Università degli studi di Salerno
Corso di Intelligenza Artificiale

GPTSCORE-DIALEVAL: UN METODO AVANZATO PER LA VALUTAZIONE DEI DIALOGHI GENERATIVI

Project work

del corso di

Intelligenza Artificiale

Carmela Pia Senatore, Gennaro Capaldo
Professore Vincenzo Deufemia
Corso di laurea magistrale in Informatica
:

Anno Accademico 2024/2025

CAMPUS FISCIANO, 2025

CARMELA PIA SENATORE, GENNARO CAPALDO

PROJECT WORK

GPTScore-DialEval: Un Metodo Avanzato per la Valutazione dei Dialoghi Generativi

FISCIANO, 2025

INDICE

ABSTRACT

Questo studio presenta GPTScore-DialEval, un framework innovativo per la valutazione della qualità delle risposte generate dai modelli di dialogo basati su GPT. In un'era in cui i chatbot e gli assistenti virtuali stanno assumendo un ruolo sempre più centrale nelle interazioni uomo-macchina, il progetto mira a superare i limiti delle metriche tradizionali come BLEU e ROUGE, incapaci di catturare pienamente elementi come coerenza, creatività e interesse. La metrica GPTScore sfrutta le capacità emergenti dei modelli GPT per valutare risposte sia a livello di turno (Turn-Level) che di dialogo (Dialogue-Level), misurando l'allineamento con il contesto e le istruzioni ricevute. Gli esperimenti sono stati condotti su diversi dataset, tra cui ConvAI2, DSCT9, Fed Data, Pc-Usr e Tc-Usr, utilizzando modelli GPT-4o-mini e Davinci-002. I risultati evidenziano correlazioni moderate tra i punteggi GPTScore e le valutazioni umane, in particolare per il modello Davinci-002, che ha ottenuto valori di correlazione significativi su ConvAI2 e Fed (a livello di turno). Il framework, pur con alcune limitazioni legate alla variabilità dei risultati, rappresenta un passo significativo verso metodologie di valutazione più affidabili, scalabili e vicine al giudizio umano, aprendo nuove possibilità per l'ottimizzazione dei modelli di dialogo generativo.

1. INTRODUZIONE

Negli ultimi anni, i modelli di linguaggio basati su architetture come GPT (Generative Pre-trained Transformers) hanno rivoluzionato il campo dell'intelligenza artificiale, in particolare nell'ambito dei sistemi di dialogo. Grazie alle loro capacità emergenti, questi modelli possono generare risposte naturali, contestuali e fluide, rendendoli sempre più utilizzati in chatbot, assistenti virtuali e piattaforme di supporto. Tuttavia, la valutazione della qualità delle risposte generate rimane una sfida aperta. Le tradizionali metriche automatiche, come BLEU e ROUGE, non riescono a catturare completamente aspetti qualitativi come l'interesse, la coerenza o la creatività delle risposte. Queste metriche, originariamente sviluppate per la traduzione automatica o il riassunto testuale, si basano sul confronto tra risposte generate e risposte di riferimento predefinite. Per colmare questa lacuna, GPTScore si propone come un nuovo paradigma per la valutazione automatica, sfruttando gli stessi modelli GPT per giudicare la qualità del testo generato. Questo approccio promette di superare i limiti delle metriche tradizionali e di avvicinarsi maggiormente alle valutazioni umane. La metrica centrale è GPTScore, è presentata nell'articolo GPTScore: Evaluate as You desire,nasce per sfruttare le capacità emergenti dei modelli GPT e valutare la qualità delle risposte in base al loro interesse. Questo approccio consente una valutazione più automatizzata e scalabile rispetto alle metodologie tradizionali.

Il progetto **GPTScore-Dial_eval** da noi proposto mira a implementare e testare la metrica GPTScore in un contesto specifico: la valutazione dell'interesse delle risposte in conversazioni generative. Gli obiettivi principali sono:

1. Sviluppare un framework che utilizza GPTScore per valutare risposte di dialoghi tra utente e bot basandoci su due livelli di dialogo:
   - **Turn-Level**: valuta ogni turno del dialogo separatamente, concentrandosi sull'accuratezza, la coerenza, o la rilevanza di una risposta rispetto al turno immediatamente precedente.
   - **Dialogue-Level** : considera l'intero scambio di dialogo, valutando la coerenza complessiva, la progressione del discorso, o l'aderenza agli obiettivi di conversazione.
2. Esplorare le correlazioni tra i punteggi GPTScore e le valutazioni umane per determinare la validità e l'affidabilità di questa metrica.
3. Analizzare come diverse configurazioni della metodologia applicata, dataset utilizzati e configurazioni del modello influenzano i risultati , offrendo un quadro completo.

