Analisi delle Metriche di Valutazione G-EVAL su Dataset di Dialogo: FED e Topical Chat

Salvatore Sirica

Università degli Studi di Salerno

January 20, 2025

Contents

1 Introduzione

Negli ultimi anni, i progressi nei modelli di *Natural Language Generation* (NLG) hanno portato alla generazione di testi di alta qualità che spesso risultano indistinguibili dai testi scritti da esseri umani. Tuttavia, la valutazione automatica della qualità di questi testi rimane una sfida aperta, soprattutto in compiti creativi e diversificati. Le metriche tradizionali, come BLEU, ROUGE e METEOR, si basano sul confronto con testi di riferimento (*reference-based evaluation*) e mostrano una correlazione limitata con i giudizi umani, in particolare nei contesti di generazione aperta. La necessità di riferimenti aumenta inoltre i costi e la complessità nei nuovi task.

Recentemente, i modelli di linguaggio di grandi dimensioni (*Large Language Models*, LLM) sono stati proposti come valutatori senza riferimento (*reference-free evaluation*). Questi approcci sfruttano la capacità degli LLM di valutare testi generati basandosi sulla probabilità di generazione, ma mancano ancora di un'adeguata corrispondenza con i giudizi umani. Per affrontare queste limitazioni, l'articolo propone **G-EVAL**, un framework innovativo che utilizza gli LLM, in particolare GPT-4, combinando il paradigma di *Chain-of-Thought* (CoT) con un processo di valutazione basato su moduli (*form-filling paradigm*).

1.1 Il framework G-EVAL

G-EVAL è progettato per fornire valutazioni della qualità del testo più affidabili e correlate con i giudizi umani. Il framework si compone di tre componenti principali:

1. **Prompt di valutazione**: definisce il compito di valutazione e i criteri specifici (ad esempio coerenza, fluency, rilevanza).

2. **Chain-of-Thought (CoT)**: guida il modello a eseguire una sequenza di passaggi intermedi per valutare i testi, migliorando l'affidabilità del processo di valutazione.

3. **Funzione di scoring**: utilizza le probabilità dei token generate dall'LLM per calcolare un punteggio continuo, fornendo valutazioni più granulari rispetto ai punteggi discreti tradizionali.

1.2 Contributi principali dell'articolo

L'articolo evidenzia i seguenti contributi principali:

- G-EVAL supera le metriche tradizionali (BLEU, ROUGE) e basate su LLM precedenti in termini di correlazione con i giudizi umani.
- L'utilizzo di CoT migliora la qualità delle valutazioni, fornendo un contesto più ricco e dettagliato per i criteri di valutazione.
- G-EVAL offre punteggi più fini gra zie alla normalizzazione delle probabilità dei token, riducendo i bias verso punteggi discreti dominanti.
- L'analisi del comportamento di G-EVAL rivela un potenziale bias verso i testi generati da LLM, evidenziando un'area per ulteriori ricerche.

1.3 Applicazioni del framework

Il framework è stato testato su due compiti principali di NLG:

- **Text summarization**: sintesi di articoli utilizzando benchmark come SummEval e QAGS. G-EVAL ha dimostrato una correlazione superiore con i giudizi umani rispetto alle metriche esistenti.
- **Dialogue generation**: generazione di risposte per conversazioni, valutata con benchmark come Topical-Chat. Anche in questo contesto, G-EVAL ha superato gli approcci tradizionali.

1.4 Limitazioni e prospettive future

L'articolo evidenzia anche alcune limitazioni di G-EVAL:

- **Bias verso i testi generati da LLM**: G-EVAL tende a favorire i testi generati dagli LLM rispetto a quelli scritti da esseri umani, anche quando i giudizi umani preferiscono quest'ultimi.
- **Necessità di ulteriori studi**: è richiesto un approfondimento per ridurre il bias e migliorare ulteriormente la corrispondenza con i giudizi umani.

