework LLM-EVAL** per valutare dialoghi open-domain utilizzando modelli linguistici di grandi dimensioni (LLM).
- **Analizzare l'efficacia di LLM-EVAL** nel misurare la qualità di risposte generate rispetto a giudizi umani utilizzando dataset di benchmark.
- **Valutare l'accuratezza di diversi modelli di IA** nell'assegnare punteggi di qualità ai dialoghi tra bot e umani. Nello specifico, verranno confrontati quattro modelli: GPT4o-mini, GPT4o, Claude 3.5 e Claude 3.
- **Confrontare le prestazioni dei modelli su un dataset di riferimento**, analizzando le differenze in termini di metriche.
- **Valutare l'impatto di diversi dataset** sulle prestazioni di un modello di IA, utilizzando Claude 3 per analizzare quattro dataset: PC, TC, DSTC9 e FED.
- **Investigare l'efficacia di un prompt specifico**, derivato dalla letteratura scientifica, per la valutazione della qualità dei dialoghi.

1.3 Struttura della relazione

La relazione è strutturata come segue:

- **Capitolo 2:** Presenta un primo approccio al problema con una panoramica generale sul background teorico necessario.
- **Capitolo 3:** Descrive la prima fase del progetto, focalizzata sulla valutazione di quattro modelli di IA su un singolo dataset.
- **Capitolo 4:** Presenta la seconda fase del progetto, in cui viene valutato l'impatto di diversi dataset sulle prestazioni di Claude 3.
- **Capitolo 5:** Riassume i risultati principali del progetto.

Nei prossimi capitoli, verranno presentati in dettaglio i dataset utilizzati, la metodologia di valutazione, i risultati ottenuti e le relative discussioni.

2. Background teorico

2.1 Introduzione

I metodi di valutazione dei sistemi di dialogo automatici tradizionali, come BLEU e ROUGE, si sono dimostrati inadeguati per catturare le sfumature delle conversazioni in linguaggio naturale. Di conseguenza, sono state sviluppate diverse metriche avanzate, ma molte di esse richiedono dati di annotazione, riferimenti umani o prompt multipli, il che può essere costoso, dispendioso in termini di tempo o soggetto a errori.

2.2 LLM-EVAL: Un metodo unificato per la valutazione multidimensionale

LLM-EVAL è un metodo di valutazione automatica unificato e multidimensionale per conversazioni open-domain con modelli linguistici di grandi dimensioni (LLM). Questo metodo affronta le limitazioni dei metodi di valutazione esistenti, offrendo un approccio efficiente e accurato che copre diverse dimensioni della qualità del dialogo, come contenuto, grammatica, pertinenza e appropriatezza, senza richiedere riferimenti umani o prompt multipli.

Schema di valutazione unificato: LLM-EVAL utilizza uno schema di valutazione in linguaggio naturale per definire il compito di valutazione e i criteri desiderati. Lo schema è progettato per coprire molteplici dimensioni della valutazione, come contenuto, grammatica, pertinenza e appropriatezza. Lo schema viene fornito come istruzione di formato, specificando la struttura e l'intervallo dei punteggi per ciascuna dimensione.

Prompt singolo: LLM-EVAL impiega un singolo prompt per la valutazione, che include il contesto del dialogo, il riferimento (se disponibile) e la risposta generata. Il prompt viene concatenato con lo schema di valutazione unificato e viene fornito a un modello linguistico di grandi dimensioni, che restituisce i punteggi per ciascuna dimensione in base allo schema definito.

Efficienza del metodo: LLM-EVAL semplifica il processo di valutazione, in quanto richiede una sola chiamata al modello linguistico per ottenere punteggi multidimensionali. Questo lo rende un metodo efficiente e scalabile per la valutazione di sistemi di dialogo.

Prompt progettati in linguaggio naturale: LLM-EVAL utilizza prompt progettati in linguaggio naturale, rendendo le richieste intuitive e facilmente comprensibili per il modello. Questo approccio riduce l'overhead computazionale rispetto ai metodi che richiedono più prompt o configurazioni complesse.

Configurazioni flessibili: LLM-EVAL supporta configurazioni di punteggio flessibili, come intervalli da 0-5 o da 0-100, per adattarsi a diverse esigenze di valutazione e per riflettere meglio i contesti di giudizio umano.

2.3 Esperimenti e risultati

LLM-EVAL è stato testato su una varietà di dataset di benchmark, come DSTC10, TopicalChat, PersonaChat e DailyDialog, rappresentativi di contesti diversi, dalla personalizzazione delle risposte alla qualità generale delle conversazioni. Gli esperimenti hanno dimostrato che LLM-EVAL supera costantemente la maggior parte dei metodi di valutazione di base.

LLM-EVAL ha ottenuto correlazioni elevate rispetto ai giudizi umani, con correlazioni Spearman ρ che superano il 50% in alcune dimensioni, dimostrando la sua affidabilità nel catturare le sfumature dei dialoghi.