In questo lavoro, abbiamo sviluppato un framework per la valutazione delle conversazioni in lingua inglese, utilizzando due diverse strategie di prompt per analizzare le risposte. I risultati ottenuti sono stati successivamente validati attraverso un confronto con i giudizi umani.

2. STATO DELL'ARTE

Le metriche tradizionali sono state ampiamente utilizzate in contesti conversazionali, ma con risultati limitati:

- BLEU (Bilingual Evaluation Understudy): Utilizzato per misurare la sovrapposizione tra frasi generate e riferimenti, BLEU non cattura la creatività o l'interesse delle risposte.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Usato per riassunti, misura la sovrapposizione di n-grammi, ma non considera la coerenza semantica o il contesto della conversazione.
- METEOR: Valuta l'allineamento tra risposte, tenendo conto di sinonimi e stemming, ma è ancora vincolato ai riferimenti predefiniti.

Queste metriche presentano significative limitazioni quando applicate alla valutazione di sistemi di dialogo:

- Incapacità di catturare aspetti qualitativi: Non considerano caratteristiche come l'interesse, la coerenza pragmatica o la creatività delle risposte.
- Dipendenza da riferimenti statici: Il confronto con un riferimento fisso penalizza la diversità nelle risposte.

- Scarsa correlazione con giudizi umani: Le metriche tradizionali spesso non riflettono accuratamente il modo in cui gli esseri umani percepiscono la qualità delle risposte.

Di fronte a queste limitazioni, emerge la necessità di sviluppare nuove metriche capaci di valutare le risposte in modo più umano-centrico, flessibile e adattabile. GPTScore è stato proposto come una risposta a questa sfida, sfruttando i modelli di linguaggio stessi per valutare la qualità del testo generato, superando i limiti delle metriche tradizionali.

3. MATERIALI E METODI

Per l'identificazione della metrica e la costruzione del framework, è stato utilizzato un approccio basato sulla linguistica computazionale, sfruttando il linguaggio di programmazione Python e librerie presenti nel file *requirements.txt*

Il progetto è stato pensato per un'esecuzione simil menù, ma è possibile modificarlo per un'esecuzione fluida manuale.

3.1 Datasets.

3.1.1 ConvAI. Il primo dataset utilizzato , *convai2_data.json - dialogue-level*, è strutturato in formato JSON ed è progettato per rappresentare dialoghi tra due partecipanti, un umano e un chatbot, con metadati utili per la valutazione e il contesto delle conversazioni. La struttura del file è la seguente:

- **Identificatore del dialogo (dialog_id):** ogni dialogo è associato a un identificatore univoco, rappresentato come una stringa alfanumerica.
- **Dialogo (dialog)**: È una lista di scambi tra due partecipanti, identificati come participant1 e participant2. Ogni scambio è rappresentato come un oggetto con i seguenti campi:
  1. id: Identificativo numerico dello scambio, incrementale.
  2. sender: Identifica il partecipante che ha inviato il messaggio (participant1 o participant2).
  3. text: Testo del messaggio inviato.

Listing 1. Esempio di dialog

```
 {
   "id": 0,
   "sender": "participant2",
   "text": "Hows it going?"
}
```

- **Profilo del bot (bot_profile)** : rappresenta una lista di informazioni di contesto relative al bot partecipante alla conversazione. Ogni elemento della lista descrive

un aspetto della personalità o delle esperienze del bot.
- **Profilo dell'utente (user_profile)**: rappresenta una lista di informazioni personali relative all'utente che partecipa alla conversazione. Ogni elemento descrive un aspetto della personalità, degli interessi o delle esperienze dell'utente.
- **Punteggio di valutazione (eval_score)**: rappresenta quanto un utente apprezza una conversazione, su una scala da 1 a 5.

3.1.2 DSCT9. Il secondo dataset utilizzato , *dsct9 - dialogue level*, contiene una serie di dialoghi. Ogni dialogo è rappresentato da una sequenza di scambi di battute (utterances) tra partecipanti, mantenendo l'ordine cronologico dello scambio.