Le prospettive future includono l'esplorazione di metodi per mitigare i bias intrinseci e migliorare l'allineamento con le preferenze umane, rendendo G-EVAL uno strumento sempre più utile per la valutazione della qualità dei testi generati.

2 Metriche e Metodologia

2.1 Introduzione

La valutazione della qualità dei testi generati da sistemi di *Natural Language Generation* (NLG) è un problema fondamentale per migliorare le prestazioni di tali modelli. Le metriche di valutazione devono essere affidabili e strettamente correlate ai giudizi umani per garantire un progresso significativo nel campo.

Nel presente lavoro, l'attenzione è focalizzata sull'uso del punteggio complessivo (*Overall Score*) come principale metrica di valutazione. L'obiettivo è ottenere una correlazione elevata tra i punteggi assegnati automaticamente dai modelli di linguaggio e quelli forniti da annotatori umani. A tale scopo, è stato adottato il framework **G-EVAL**, che utilizza tecniche avanzate di ragionamento per migliorare l'affidabilità delle valutazioni.

2.2 Descrizione della Metrica di Valutazione

Nel nostro studio, abbiamo adottato il punteggio complessivo (**Overall Score**) per valutare la qualità delle risposte generate dai modelli NLG. Questo punteggio fornisce una valutazione unificata che tiene conto di diversi aspetti, tra cui la naturalezza, la coerenza, l'interesse e la pertinenza rispetto al contesto.

- **Obiettivo**: Fornire un'indicazione globale della qualità della risposta, considerando tutti gli aspetti rilevanti in un unico valore.
- **Calcolo**: Il punteggio viene assegnato utilizzando il framework G-EVAL, che genera una valutazione basata su un prompt specifico e sulla probabilità di generazione del modello.
- **Limiti**: La valutazione complessiva potrebbe nascondere dettagli specifici relativi a singole caratteristiche della risposta, rendendo più difficile l'interpretazione di aree di miglioramento specifiche.

2.3 Processo di Valutazione

Il processo di valutazione si articola nei seguenti passaggi:

1. **Caricamento del dataset**: I dati sono stati letti da un file JSON contenente il contesto del dialogo, la risposta generata e il punteggio complessivo annotato dagli esperti umani.

2. **Generazione dei prompt**: È stato utilizzato un template specifico per generare i prompt richiesti per la valutazione.

```
# Genera il prompt per la valutazione
prompt = g_eval.generate_prompt(prompt_template,
    full_conversation, response)
```

3. **Valutazione tramite G-EVAL**: Ogni prompt è stato inviato al modello per ottenere il punteggio di qualità complessiva.

```
# Effettua la richiesta al modello
evaluation = g_eval.send_request(prompt)
```

4. **Parsing dei risultati**: Il punteggio complessivo è stato estratto e confrontato con il punteggio umano per l'analisi della correlazione.

2.4 Esempio di Valutazione

Un esempio di valutazione eseguita sul dataset è riportato di seguito:

- **Contesto del dialogo**: "User: Qual è il tempo oggi?
  System: Oggi sarà soleggiato con temperature attorno ai 25°C."
- **Risposta generata**: "System: È una giornata perfetta per una passeggiata."
- **Punteggi ottenuti**:
  - Naturalness: 5
  - Coherence: 4
  - Engagingness: 5
  - Groundedness: 4

2.5 Calcolo delle Correlazioni e Distanze

Per valutare l'accuratezza del sistema di valutazione, sono state calcolate le correlazioni tra il punteggio generato automaticamente (*evaluation overall*) e il punteggio umano (*overall score*). Le correlazioni sono state misurate utilizzando le seguenti metriche:

- **Pearson**: Valuta la relazione lineare tra le due serie di punteggi.
- **Spearman**: Misura la relazione di monotonicità tra i punteggi assegnati.
- **Kendall-Tau**: Analizza la concordanza dei ranghi tra le valutazioni umane e automatiche.