Particolarmente interessante è l'efficacia del metodo nella modalità senza riferimenti umani (reference-free). Questo lo rende ideale per scenari in cui non sono disponibili risposte di riferimento.

2.4 Analisi

L'analisi ha dimostrato che i modelli ottimizzati per applicazioni di chat, come Claude e ChatGPT, forniscono risultati di valutazione più accurati rispetto a modelli generici come GPT-3.5. Questo suggerisce che la scelta del modello è cruciale per ottenere valutazioni affidabili.

La decodifica greedy, che seleziona il token con la probabilità più alta a ogni passo, genera risposte più coerenti rispetto al nucleus sampling, che introduce casualità. Ad esempio, nel dataset DailyDialog, greedy decoding ha prodotto punteggi di coerenza più alti del 10% rispetto al nucleus sampling.

2.5 Limiti e considerazioni etiche

LLM-EVAL presenta alcune limitazioni, tra cui la dipendenza dalle prestazioni del modello linguistico sottostante, che può essere influenzato da bias o generare output inaspettati. Inoltre, la scelta del LLM influenza significativamente i risultati della valutazione. È importante considerare attentamente le possibili implicazioni etiche legate all'uso di LLM, come la perpetuazione di bias presenti nei dati di addestramento.

La qualità dei risultati di LLM-EVAL dipende fortemente dalla progettazione dei prompt e dallo schema di valutazione. Creare prompt efficaci richiede competenze specifiche, il che potrebbe limitare l'accessibilità del metodo per utenti non esperti.

Un aspetto critico è la possibilità che LLM-EVAL perpetui bias presenti nei dati di addestramento del modello. Tali bias possono influenzare i punteggi e portare a valutazioni non equilibrate, con potenziali implicazioni etiche.

2.6 Conclusioni

LLM-EVAL offre una soluzione versatile e robusta per la valutazione dei sistemi di conversazione open-domain, semplificando il processo di valutazione e fornendo prestazioni coerenti in diversi scenari. Il metodo proposto contribuisce alla ricerca sui sistemi di dialogo, fornendo uno strumento efficiente e accurato per la valutazione della qualità delle conversazioni generate dai modelli linguistici di grandi dimensioni.

Un'area di miglioramento futura per LLM-EVAL è l'inclusione di dimensioni di valutazione aggiuntive, come empatia, creatività e umorismo. Queste dimensioni richiedono interpretazioni più soggettive e una maggiore sensibilità alle sfumature dei dialoghi.

Un altro sviluppo promettente è l'integrazione di LLM-EVAL con tecniche di apprendimento per rinforzo basate sui feedback dei modelli stessi. Questo approccio potrebbe migliorare ulteriormente la qualità e l'efficacia delle valutazioni.

3. Fase 1: Valutazione dei modelli di IA

3.1 Descrizione del dataset

In questa fase del progetto, è stato utilizzato il dataset "convai2_data.json" per valutare le prestazioni dei modelli di IA. Questo dataset contiene un insieme di dialoghi, contrassegnati da un “dialog_id”, tra due profili (participant1 e participant2), ogni messaggio contiene un proprio “id”.

Sono presenti ulteriori parametri nel dataset, il più importante per la nostra analisi è "score_eval". Questo parametro rappresenta un punteggio di qualità assegnato ai dialoghi. L'obiettivo dei modelli di IA è quello di predire questo punteggio in modo accurato, replicandolo.

Oltre ai parametri utilizzati in questo studio, ce ne sono altri che risultano ugualmente interessanti.

Infatti, abbiamo “profile_match” che valuta l’attinenza dei messaggi inviati dal bot con la descrizione del profilo presente nel parametro “bot_profile”. Tale valore è 0 in caso di non attinenza e 1 in caso contrario.

Oltre a “bot_profile” è presente anche “user_profile” che come per l’altro, presenta delle voci in cui viene descritto il profilo dell’utente.

3.2 Presentazione dei modelli di IA

Sono stati valutati quattro modelli di IA, tutti basati su architetture di deep learning e addestrati su enormi quantità di dati testuali:

- **GPT4o-mini:** Una versione ridotta del modello GPT-4, ottimizzata per applicazioni con risorse computazionali limitate.
- **GPT4o:** Un modello linguistico di grandi dimensioni sviluppato da Google, noto per la sua capacità di generare testo fluente e coerente.
- **Claude-3-5-haiku-20241022:** Un modello linguistico avanzato sviluppato da Anthropic, progettato per essere più sicuro e meno suscettibile a generare output tossici o dannosi.
- **Claude-3-haiku-20240307:** La versione precedente di Claude 3.5, anch'essa sviluppata da Anthropic.