3.1.3 Fed Data. Contiene sia valutazioni a livello di turno che a livello di dialogo, effettuate da annotatori umani per fluency, coerenza e capacità informativa. La metrica complessiva considerata è la media di tutti i punteggi.

3.1.4 Pc-Usr e Tc-usr. Dataset a livello di turno contenenti contesti conversazionali associati a diverse risposte generate da modelli differenti. Ogni risposta viene valutata su diverse dimensioni, tra cui Comprensibilità, Naturalità, Mantenimento del Contesto ed Engagement.

3.2 Workflow. L'approccio metodologico si articola in tre fasi principali:

1. **Interazione con un chatbot e valutazione della conversazione**: Gli utenti possono interagire con un chatbot basato su un modello pre-addestrato, disponibile sulla piattaforma Hugging Face, per generare conversazioni. Al termine dell'interazione, viene richiesto di assegnare un punteggio (score) che rifletta il grado di interesse delle risposte fornite dal bot e della conversazione nel suo complesso. I dettagli del dialogo e i punteggi assegnati vengono salvati in un file JSON strutturato secondo un formato predefinito, facilitando così analisi successive sui dati raccolti.
2. **Implementazione della metrica GPTScore**: L'utente può specificare il percorso del dataset su cui applicare la valutazione (ad esempio, `convai2_data.json`), la `openai_key` necessaria per accedere al modello e il tipo di modello desiderato (tra `gpt-3.5-turbo`, `chatgpt-4`, `davinci-002` o `gpt-4o-mini`, quest'ultimo impostato come predefinito).
   Una volta avviata l'analisi, il sistema:
   - Effettua richieste al modello GPT selezionato.
   - Estrae le probabilità condizionate dell'output generato (utilizzando il prompt differente a seconda se il metodo è Turn-Level o Dialogue-Level).
   - Calcola la metrica GPTScore e il punteggio assegnato dall'intelligenza artificiale.

- Salva i risultati nel file JSON originale, aggiungendo nuove feature con una nomenclatura che identifica il metodo e il modello utilizzati.

Ad esempio, nel caso si utilizzi il modello `gpt-4o-mini`, il salvataggio delle feature avviene nel seguente formato:

Listing 2. Esempio di risultato della metrica GPTScore

```
{
    "gpt_score_gpt-4o-mini": 0.28,

}
```

3. **Analisi delle correlazioni statistiche**: L'utente può selezionare il dataset su cui effettuare l'analisi e specificare il percorso di output dove salvare i risultati. Durante l'analisi, vengono calcolate le correlazioni statistiche tra gli score generati dal sistema e le valutazioni fornite dagli esseri umani. Le misure di correlazione includono:
   - **Spearman's Rank Correlation**: per valutare la correlazione tra ranghi ordinali.
   - **Pearson's Correlation**: per misurare la relazione lineare tra due variabili.
   - **Kendall's Tau**: per analizzare la concordanza tra classifiche.

   I risultati vengono infine printati e salvati in formato csv nel file di output specificato dall'utente.

4. IMPLEMENTAZIONE

4.1 ***Try a bot and evaluate*** **.** In questa sezione, viene dettagliata l'operazione che consente agli utenti di interagire con un chatbot e di valutare la qualità della conversazione. L'obiettivo principale è registrare i dialoghi in un formato strutturato per ulteriori analisi, includendo una valutazione soggettiva da parte dell'utente.

Il chatbot utilizza un modello pre-addestrato chiamato `DialoGPT` nella versione `microsoft/DialoGPT-small`, caricato tramite la libreria *transformers*.

La funzione `generate_response` gestisce il processo di creazione delle risposte del bot. Essa combina la cronologia della conversazione con l'input fornito dall'utente per generare un output. Alcuni parametri importanti di questa funzione includono:
- temperature: controlla la creatività nelle risposte del bot (valore impostato a 0.8).
- top_p: definisce la probabilità cumulativa per la selezione dei token successivi (0.9 di default).

L'interazione è gestita all'interno della funzione `generate_bot`. Gli utenti possono

inserire input nel terminale, ricevere risposte dal bot e terminare la conversazione digitando "exit" o "quit".
Al termine del dialogo, l'utente assegna un punteggio alla conversazione tramite la funzione `ask_for_rating`. Questo punteggio, compreso tra 1 e 5, rappresenta una valutazione soggettiva su quanto la conversazione sia risultata interessante.