Inoltre, è stata analizzata la distanza assoluta media tra i punteggi umani e quelli generati automaticamente per identificare discrepanze significative.

2.6 Esempio di Valutazione

Un esempio di valutazione eseguita sul dataset è riportato di seguito:

- **Contesto del dialogo**: "User: Qual è il tempo oggi?
  System: Oggi sarà soleggiato con temperature attorno ai 25°C."
- **Risposta generata**: "System: È una giornata perfetta per una passeggiata."
- **Punteggi ottenuti**:
  - Overall Score (umano): 4
  - Overall Score (modello): 3

3 Implementazione

3.1 Implementazione del Framework G-EVAL

Per la valutazione della qualità dei dialoghi generati, è stato sviluppato un framework basato su G-EVAL, integrando il modello GPT-4 per ottenere un punteggio complessivo (**Overall Quality**) su una scala da 1 a 3. Il framework si basa su una classe Python, `GEvalAPI`, progettata per:

- Caricare template di prompt specifici per la valutazione.
- Generare prompt personalizzati sostituendo i placeholder con i contenuti dei dialoghi.
- Inviare richieste al modello GPT-4 e raccogliere le valutazioni.

La struttura del framework è modulare, permettendo l'estensione a nuovi scenari di valutazione.

3.1.1 Caricamento dei Template

Il primo passo del framework consiste nel caricamento dei template di prompt, definiti in file di testo. Questo consente di mantenere i prompt separati dal codice e facilmente modificabili. Il caricamento avviene tramite un metodo dedicato:

```
def load_prompt_template(self, file_path):
    with open(file_path, "r") as f:
        return f.read()
```

3.1.2 Generazione del Prompt

Per ogni dialogo o risposta, viene generato un prompt personalizzato sostituendo i placeholder ({{`context`}}, {{`response`}}) con i dati specifici. Ad esempio:

```
def generate_prompt(self, template, context, response):
    prompt = template.replace("{{context}}", context)
                     .replace("{{response}}", response)
    return prompt
```

3.1.3 Integrazione con il Modello GPT-4

Il prompt generato viene inviato al modello GPT-4 per ottenere il punteggio complessivo. Il framework supporta parametri configurabili come il numero di risposte (`n`) e la temperatura (`temperature`) per gestire la generazione:

```
def send_request(self, prompt, n=1, max_tokens=50, temperature=0):
    response = self.client.chat.completions.create(
        model=self.model,
        messages=[{"role": "system", "content": prompt}],
```

```
        n=n,
        max_tokens=max_tokens,
        temperature=temperature,
    )
    return [choice.message.content for choice in response.choices]
```

3.2 Prompt Utilizzati

I prompt sono stati progettati per adattarsi a diversi contesti di valutazione. Ogni prompt guida il modello nella valutazione della qualità complessiva, utilizzando un approccio *Chain-of-Thought* (CoT) per migliorare l'accuratezza.

3.2.1 Esempio di prompt

Utilizzato per valutare una singola risposta rispetto al contesto fornito:

```
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.

Please follow these instructions carefully and respond ONLY with the
    score in the format provided below. Do not include any additional
    text, comments, or explanations. If the response cannot be
    evaluated, assign a score of ‘0‘.

Evaluation Criteria:
Overall Quality (1-3): Is the overall quality of the response
    satisfactory?
- Score 1 (Unsatisfactory): The response does not align with the tone,
    context, or intent of the conversation. It may include irrelevant
    or incoherent content, disrupting the flow.
- Score 2 (Satisfactory): The response is generally appropriate,
    maintaining coherence and relevance. It may have minor issues, such
    as awkward phrasing or missing some context, but it still
    contributes to the conversation.
- Score 3 (Excellent): The response is highly relevant, engaging, and
    natural. It clearly addresses the topic, enhances the conversation,
    and has no significant issues in tone or content.
```

```
12
13 Evaluation Steps:
14 1. Read the conversation and response carefully.
15 2. Assign a score between 1 and 3 based on the criteria above.
16
17 Dialogue Context:
18 {{context}}
19
20 Response:
21 {{response}}
22
23 Response Format:
24 Overall Quality: <score>
```

3.3 Ruolo del Chain-of-Thought (CoT)

Il framework utilizza un approccio *Chain-of-Thought* (CoT), che permette al modello di ragionare passo dopo passo. Questo migliora l'accuratezza della valutazione, in particolare per:

- **Dialoghi completi**: Mantiene la coerenza tra i turni di dialogo.
- **Singole risposte**: Analizza dettagliatamente il contesto per valutazioni più precise.