Nella scelta del modello linguistico per il nostro progetto, abbiamo optato per Claude versione Haiku, scartando le alternative Sonnet e Opus. Questa decisione è stata guidata da una serie di motivazioni legate alle specifiche caratteristiche di Haiku e agli obiettivi del nostro progetto.

Innanzitutto, Haiku è la variante di Claude progettata specificamente per l'analisi e la generazione di testo in linguaggio naturale. Considerando che il nostro obiettivo principale è la valutazione della qualità dei dialoghi, la scelta di Haiku si è rivelata naturale. Al contrario, Sonnet e Opus, pur essendo modelli linguistici potenti, sono focalizzati su altri ambiti, come la generazione di codice o la traduzione, e quindi meno adatti all'analisi del dialogo.

In secondo luogo, Haiku è stato addestrato su un vasto dataset di dialoghi e conversazioni, il che gli conferisce una maggiore sensibilità nel valutare la qualità e la coerenza di un dialogo. Questo addestramento specifico lo rende in grado di cogliere le sottigliezze e le dinamiche conversazionali che sono cruciali per una valutazione accurata, capacità che Sonnet e Opus potrebbero non possedere nella stessa misura.

Inoltre, la scelta di Haiku si allinea con le raccomandazioni dell'articolo su LLM-EVAL.

Infine, Haiku, essendo una versione più leggera di Claude rispetto a Sonnet e Opus, offre vantaggi in termini di efficienza computazionale. Questo aspetto è rilevante nel nostro progetto, che prevede l'analisi di un grande volume di dialoghi.

In conclusione, la scelta di Claude-Haiku è stata dettata da una combinazione di fattori: la sua specializzazione nell'elaborazione del linguaggio naturale, la sua capacità di valutare la qualità dei dialoghi, il suo allineamento con le raccomandazioni dell'articolo su LLM-EVAL e la sua efficienza computazionale. Riteniamo che Haiku sia il modello più adatto per raggiungere gli obiettivi del nostro progetto e per fornire una valutazione accurata e affidabile della qualità dei dialoghi.

*Per una questione di praticità, nell'intera documentazione i modelli **Claude-3-5-haiku-20241022** e **Claude-3-haiku-20240307** verranno nominati nelle forme abbreviate: **Claude 3.5** e **Claude 3**.

3.3 Metodologia di valutazione

Per valutare le prestazioni dei modelli linguistici di grandi dimensioni (LLM) nella valutazione della qualità dei dialoghi, sono state utilizzate diverse metriche di valutazione. Queste metriche sono state scelte per la loro capacità di catturare diversi aspetti della qualità del dialogo, come la coerenza, la fluenza, l'informatività e la pertinenza.

Le metriche utilizzate in questo progetto sono:

- **Accuratezza:** L'accuratezza misura la percentuale di risposte corrette fornite dal modello. Nel contesto della valutazione del dialogo, l'accuratezza indica la percentuale di volte in cui il modello ha assegnato il punteggio corretto al dialogo, ossia, il punteggio che corrisponde all'eval_score presente nel dataset.
- **Kappa di Cohen:** La Kappa di Cohen è una misura statistica che misura l'accordo tra due valutatori, tenendo conto dell'accordo casuale. Nel contesto della valutazione del dialogo, la Kappa di Cohen misura l'accordo tra il modello e un valutatore umano nell'assegnazione dei punteggi di qualità ai dialoghi. Un valore di Kappa pari a 1 indica un accordo perfetto, mentre un valore pari a 0 indica un accordo non migliore del caso.
- **Correlazione di Spearman:** La correlazione di Spearman è una misura statistica che misura la correlazione tra due variabili ordinali. Nel contesto della valutazione del dialogo, la correlazione di Spearman misura la correlazione tra i punteggi di qualità assegnati dal modello e quelli assegnati da un valutatore umano. Un valore di correlazione pari a 1 indica una correlazione perfetta, mentre un valore pari a 0 indica l'assenza di correlazione.
- **Correlazione di Pearson:** La correlazione di Pearson è una misura statistica che misura la correlazione lineare tra due variabili continue. Nel contesto della valutazione del dialogo, la correlazione di Pearson misura la correlazione lineare tra i punteggi di qualità assegnati dal modello e quelli assegnati da un valutatore umano. Un valore di correlazione pari a 1 indica una correlazione positiva perfetta, un valore pari a -1 indica una correlazione negativa perfetta e un valore pari a 0 indica l'assenza di correlazione.
- **Correlazione di Kendall-Tau:** La correlazione di Kendall-Tau è una misura statistica che misura la correlazione tra due variabili ordinali, basandosi sul numero di coppie concordanti e discordanti. Nel contesto della valutazione del dialogo, la correlazione di Kendall-Tau misura la correlazione tra i punteggi di qualità assegnati dal modello e quelli assegnati da un valutatore umano, considerando il numero di coppie di dialoghi per cui il modello e il valutatore umano sono d'accordo o in disaccordo sull'ordine dei punteggi.

