ANALISI DELLE METRICHE DI VALUTAZIONE G-EVAL SU SU DIVERSI DATASET DI BENCHMARK

Salvatore Sirica

01

PROBLEMA

I sistemi di Natural Language Generation (NLG) generano dialoghi che necessitano di valutazioni per migliorarne la qualità.

01

02

Le metriche tradizionali (es. BLEU, ROUGE) presentano limiti, non correlando sempre bene con i giudizi umani.

02

OBIETTIVO

Sviluppare un framework di valutazione basato su G-EVAL per misurare la qualità dei dialoghi in modo più affidabile

03

G-EVAL

01 Valutazione centrata sull'uomo: Utilizza GPT-4 con il ragionamento Chain-of-Thought (CoT) per ottenere valutazioni strettamente allineate ai giudizi umani, applicabili a vari compiti NLG, come la generazione di dialoghi e la sintesi testuale.

02 Paradigma di compilazione Form-Filling: Impiega prompt strutturati e passi intermedi CoT per standardizzare i punteggi e fornire un processo di valutazione dettagliato e spiegabile.

03 Punteggi dettagliati e precisi: Introduce una normalizzazione basata sulle probabilità per generare punteggi continui e accurati, superando metriche tradizionali come BLEU e ROUGE in termini di correlazione con i giudizi umani.

04

DATASET UTILIZZATI

FED (Feedback Evaluation Dataset): analizza la qualità complessiva dei dialoghi con annotazioni umane

TC_USR (Topical Chat): dialoghi su argomenti specifici con valutazioni di engagingness e groundedness.

PC_USR: conversazioni orientate alla precisione e comprensione contestuale.

ConvAI2: dialoghi di chatbot con annotazioni di qualità umana

DSTC9: dataset per benchmarking di dialoghi orientati al task.

05

FLUSSO G-EVAL

Fase 2

Generazione dei prompt per valutare le metriche definite

Fase 1

Caricamento del dataset

Fase 3

Invio a G-EVAL per ottenere punteggi

06

FASE 1

(CARICAMENTO DEL DATASET)

I dataset vengono caricati in formato JSON ed ogni record include:

- Contesto del dialogo.
- Risposta generata dal sistema (In caso di turn-level).
- Annotazioni umane (Overall Score).

Questi dati costituiscono la base per la generazione dei prompt e l'analisi successiva.

07

FASE 2

(GENERAZIONE DEI PROMPT)

Per ogni dialogo o risposta, viene generato un prompt personalizzato. I prompt definiscono le istruzione da dare al modello per generare un punteggio complessivo in base al tipo di task (turn-level e dialog-level)

Esempio di sostituzione nei template:

- {{context}} → Dialogo completo.
- {{response}} → Risposta generata (in caso di turn level).

08

FASE 3
(VALUTAZIONE)

I prompt vengono costruiti ed inviati al modello GPT4. Possiamo impostare valori come temperature e top_p per guidare la generazione della risposta.

Il risultato viene restituito come un punteggio complessivo che indica la qualità del dialogo o della risposta.

Questo processo automatizzato consente una valutazione rapida e standardizzata dei dialoghi.

09

ESEMPIO DI PROMPT

You will be given a conversation between two individuals and one potential response for the next turn in the conversation.

Your task is to evaluate the overall quality of the response.

Evaluation Criteria:
Overall Quality (1-3): Is the overall quality of the response satisfactory?
- Score 1 (Unsatisfactory): ...
- Score 2 (Satisfactory): ...
- Score 3 (Excellent): ...

Dialogue Context: {{context}}
Response: {{response}}

ESEMPIO DI COT

Evaluation Steps:

1. Read the conversation and response carefully.
2. Assign a score between 1 and 3 based on the criteria above.
3. Provide a brief explanation for your rating, referring to specific aspects of the response and the conversation.

FLUSSO DI ANALISI

Fase 1

Calcolo delle distanze

Fase 2

Generazione risultati

FASE 1

(CALCOLO DELLE DISTANZE)