3.4 Calcolo delle Correlazioni e Distanze

Per valutare l'accuratezza del sistema di valutazione, sono state calcolate le correlazioni tra i punteggi generati automaticamente (*Overall Quality predetta*) e quelli forniti dagli annotatori umani (*Overall Quality umana*). Le correlazioni sono state misurate utilizzando le seguenti metriche:

- **Pearson**: Misura la relazione lineare tra i punteggi predetti e quelli umani.
- **Spearman**: Analizza la relazione monotona tra i punteggi ordinati.
- **Kendall-Tau**: Valuta la concordanza nei ranghi dei punteggi.

4 I Dataset

4.1 Introduzione

Per valutare le prestazioni del framework G-EVAL, sono stati utilizzati cinque dataset distinti: **FED (Feedback Evaluation Dataset)**, **Topical-Chat (TC-USR)**, **Persona-Chat (PC-USR)**, **ConvAI2** e **DSTC9**. Questi dataset contengono dialoghi annotati con valutazioni umane che consentono di confrontare le prestazioni del modello rispetto ai giudizi forniti dagli annotatori.

4.2 Descrizione dei Dataset

4.2.1 FED (Feedback Evaluation Dataset)

Il dataset FED è stato sviluppato per valutare modelli di dialogo open-domain, fornendo annotazioni su diversi aspetti qualitativi della conversazione. Ogni esempio include:

- **context**: la sequenza completa del dialogo tra l'utente e il sistema.
- **response**: la risposta generata dal sistema.
- **annotations**: punteggi assegnati dagli annotatori per criteri come *Interesting*, *Engaging*, *Specific*, *Relevant*, *Correct*, *Fluent*, e un *Overall score*.
- **system**: il nome del sistema che ha generato la risposta.

4.2.2 Topical-Chat (TC-USR)

Il dataset TC-USR contiene dialoghi informativi tra utenti su argomenti specifici, accompagnati da annotazioni su metriche chiave. Ogni esempio è strutturato come segue:

- **context**: un insieme di scambi testuali tra due partecipanti.
- **fact**: un fatto di supporto relativo al tema della conversazione.
- **responses**: una lista di risposte generate da vari modelli, con le seguenti annotazioni umane:
  - *Understandable*
  - *Natural*

- *Maintains Context*
- *Engaging*
- *Uses Knowledge*
- *Overall score*

4.2.3 Persona-Chat (PC-USR)

Il dataset PC-USR contiene dialoghi tra interlocutori con una specifica personalità assegnata, con l'obiettivo di valutare la coerenza e la personalizzazione delle risposte. Le proprietà principali includono:

- **context**: il dialogo tra due partecipanti.
- **responses**: risposte generate con valutazioni sui seguenti aspetti:
  - *Fluency*
  - *Coherence*
  - *Engagingness*
  - *Consistency*
  - *Overall score*

4.2.4 ConvAI2

Il dataset ConvAI2 è stato sviluppato per la valutazione di modelli di dialogo focalizzati sulla personalizzazione. Ogni dialogo include:

- **dialog**: una sequenza di turni di conversazione tra due partecipanti.
- **eval_score**: un punteggio complessivo assegnato dagli annotatori per valutare la qualità della conversazione.