Queste metriche forniscono una valutazione completa delle prestazioni dei modelli linguistici nella valutazione della qualità dei dialoghi, tenendo conto di diversi aspetti come l'accuratezza, l'accordo con i valutatori umani e la correlazione tra i punteggi assegnati.

L'implementazione degli esperimenti è stata effettuata utilizzando il linguaggio di programmazione Python. I modelli di IA sono stati interrogati tramite API, e i risultati ottenuti sono stati confrontati con i valori di "score_eval" presenti nel dataset.

3.4 Prompt utilizzato

Come accennato, per tutti i modelli è stato utilizzato lo stesso prompt presente nell'articolo citato all'interno del secondo capitolo.

Il prompt completo viene concatenato e fornito a un modello linguistico di grandi dimensioni (LLM), che restituisce uno score che valuti la bontà del dialogo.

Questo è il prompt utilizzato:

```
Score the following dialogue generated on a continuous scale from 1 to 5.
Dialogue: {dialogue}
```

Esempio:

```
Score the following dialogue generated on a continuous scale from 1 to 5.
Dialogue:
Participant1: Hi, how are you doing today?
Participant2: I'm good, thanks! How about you?
Participant1: I'm doing well, just enjoying the nice weather.
Participant2: That sounds lovely. Have you been outside much?
```

Spiegazione:

Contesto del Dialogo:

Il dialogo è tra due partecipanti, Participant1 e Participant2, che scambiano messaggi.

Il contenuto è casuale e rappresenta una conversazione.

Struttura del Prompt:

La frase iniziale specifica il compito: "Score the following dialogue on a continuous scale from 1 to 5". Questo indica che il modello deve assegnare un punteggio da 1 a 5 per valutare il dialogo.

Il dialogo è presentato in un formato leggibile e semplice, con i turni di conversazione chiaramente attribuiti a ciascun partecipante.

Risultato Atteso:

Il modello dovrebbe restituire un punteggio numerico compreso tra 1 e 5, denominato eval_score

Esempio di Output:

eval_score: 4.5

3.5 Descrizione codice

3.5.1 Codice Claude

Di seguito viene descritto il funzionamento del codice implementato per il modello Claude:

1. Setup e configurazione

- **Import delle librerie**:

    - os, json, re: Utilizzate per gestire file, lavorare con JSON, e manipolare stringhe con espressioni regolari.
    - dotenv: Per caricare la chiave API da un file .env.
    - anthropic: Libreria client per interagire con il modello di linguaggio Claude.
    - tqdm: Fornisce barre di progresso per il monitoraggio dei processi.
- **Caricamento della chiave API**: La chiave API viene caricata dal file .env, assicurando che le credenziali non siano esposte nel codice.
- **Inizializzazione del client Anthropic**

2. Funzione estrai_json

Questa funzione ripulisce e normalizza l'output del modello di linguaggio per estrarre un JSON valido.

Passaggi principali:

1. **Ricerca del JSON nell'output**: Utilizza un'espressione regolare per trovare un blocco JSON nell'output del modello.
2. **Pre-processamento**: Corregge eventuali valori non validi, ad esempio sostituendo high con "high".
3. **Parsing e gestione errori**: Tenta di convertire il testo JSON in un dizionario Python.
4. **Normalizzazione del punteggio**:
    - Cerca una delle possibili chiavi che rappresentano il punteggio:
    - Arrotonda il valore trovato e lo assegna al campo eval_score.
5. **Costruzione del risultato**:
    - Restituisce un dizionario contenente il punteggio normalizzato e l'ID del dialogo:

3. Funzione valuta_dialogo

Valuta un singolo dialogo utilizzando il modello Claude.

Passaggi principali:

1. **Preparazione del testo del dialogo**: Combina tutte le battute in un unico blocco di testo:
2. **Definizione del prompt**: Crea un messaggio strutturato per il modello Claude, chiedendogli di valutare il dialogo:
3. **Chiamata al modello**: Invia il prompt al modello Claude e ottiene una risposta.
4. **Estrazione del risultato**: Chiama estrai_json per ripulire e interpretare la risposta del modello.
5. **Gestione errori**: Se il risultato non è valido o si verifica un'eccezione, registra l'errore.

4. Funzione valuta_intero_dataset

Valuta tutti i dialoghi nel dataset e salva i risultati incrementali in due file: uno per i risultati corretti e uno per gli errori.

Passaggi principali:

1. **Caricamento dei risultati esistenti**: Controlla se i file di output ed errori esistono, e li carica:
2. **Evita duplicati**: Traccia i dialoghi già processati:
3. **Processa i dialoghi**:
    - Per ogni dialogo non ancora processato, chiama valuta_dialogo.
    - Salva i risultati incrementali in output_file o registra gli errori in error_file.
4. **Salvataggio incrementale**: Aggiorna i file ad ogni dialogo processato, riducendo il rischio di perdita di dati.