Le conversazioni vengono salvate in un file JSON (conversation.json) per essere analizzate successivamente. Il salvataggio gestisce anche il caricamento di dati esistenti nel file per preservare dialoghi precedenti.

Al termine dell'operazione viene chiesto se l'utente vuole ottenere anche la valutazione da un modello, eseguendo il task successivo.

4.2 *Test models for GPT score*. L'idea alla base della metrica GPTSCORE è che un modello di pre-addestramento generativo assegna una probabilità più alta a un testo generato di alta qualità che segue una data istruzione e contesto. In questa metodologia, l'istruzione è composta dalla descrizione del task *d* e dalla definizione dell'aspetto *a*. Supponendo che il testo da valutare sia *h* (es., testo sorgente o testo di riferimento), il GPTSCORE è definito come una probabilità condizionale calcolata secondo la formula riportata.

$$\text{GPTScore}(h \mid d, a, S) = \sum_{t=1}^{m} w_t \log p(h_t \mid h_{<t}, T(d, a, S), \theta),$$

dove :
- w_t: è il peso del token nella posizione t,
- $h_{<t}$ rappresenta la sequenza di token precedenti,
- *T(d,a,S)* rappresenta l'integrazione delle informazioni del task, aspetto e contesto,
- θ è l'insieme dei parametri del modello.

L'idea centrale del GPTScore è quella di stimare quanto bene un modello generativo segue istruzioni e contesto, assegnando un punteggio alle risposte generate.

La valutazione avviene attraverso il calcolo della probabilità condizionale delle risposte date, rappresentata come una somma logaritmica pesata sui token che compongono il testo. In pratica, un punteggio alto indica che il modello considera le risposte coerenti e rilevanti rispetto al dialogo e agli aspetti specifici che si vogliono misurare.
Il modello descritto è progettato per valutare la qualità delle risposte generate da un bot in un dialogo con un utente umano, utilizzando GPTScore. L'obiettivo principale è misurare quanto le risposte del bot siano interessanti, con l'implementazione di due tipi di prompt che definiscono diversi metodi di valutazione.

L'interazione inizia con l'utente che fornisce alcune informazioni necessarie per l'esecuzione del modello: il percorso del file JSON contenente i dialoghi, la propria API key per accedere ai modelli GPT, e il tipo di modello da utilizzare (ad esempio, gpt-3.5-turbo, gpt-4, davinci-002 o gpt-4o-mini).Il file JSON fornito in input viene preprocessato per estrarre e formattare un dialogo strutturato, organizzato come una sequenza di domande e risposte tra un essere umano e il bot. Questa struttura viene poi utilizzata per costruire il prompt da passare al modello. A seconda del dataset scelto, il prompt varia per riflettere la modalità di valutazione (TURN o DIALOGUE).
L'aspetto valutato è la qualità generale delle risposte da parte del bot, tenendo conto dei prompt proposti nell'articolo. Il modello supporta due approcci principali per valutare il dialogo:

- **Metodo TURN:**
  - L'istruzione volta a valutare la qualità della risposta più recente fornita da un'intelligenza artificiale all'interno di una conversazione. In particolare, si chiede di rispondere alla domanda se la risposta dell'IA sia soddisfacente o meno, con le opzioni "Sì" o "No". L'analisi si basa sul dialogo precedente tra l'utente e l'intelligenza artificiale, che funge da contesto per giudicare la pertinenza e l'efficacia della risposta fornita.

Listing 3. Prompt per OVERALL , TURN-LEVEL

```
if method == "turn-level":
    prompt = (
        f"Answer␣the␣question␣based␣on␣the␣conversation␣between␣a␣human␣
            and␣AI.\nQuestion:␣Is␣the␣overall␣quality␣of␣the␣AI's␣most␣
            recent␣response␣satisfactory?␣(a)␣Yes.␣(b)␣No.\nConversation:␣{
            dialogue}\nAnswer:␣Yes."
    )
```

- **Metodo DIALOGUE:**
  - Questo approccio include una domanda esplicita che invita il modello a focalizzarsi sulla qualità generale del dialogo.