4.2.5 DSTC9

Il dataset DSTC9 (Dialog System Technology Challenge) è focalizzato sulla valutazione di modelli di dialogo in contesti informativi e conversazionali. La struttura del dataset include:

- **contexts**: sequenza di messaggi in una conversazione.
- **scores**: punteggi umani assegnati all'intero dialogo per riflettere la qualità complessiva.

4.3 Preprocessing dei Dataset

I dataset sono stati preprocessati per garantire uniformità e coerenza nelle analisi. Le principali operazioni di preprocessing includono:

- **Normalizzazione del testo**: rimozione di caratteri speciali e uniformazione del formato del dialogo.
- **Conversione dei dati**: adattamento dei formati per l'elaborazione con il framework G-EVAL.

4.4 Modalità di Valutazione

Le conversazioni nei dataset sono state valutate in due modalità principali:

- **Turn-level**: valutazione delle singole risposte in un dialogo.
- **Dialog-level**: valutazione dell'intero dialogo come un'unica unità.

5 Risultati

5.1 Valutazioni dei Dataset

Per analizzare la qualità della valutazione automatica rispetto ai giudizi umani, sono stati calcolati i coefficienti di correlazione tra i punteggi generati dal sistema (*evaluation mean*) e i punteggi assegnati dagli annotatori umani (*overall score*). I dataset utilizzati per l'analisi includono **FED**, **TC_USR**, **PC_USR**, **DSTC9** e **ConvAI2**. Le metriche di correlazione calcolate sono:

- **Pearson**: Misura della relazione lineare tra le due variabili.
- **Spearman**: Valuta la monotonicità tra i punteggi.
- **Kendall-Tau**: Analizza la concordanza tra le variabili ordinali.

5.1.1 FED (Feedback Evaluation Dataset)

Il dataset FED è stato analizzato considerando sia la valutazione *turn-level* che *dialog-level*. I risultati delle correlazioni sono riportati di seguito.

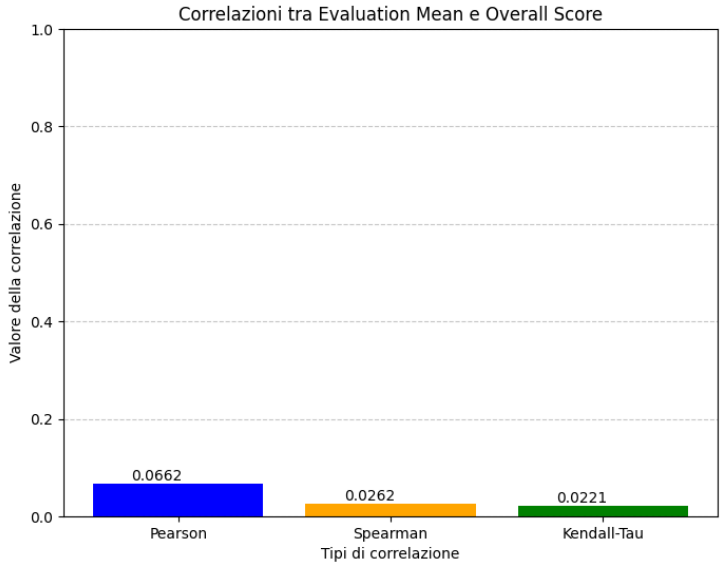

Figure 1: Correlazioni tra *Evaluation Mean* e *Overall Score* per il dataset FED (Turn-Level).