3.5.2 Codice GPT4o

Le differenze sostanziali dal codice di Claude sono le seguenti:

Struttura della richiesta al modello

Per quanto riguarda Claude, rispetto a GPT4o, permette di specificare istruzioni specifiche che definiscono il comportamento del modello.

Gestione delle risposte del modello

Claude si concentra su risposte in formato JSON e prevede un'accurata fase di normalizzazione per correggere eventuali errori o incongruenze, come valori non validi.

GPT4o, invece, è più flessibile nel trattare risposte non strutturate. Se non trova un JSON valido, tenta di estrarre manualmente il punteggio dalla risposta testuale utilizzando una serie di regole basate su parole chiave e numeri. Questo approccio consente di gestire risposte meno prevedibili, ma richiede meno rigore nella normalizzazione.

Flessibilità dell'estrazione del punteggio

Claude assume che la risposta del modello sarà sempre in formato JSON e segue un approccio rigoroso per verificare e correggere eventuali errori in questo formato. GPT4o, invece, è più adattabile a variazioni nell'output del modello grazie al suo estrattore manuale, che può funzionare anche quando il modello fornisce risposte in linguaggio naturale.

Gestione del dataset

Entrambi i modelli adottano un approccio incrementale per processare un dataset di dialoghi. Salvano i risultati man mano che vengono elaborati e gestiscono eventuali errori salvando i dialoghi problematici in un file separato

3.6 Risultati

I risultati ottenuti dai quattro modelli di AI nella valutazione del dataset convai2_data.json sono presentati nella seguente tabella:

Modello	Accuratezza (%)	Kappa di Cohen	Correlazione di Spearman	Correlazione di Pearson	Correlazione di Kendall-Tau
GPT4o-mini	21.17	0.06	**0.39**	**0.36**	**0.33**
GPT4o	23.57	0.04	0.16	0.15	0.13
Claude 3.5	**25.99**	**0.07**	0.28	0.28	0.26
Claude 3	20.32	0.04	0.19	0.19	0.17

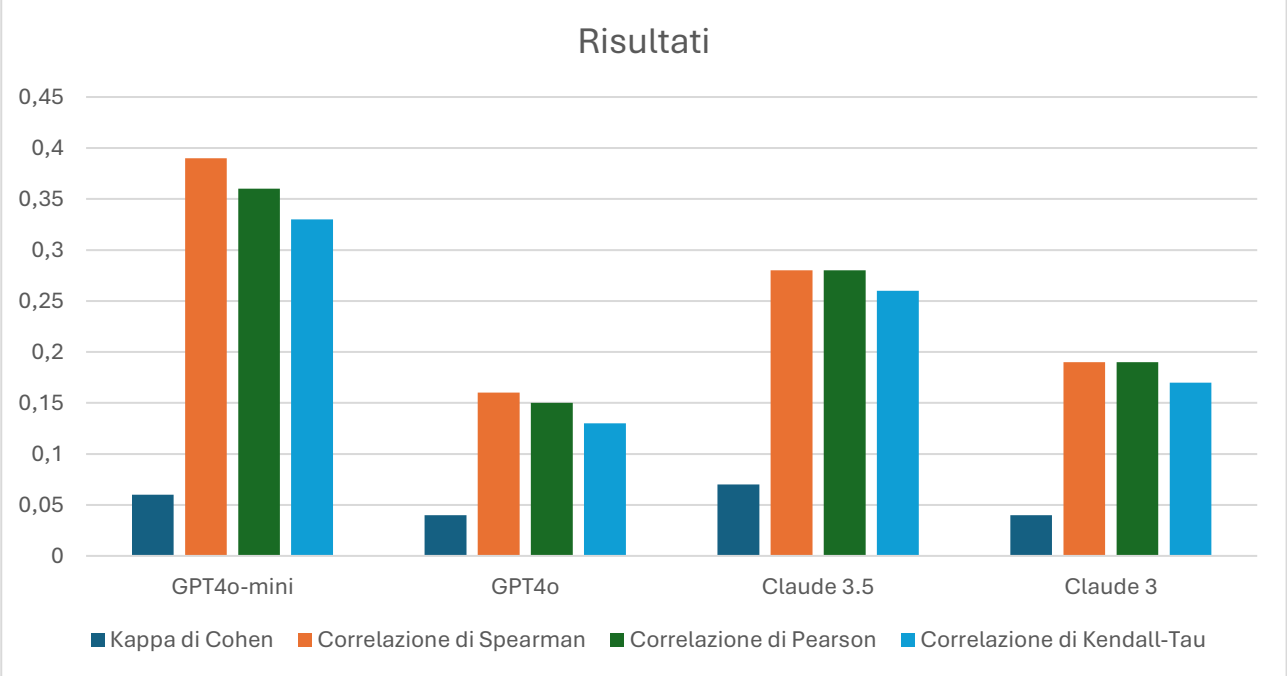
Come si può osservare, **Claude 3.5** ha ottenuto la migliore accuratezza (25.99%), seguito da **GPT4o** (23.57%). Tuttavia, l'accuratezza di tutti i modelli è relativamente bassa, indicando che nessuno di essi è stato in grado di predire il punteggio di qualità dei dialoghi in modo sufficientemente accurato.