Listing 4. Prompt per OVERALL, DIALOGUE-LEVEL

```
elif method == "dialogue-level":
    prompt = (
        f"Answer␣the␣question␣based␣on␣the␣conversation␣between␣a␣human␣
            and␣AI.\nQuestion:␣Is␣the␣overall␣quality␣of␣the␣dialogue␣
            satisfactory?␣(a)␣Yes.␣(b)␣No.\nConversation:␣{dialogue}\
            nAnswer:␣Yes."
    )
```

Funzionamento del modello:

1. **Input dell'utente.** L'utente fornisce il percorso del file JSON, l'API key, il tipo di

dialogo e il nome del modello GPT.

2. **Costruzione del prompt.** Il dialogo estratto dal file JSON viene organizzato in un formato strutturato (domande/risposte) e inserito in un prompt, che varia in base al metodo di valutazione scelto.
3. **Inferenza del modello.** Il modello esegue inferenza sul prompt, calcolando la log-probabilità delle risposte in base al contesto fornito.
   Si estrae il punteggio grezzo che rappresenta il GPTScore.
4. Al termine delle valutazioni, il sistema chiede all'utente se desidera effettuare un'analisi di correlazione tra i punteggi generati dal modello e quelli forniti dalle valutazioni umane. Questo passaggio consente di verificare l'allineamento tra il giudizio del GPTScore e la percezione umana, valutando la validità della metrica. Si esegue il task successivo.

4.3 ***Calculate correlations (val_score ↔ gpt_score).*** In questa sezione, si esplora come calcolare e analizzare le correlazioni statistiche tra i punteggi generati dal modello GPT e i punteggi attribuiti dagli esseri umani nei dialoghi valutati. L'obiettivo è determinare se le risposte fornite dal modello siano in linea con le valutazioni umane, cercando di quantificare l'affinità tra i giudizi e fornire una metrica che aiuti a validare il comportamento del bot.

L'analisi della correlazione è stata scelta per diversi motivi:

- Valutazione della Coerenza: Misurare quanto i punteggi generati automaticamente siano coerenti con i giudizi umani è fondamentale per validare il GPTScore.
- Analisi Comparativa: Confrontare i risultati ottenuti su dataset diversi, metodi diversi e modelli diversi consente di validare la metodologia.

Le metriche utilizzate sono : `Spearman correlation, Pearson correlation e Kendall-Tau Correlation`. Queste le operazioni svolte:

- Il dataset è fornito in un file JSON, contiene i punteggi generati dal modello (GPTScore) e i giudizi umani (eval_score) per ogni conversazione. I dati vengono filtrati per garantire che tutte le osservazioni abbiano un punteggio umano valido.
- L'utente può scegliere tra tre opzioni:
- Il sistema calcola le correlazioni utilizzando i punteggi umani e i punteggi del modello GPT corrispondenti. La funzione compute_correlations esegue i calcoli delle correlazioni, visualizzando i risultati in una tabella e salvando i valori in un file CSV.
- I risultati delle correlazioni, tra cui il valore della correlazione e i relativi p-value, vengono visualizzati in una tabella formattata. Inoltre, vengono salvati in un file CSV per permettere ulteriori analisi o riferimenti futuri.

Interpretazione dei risultati. I risultati ottenuti dalla correlazione tra i punteggi GPT e i punteggi umani possono essere interpretati come segue:

- Correlazione di Spearman e Pearson: Valori positivi indicano che le risposte del modello tendono a seguire la stessa direzione delle valutazioni umane, mentre valori negativi suggeriscono il contrario. Un valore di 0 indica assenza di correlazione.

- Correlazione di Kendall-Tau: Simile alla correlazione di Spearman, ma più robusta in presenza di legami nei dati. Valori positivi indicano concordanza nelle classifiche, mentre valori negativi suggeriscono disaccordo.

Per Spearman, Pearson e Kendall-Tau, il valore p indica la significatività statistica della correlazione:

- $p \leq 0.05$: La correlazione è significativa e non dovuta al caso, si rigetta l'ipotesi nulla di non presena di correlazione tra gli items;
- $p > 0.05$: La correlazione non è statisticamente significativa, si accetta l'ipotesi nulla di non correlazione tra gli items.