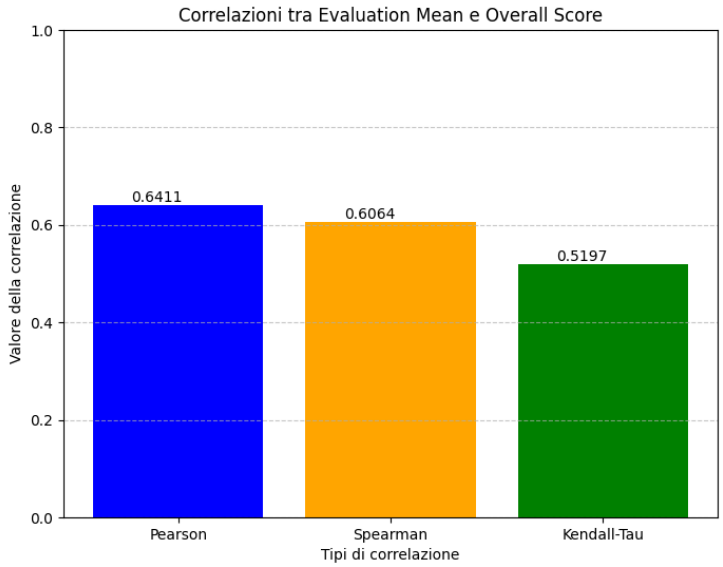

Figure 2: Correlazioni tra *Evaluation Mean* e *Overall Score* per il dataset FED (Dialog-Level).

5.1.2 TC_USR

Il dataset TC_USR è stato analizzato per comprendere la relazione tra i punteggi generati e i giudizi umani.

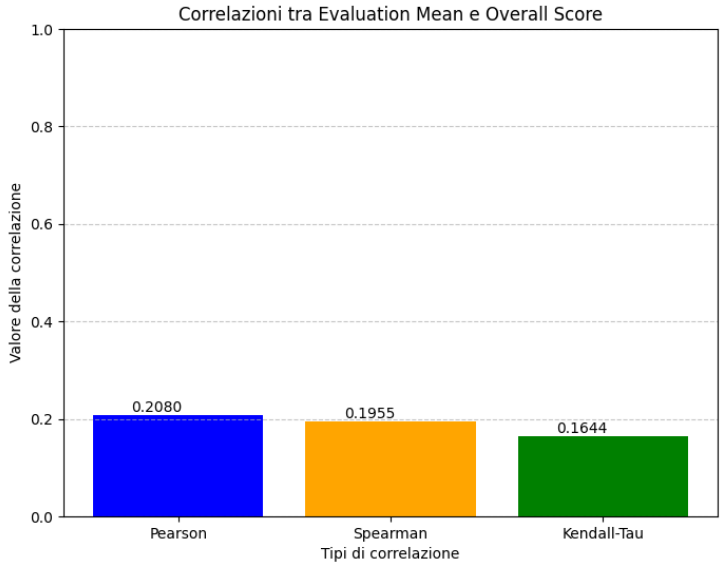

Figure 3: Correlazioni tra *Evaluation Mean* e *Overall Score* per il dataset TC_USR.

5.1.3 PC_USR

Il dataset PC_USR è stato analizzato per verificare la robustezza delle metriche di valutazione.

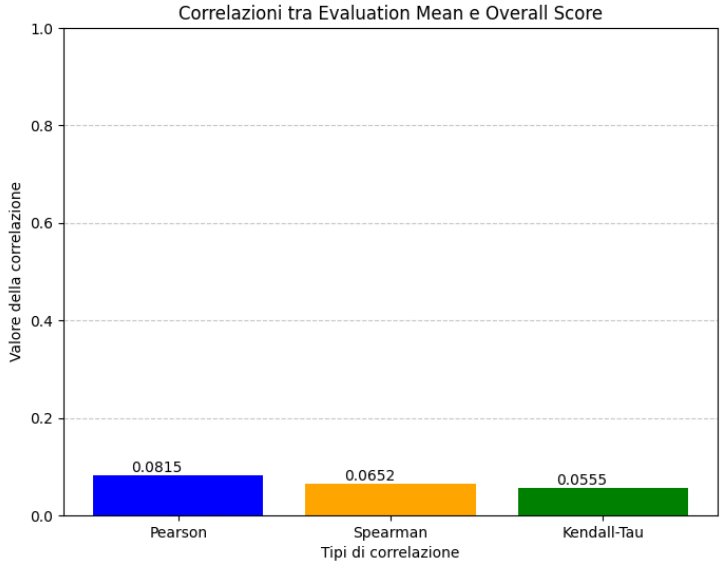

Figure 4: Correlazioni tra *Evaluation Mean* e *Overall Score* per il dataset PC_USR.