Per quanto riguarda le metriche di correlazione, **GPT4o-mini** ha ottenuto i valori più alti per le tre metriche (Spearman, Pearson e Kendall-Tau). Questo risultato suggerisce che GPT4o-mini mostra la maggior coerenza con i giudizi umani nella classificazione generale dei dialoghi pur non essendo il modello più accurato.

Ritornando a **Claude 3.5**, oltre all'accuratezza maggiore, presenta il migliore accordo generale (Kappa di Cohen) e buoni risultati nelle altre metriche. Quindi, è il modello con il miglior equilibrio tra correlazione e accuratezza.

Per quanto riguarda **Claude 3** presenta risultati modesti in tutte le metriche, con un miglioramento evidente in Claude 3.5.

Infine, per quanto riguarda **GPT4o** ha le correlazioni più basse, indicando che i suoi punteggi sono meno allineati ai giudizi umani.



3.7 Discussione

I risultati ottenuti evidenziano la difficoltà di valutare automaticamente la qualità dei dialoghi tra bot e umani. Nessuno dei modelli testati ha raggiunto un'accuratezza soddisfacente, indicando che la semplice applicazione di un prompt, seppur derivato dalla letteratura scientifica, non è sufficiente per ottenere prestazioni elevate.

Le ragioni di queste difficoltà possono essere molteplici:

- **Complessità del linguaggio naturale:** Il linguaggio naturale è intrinsecamente ambiguo e ricco di sfumature, e i modelli di IA possono avere difficoltà a interpretarlo correttamente, soprattutto in contesti conversazionali.
- **Soggettività della valutazione:** La qualità di un dialogo è spesso soggettiva e dipende da diversi fattori, come le aspettative dell'utente, il contesto della conversazione e le preferenze personali.

- **Limiti dei modelli:** I modelli di IA, seppur addestrati su enormi quantità di dati, possono presentare limiti nella comprensione di concetti complessi o nella gestione di situazioni nuove o inaspettate.

Nonostante le difficoltà incontrate, il progetto ha permesso di ottenere alcuni risultati interessanti:

- **Claude 3.5:** Si è dimostrato il modello più accurato nella predizione del punteggio di qualità dei dialoghi.
- **GPT4o-mini:** Ha mostrato la migliore correlazione con i punteggi assegnati dai valutatori umani.

Questi risultati suggeriscono che la scelta del modello di IA può influenzare significativamente le prestazioni nella valutazione automatica dei dialoghi.

4. Fase 2: Valutazione dei dataset

4.1 Presentazione dei dataset

Nella seconda fase del progetto, abbiamo valutato le prestazioni di Claude 3 su quattro dataset diversi: PC, TC, DSTC9 e FED. Ciascun dataset presenta caratteristiche specifiche in termini di dimensione, tipologia di dialoghi, dominio di applicazione e struttura.

- **DSTC9 (Dialogue-Level):** Questo dataset contiene 2200 dialoghi. I dialoghi sono informativi e open-domain, e coprono una vasta gamma di argomenti. Le valutazioni fornite si riferiscono alla qualità generale del dialogo, senza distinzione tra turni individuali.
- **FED(Turn-Level e Dialogue-Level):** Il dataset FED contiene 125 dialoghi e 375 turn. Include dialoghi provenienti da diversi sistemi di dialogo e copre sia la valutazione a livello di dialogo che a livello di turno. Oltre alla valutazione complessiva, sono presenti valutazioni relative a parametri specifici, come la coerenza, la fluenza e la pertinenza delle risposte del sistema.
- **PC(Turn-Level):** Questo dataset contiene 300 turn e si concentra sulla valutazione della qualità delle risposte del sistema a livello di turno. Include informazioni sul contesto del dialogo, la risposta del sistema e valutazioni su parametri specifici.
- **TC(Turn-Level):** Il dataset TC è simile al dataset PC e contiene turn (360). Anche in questo caso, la valutazione si concentra sulla qualità delle risposte del sistema a livello di turno.

Le differenze tra questi dataset in termini di tipologia di dialoghi, dominio di applicazione e complessità ci permetteranno di valutare le prestazioni di Claude 3 in diversi contesti e di comprenderne meglio i punti di forza e di debolezza.

Un'ulteriore differenza sorge andando a visualizzare il numero di valutazioni effettuate per ogni caratteristica e/o overall. Difatti, FED include 5 valutazioni per caratteristica, mentre PC e TC ne hanno 3, e DSTC9 solo 1.