5. RISULTATI E DISCUSSIONE

Gli esperimenti sono stati condotti su una macchina equipaggiata con un processore Intel Core i9-14900K, che dispone di 24 core suddivisi in 8 Performance-cores (P-core) e 16 Efficiency-cores (E-core), operanti a una frequenza di 6.0 GHz, con supporto per un totale di 32 thread. La macchina è dotata di 32 GB di memoria RAM DDR5 a 7200 MHz e di una scheda grafica NVIDIA GeForce RTX 4070 Ti Super con 16 GB di memoria GDDR6 dedicata. La scheda madre utilizzata è una MSI MAG Z790 TOMAHAWK MAX WIFI, mentre il sistema di raffreddamento è un dissipatore a liquido NZXT Kraken 360. L'alimentazione è garantita da un alimentatore MSI MPG A1000G PCIE5. Il sistema operativo installato è Windows 11 Pro. Per l'esecuzione degli esperimenti è stato utilizzato Python nella versione 3.11.

La Tabella 1 presenta le correlazioni calcolate tra i giudizi umani e le valutazioni fornite dalla metrica GPTScore, analizzate attraverso tre diverse misure statistiche: Spearman, Pearson e Kendall-Tau. I risultati sono suddivisi per dataset e livelli di analisi (dialogue o turn) e considerano due modelli, GPT-4o-mini (abbreviato in Gpt-4o-m) e Davinci-002 (abbreviato in d-002). I dati riportati nella Tabella 1 evidenziano una performance variabile della metrica con differenze sostanziali in base al dataset, al livello di analisi e al modello considerato.

- Per il dataset *convai_2* l+il modello d-002 mostra correlazioni positive e altamente significative con i giudizi umani, in particolare su tutte e tre le metriche statistiche, con valori di correlazione moderati (fino a 0.26 per Spearman). Poichè dalla prima sperimentazione il modello gpt-4o-mini non ha ottenuto risultati rilevanti,

non è stato impiegato per le successive analisi.

- Per il dataset DSCT9, le correlazioni per d-002 sono statisticamente significative ma deboli (0.10 per Spearman).
- Per il dataset Fed, a livello di dialogue, le correlazioni per d-002 non sono significative per nessuna delle tre metriche. Tuttavia, a livello di turn, si osservano correlazioni moderate e significative per tutte le misure, con valori di Spearman e Pearson pari a 0.20 ($p<0.001$) e di Kendall-Tau pari a 0.14 ($p<0.001$).
- Nei dataset Pc-usr e Tc-usr, le correlazioni sono debolmente negative ma significative, seppur sulla soglia del rigetto dell'ipotesi nulla di non correlazione tra le features (alpha=0.05) per il dataset Tc-usr. Un'interpretazione plausibile per le deboli correlazioni negative potrebbe risiedere nella mancanza di una netta distinzione tra contributi umani e generati dall'AI all'interno dei dati. In questi dataset, infatti, il confine tra i giudizi espressi dagli utenti umani e quelli attribuibili all'intelligenza artificiale è poco chiaro, rendendo più complesso per il modello catturare una relazione coerente.

Table 1. Correlazioni tra vari modelli e dataset utilizzando le misure di Spearman, Pearson e Kendall-Tau.

Dataset	Level	Model	Spearman		Pearson		Kendall-Tau	
			Correlation	P-value	Correlation	P-value	Correlation	P-value
Convai_2	Dialogue	Gpt-4o-m	-0.08	0.002***	-0.03	0.25	-0.06	0.002***
Convai_2	Dialogue	d-002	0.26	3.05e-23***	0.18	4.21e-12***	0.19	2.79e-23***
DSCT9	Dialogue	d-002	0.10	6.63e-10***	0.10	5.64e-06***	0.07	6.11e-06***
Fed	Dialogue	d-002	0.03	0.72	0.05	0.56	0.03	0.56
Fed	Turn	d-002	0.20	9.42e-06***	0.20	9.12e-05***	0.14	8.60e-05***
Pc-usr	Turn	d-002	-0.11	0.016**	-0.12	0.012**	-0.10	0.012**
Tc-usr	Turn	d-002	-0.10	0.04*	-0.10	0.05*	-0.07	0.05*

GPT-4o-m e d-002 denotano i modelli GPT-4o-mini e davinci-002, rispettivamente. I simboli ***, ** e * indicano la significatività statistica rispettivamente a : alpha 0, 0.01 e 0.05, rispettivamente.

CONCLUSIONI

Il progetto GPTScore-DialEval si propone come un avanzamento cruciale nel campo della valutazione dei dialoghi generativi, affrontando direttamente alcune delle sfide più pressanti nell'ambito dei modelli di linguaggio naturale. I risultati ottenuti dimostrano l'utilità della metrica GPTScore nel fornire una valutazione quantitativa delle risposte dei chatbot, con correlazioni significative rispetto ai giudizi umani in determinati contesti.