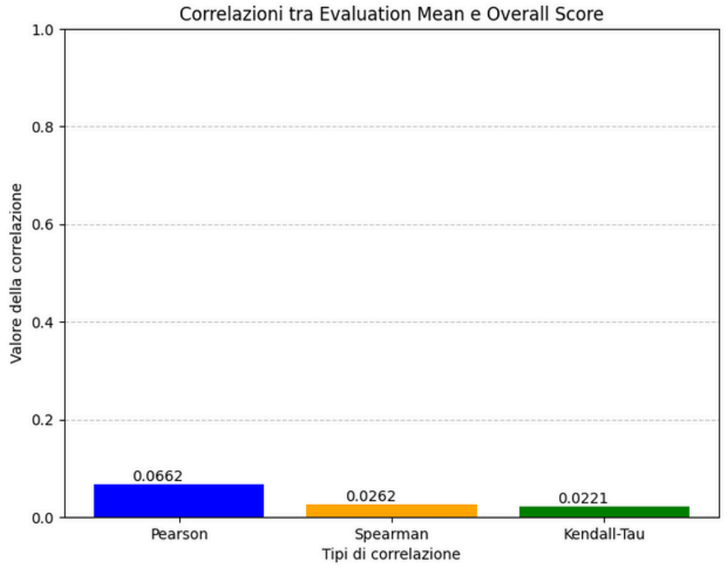
Valutare la relazione tra i punteggi generati automaticamente (evaluation mean) e quelli forniti dagli annotatori umani (overall score).

Metriche Utilizzate:

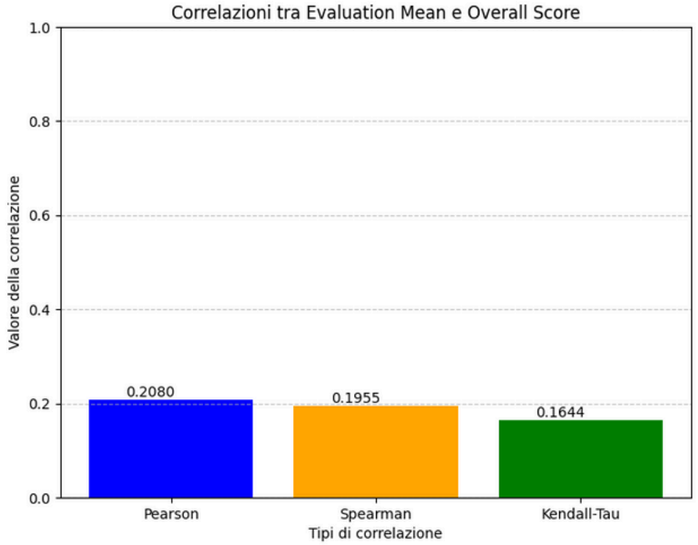
- Pearson: Correlazione lineare tra i punteggi.
- Spearman: Concordanza nei ranghi.
- Kendall-Tau: Monotonicità tra i dati.

RISULTATI TURN LEVEL

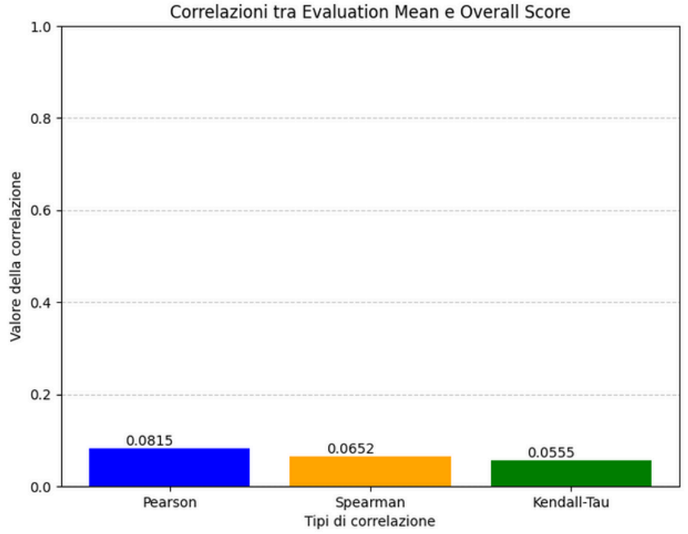
Le correlazioni risultano generalmente basse in tutti i dataset a livello di singolo turno. Questo indica che la sola metrica di Overall Score non è sufficiente per catturare gli aspetti complessi della qualità del dialogo. La valutazione a livello di singola risposta può perdere informazioni cruciali come il contesto e la coerenza globale della conversazione. È necessario considerare metriche più dettagliate per migliorare l'allineamento con i giudizi umani.