Di conseguenza, per effettuare un confronto per i valori di overall, è stata fatta una media di queste valutazioni, generandone una singola che potesse essere confrontabile con il valore generato dal nostro modello.

4.2 Metodologia di valutazione

In questa fase, abbiamo utilizzato Claude 3 per valutare le prestazioni sui diversi dataset.

Per valutare i dataset, abbiamo utilizzato due tipi di prompt a seconda della tipologia del dataset:

Per i dataset di tipo **turn-level** abbiamo usato:

```
Score the following dialogue response generated on a continuous scale from 1 to 5.

Context: {context}

Dialogue response: {response}
```

Per i dataset di tipo **dialogue-level**, invece:

```
Score the following dialogue generated on a continuous scale from 1 to 5.

Dialogue: {dialog}
```

Le metriche utilizzate per valutare le prestazioni di Claude 3 sono le stesse della Fase 1, ossia **Kappa di Cohen, Correlazione di Spearman, Correlazione di Pearson** e **Correlazione di Kendall-Tau.**

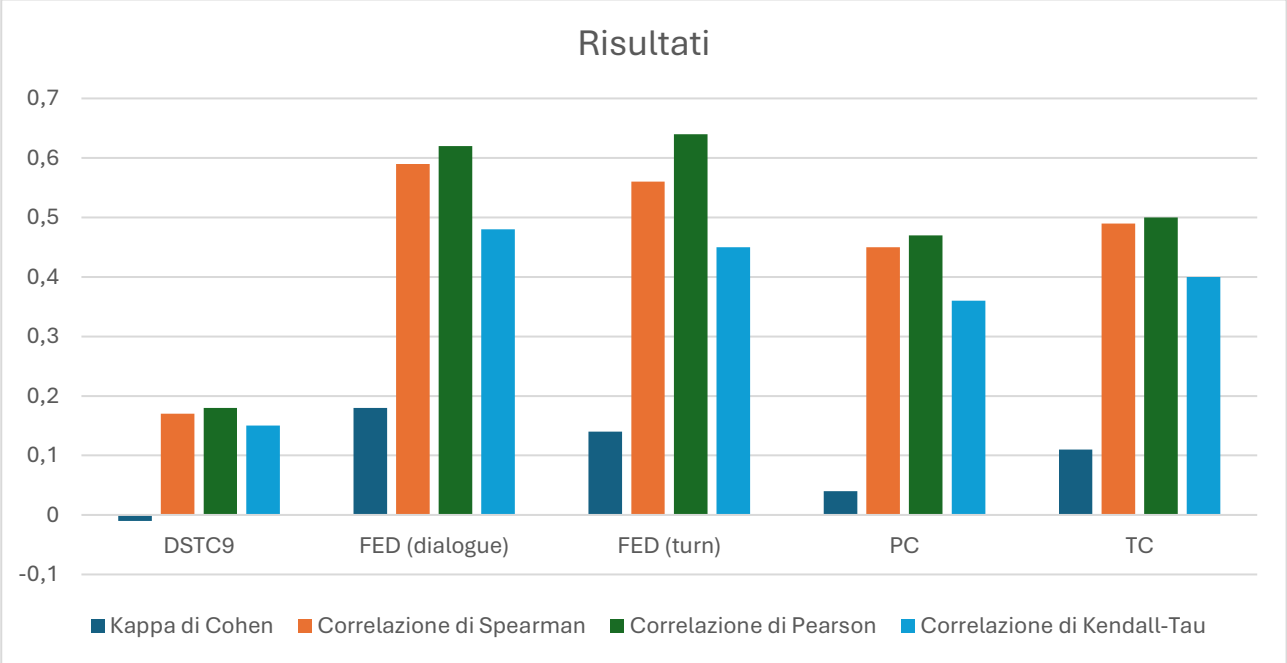
4.3 Risultati

I risultati ottenuti da Claude 3 sui diversi dataset sono riassunti nella seguente tabella:

Dataset	Accuratezza (%)	Kappa di Cohen	Correlazione di Spearman	Correlazione di Pearson	Correlazione di Kendall-Tau
DSTC9	4.23	-0.01	0.17	0.18	0.15
FED (dialogue)	**39.20**	**0.18**	**0.59**	0.62	**0.48**
FED (turn)	34.13	0.14	0.56	**0.64**	0.45
PC	25.67	0.04	0.45	0.47	0.36
TC	30.56	0.11	0.49	0.50	0.40

Come si può osservare, Claude 3 ha ottenuto un'accuratezza molto bassa sul dataset DSTC9 (4.23%), mentre le prestazioni sono significativamente migliori sugli altri dataset, con accuratezze intorno al 30%.

Le metriche di correlazione mostrano un andamento simile, con valori bassi per DSTC9 e valori più alti per gli altri dataset. In particolare, FED (dialogue) ha ottenuto i valori più alti per tutte le metriche tranne la Correlazione di Pearson, che per 0,02, risulta maggiore in FED (turn).