Una delle caratteristiche principali di GPTScore è la sua flessibilità nell'adattarsi a diversi contesti di dialogo grazie a un approccio basato su prompt personalizzabili. I risultati ottenuti hanno dimostrato che il framework può essere configurato per valutare risposte su due livelli:

- Turn-Level: focalizzato sulla qualità della risposta più recente, in relazione al turno immediatamente precedente.
- Dialogue-Level: orientato all'analisi della coerenza complessiva di un'intera conversazione.

Questa capacità di personalizzazione si manifesta nelle differenze di correlazione osservate in vari dataset e livelli di analisi.

Nei dataset ConvAI2 e Fed (Turn-Level), GPTScore ha mostrato correlazioni significative con i giudizi umani (Spearman = 0,26, $p < 0,001$ per ConvAI2 e Spearman = 0,20, $p < 0,001$ per Fed). Ciò evidenzia che la metrica può adattarsi efficacemente a situazioni in cui la qualità immediata delle risposte è importante.

In altri dataset, come nei dataset DSCT9 e Pc-Usr, la metrica ha faticato a catturare il giudizio umano con lo stesso livello di precisione, notando che le sue capacità di personalizzazione non sempre si traducono in performance uniformi.

GPTScore permette inoltre di configurare aspetti come il tipo di modello utilizzato (GPT-4o-mini o Davinci-002) e i parametri specifici per il calcolo della probabilità condizionale dei token. Questo lo rende adattabile a diversi scenari di valutazione e tipi di interazione, un vantaggio evidente rispetto alle metriche tradizionali, statiche e meno flessibili.

Nonostante i risultati incoraggianti, sono emerse alcune limitazioni che meritano ulteriore attenzione:

- Lingua dei dialoghi: Il framework è stato testato esclusivamente su conversazioni in inglese, escludendo altre lingue. L'implementazione potrebbe rappresentare un'ostacolo per la generalizzazione del metodo a livello globale.
- Difficoltà nel distinguere risposte generate e giudizi umani: La qualità dei dataset, soprattutto Pc-Usr e Tc-Usr, non ha sempre fornito un confine chiaro tra il contributo umano e quello generativo, rendendo meno affidabili i punteggi ottenuti. La mancanza di una chiara distinzione tra contributi umani e generati dall'AI nei dataset compromette la capacità del modello di fornire una valutazione coerente.

- Limiti della personalizzazione. Sebbene GPTScore permetta di definire diversi prompt e configurazioni, la sua capacità di adattarsi a specifiche esigenze dipende fortemente dalla qualità e dalla chiarezza dei dataset utilizzati. Dataset poco strutturati o privi di annotazioni dettagliate limitano l'efficacia della personalizzazione. L'uso esclusivo di prompt strutturati limita l'applicabilità del framework in scenari che richiedono una valutazione più dinamica o creativa, come la valutazione della "umanità" delle risposte o della loro empatia.

Questo lavoro rappresenta una base solida per l'automazione della valutazione qualitativa dei dialoghi generativi. Tuttavia, il futuro di GPTScore-DialEval potrebbe includere:

1. **Espansione verso dialoghi multilingue.** L'adattamento del framework a dialoghi in lingue diverse dall'inglese rappresenta una priorità assoluta. Questo non solo aumenterebbe la generalizzabilità del metodo, ma aprirebbe anche nuove opportunità per applicazioni in contesti globali. Parallelamente, l'adozione di dataset che riflettano entità culturali differenti consentirebbe di catturare aspetti qualitativi unici delle conversazioni umane.
2. **Valutazione di caratteristiche differenti.** Nel caso specifico la metrica è stata testata per valutare la soddisfazione generale del discorso o dell'ultima risposta, ma potrebbe essere adattata per valutare contesti e/o emozioni diverse. Questo richiede la modifica del prompt in maniera dettagliata e una formattazione non banale dai dati in input.
3. **Implementazione in applicazioni interattive**. Un possibile sviluppo è rappresentato dall'integrazione di GPTScore in applicazioni pratiche, come piattaforme educative, assistenti sanitari virtuali o sistemi di customer service. In questi contesti, la metrica potrebbe essere utilizzata per monitorare e migliorare continuamente l'efficacia e la qualità delle interazioni con gli utenti finali, mantenendo i feedback.