Risultati FED (Turn-Level)



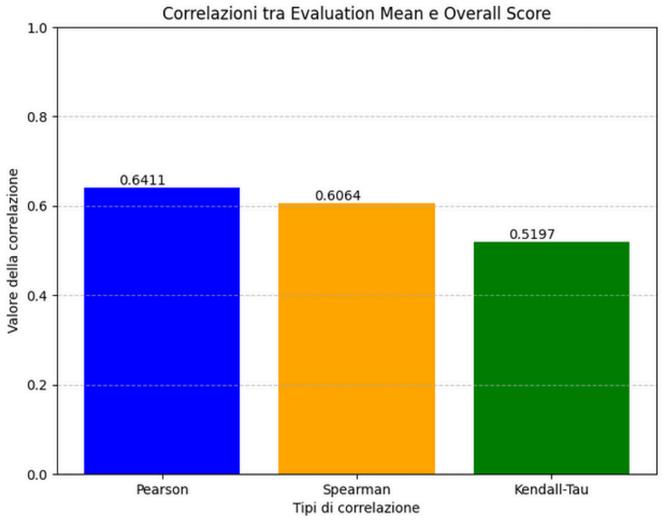
Risultati TC_USR



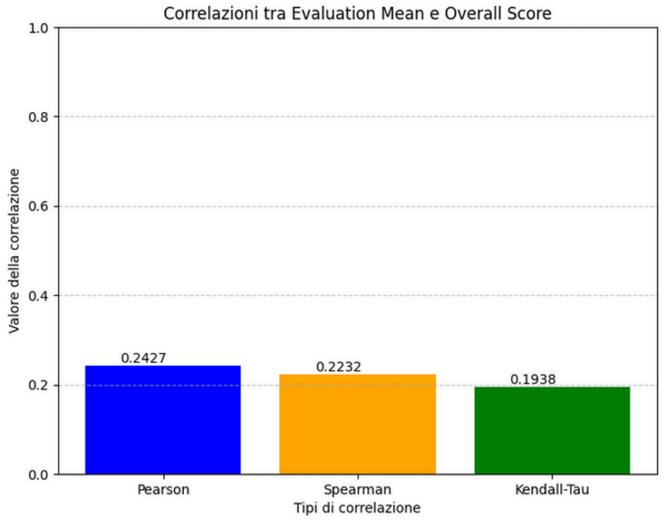
Risultati PC_USR

RISULTATI DIALOG LEVEL

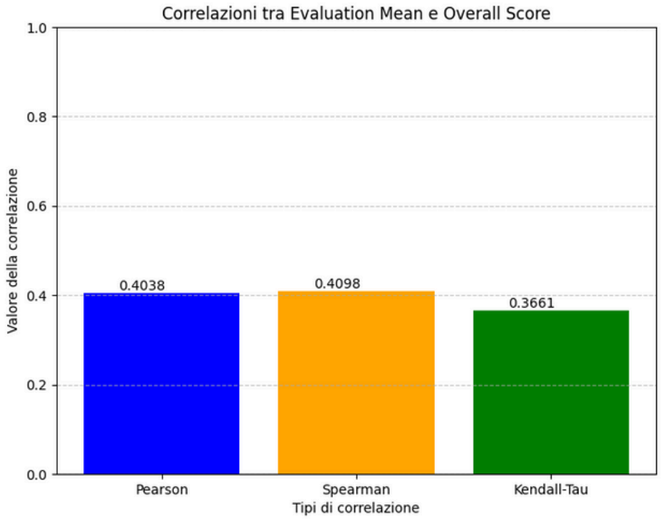
L'analisi dei dataset a livello di dialogo evidenzia una correlazione significativamente più alta rispetto al livello di turn. Questo suggerisce che la valutazione dell'intero dialogo permette di cogliere meglio gli aspetti globali della qualità delle risposte. Tuttavia, la metrica complessiva (Overall) non è ancora sufficiente a catturare tutti gli aspetti della qualità, come l'engagingness e la groundedness, che sono cruciali per una valutazione più accurata.



Risultati FED (Dialog-Level)



Risultati DSTC



Risultati CONVAI

CONCLUSION

01 Limiti Identificati:

- Le correlazioni ottenute per i dataset turn-level risultano generalmente basse, indicando che l'Overall Score da solo non è sufficiente a catturare aspetti cruciali della qualità del dialogo.
- Nei dataset dialog-level, sebbene le correlazioni siano più elevate, rimangono margini di miglioramento per riflettere accuratamente le valutazioni umane.
- È necessaria una maggiore granularità nelle metriche per catturare sfumature come l'engagingness e la groundedness.

02 Prospettive Future:

- Integrare nuove metriche specifiche per aspetti chiave del dialogo, come la coerenza, la pertinenza e l'engagement.
- Validare il framework su un insieme più diversificato di dataset per migliorare la generalizzabilità delle valutazioni.

GRAZIE PER L'ATTENZIONE