4.4 Discussione

Le basse prestazioni su DSTC9 potrebbero essere attribuite alla mancanza di annotazioni dettagliate e al fatto che i dialoghi open-domain includono argomenti eterogenei e spesso complessi, che richiedono una comprensione contestuale approfondita.

I risultati più elevati sui dataset FED, PC e TC indicano che Claude 3 è più efficace nella valutazione di dialoghi strutturati e task-oriented. Questo potrebbe essere dovuto al fatto che tali dataset presentano un contesto più definito e obiettivi specifici, facilitando la valutazione del modello.

Rispetto alla Fase 1, dove i modelli sono stati testati su un singolo dataset, questa analisi evidenzia come la scelta del dataset influenzi significativamente le prestazioni del modello. Claude 3 sembra trarre vantaggio da annotazioni dettagliate e contesti più strutturati, come evidenziato dai risultati sui dataset FED e TC.

5. Conclusioni

In questo progetto, abbiamo affrontato la sfida della valutazione automatica dei dialoghi tra bot e umani, con l'obiettivo di valutare l'accuratezza di diversi modelli di IA e l'impatto di diversi dataset sulle prestazioni.

5.1 Riepilogo dei risultati principali

Nella prima fase del progetto, abbiamo confrontato quattro modelli di IA (GPT4o-mini, GPT4o, Claude 3.5 e Claude 3) su un dataset di riferimento. I risultati hanno mostrato che Claude 3.5 ha ottenuto la migliore accuratezza, mentre GPT4o-mini ha mostrato la correlazione più alta con i punteggi umani. Nessun modello ha raggiunto un'accuratezza elevata, evidenziando la complessità del task di valutazione.

Nella seconda fase, abbiamo valutato le prestazioni di Claude 3 su quattro dataset diversi (PC, TC, DSTC9 e FED). I risultati hanno mostrato che Claude 3 è più efficace nella valutazione di dialoghi task-oriented e con annotazioni dettagliate, come quelli presenti nei dataset FED, PC e TC. Al contrario, le prestazioni su DSTC9, un dataset open-domain con annotazioni a livello di dialogo, sono state scarse.

5.2 Punti di forza dei modelli

- **GPT4o-mini:** Sebbene non sia stato il modello più accurato nella prima fase, GPT4o-mini ha mostrato la migliore correlazione con i punteggi umani, suggerendo una buona capacità di comprendere la qualità del dialogo. La sua versione ridotta lo rende inoltre più efficiente in termini di risorse computazionali.
- **GPT4o:** Questo modello ha ottenuto un'accuratezza leggermente inferiore a Claude 3.5, ma si distingue per la sua capacità di generare testo fluente e coerente, che potrebbe essere utile in altri task di elaborazione del linguaggio naturale.
- **Claude 3.5:** Si è dimostrato il modello più accurato nella prima fase, indicando una buona capacità di valutazione della qualità del dialogo.
- **Claude 3:** Pur essendo la versione precedente di Claude 3.5, ha ottenuto risultati comparabili agli altri modelli, dimostrando la solidità dell'architettura di base.

5.3 Modello migliore

Sulla base dei risultati ottenuti, **Claude 3.5** sembra essere il modello più promettente per la valutazione automatica dei dialoghi. La sua accuratezza superiore nella prima fase e le buone prestazioni sui dataset FED, PC e TC suggeriscono una buona capacità di adattamento a diversi tipi di dialoghi e di annotazioni.

5.4 Panoramica dei risultati della seconda fase

La seconda fase del progetto ha evidenziato l'importanza del dataset nella valutazione dei modelli di IA. Le prestazioni di Claude 3 sono variate significativamente a seconda del dataset utilizzato, dimostrando che la scelta del dataset è cruciale per ottenere risultati affidabili. In generale, Claude 3 ha ottenuto risultati migliori su dataset con dialoghi task-oriented e annotazioni dettagliate.

5.5 Considerazioni finali

Questo progetto ha fornito un contributo alla ricerca sulla valutazione automatica dei dialoghi, evidenziando le sfide e le opportunità di questo campo. I risultati ottenuti possono essere utili per guidare la scelta dei modelli e dei dataset nella valutazione dei chatbot e per lo sviluppo di nuove tecniche di valutazione.

Per quanto riguarda i problemi che sono emersi durante il progetto, si potrebbe dire che l'utilizzo di prompt generici e quindi non specifici per ciascun modello, abbiano influenzato negativamente i risultati.
Discorso analogo può essere fatto per il non utilizzo dei parametri del dataset "convai2_data" della prima fase di progetto, in quanto, alcuni parametri come "profile_match", "user_profile" e "bot_profile", avrebbero sicuramente potuto aiutare il modello in una valutazione più precisa.