5.1.4 DSTC9

Il dataset DSTC9 è stato valutato per analizzare la qualità delle risposte in un contesto di dialogo strutturato.

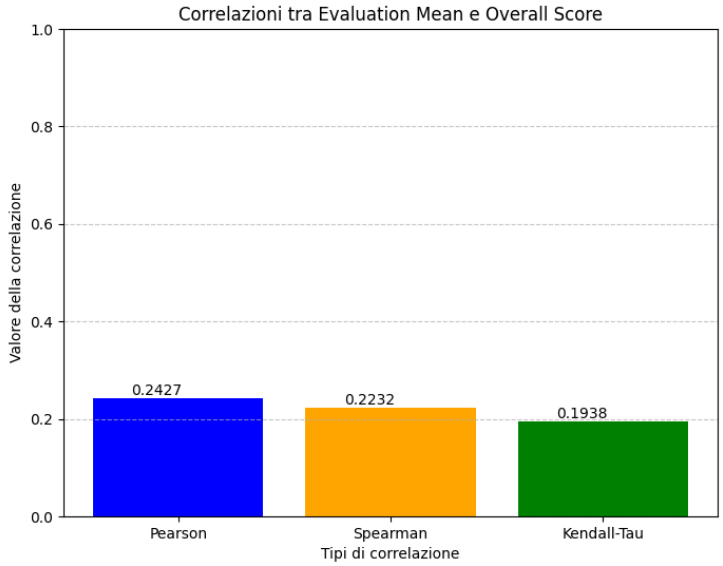

Figure 5: Correlazioni tra *Evaluation Mean* e *Overall Score* per il dataset DSTC9.

5.1.5 ConvAI2

Il dataset ConvAI2 è stato utilizzato per testare le capacità di correlazione tra punteggi automatici e annotazioni umane.

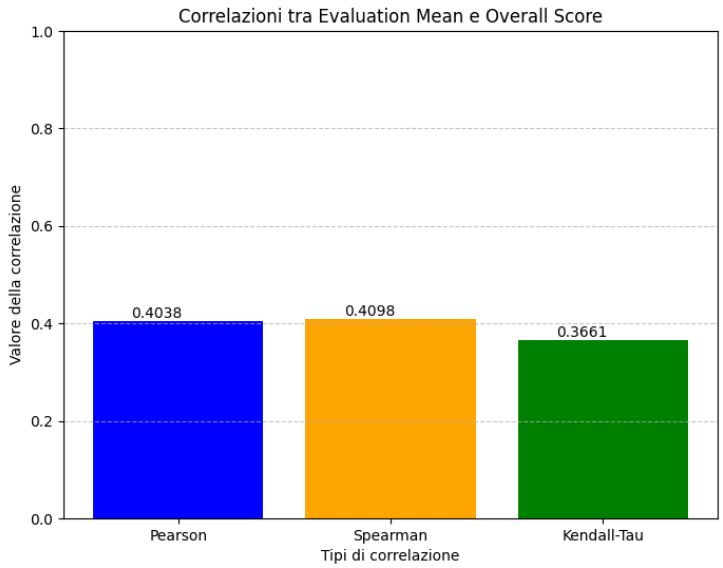

Figure 6: Correlazioni tra *Evaluation Mean* e *Overall Score* per il dataset ConvAI2.

5.2 Discussione dei Risultati

Dall'analisi dei risultati emerge quanto segue:

- **FED**: Le correlazioni per il turn-level risultano inferiori rispetto al dialog-level, suggerendo che la valutazione complessiva del dialogo potrebbe essere più affidabile rispetto alla valutazione di singole risposte.
- **TC_USR**: Mostra correlazioni relativamente più alte, indicando una buona relazione tra punteggi generati e giudizi umani.
- **PC_USR**: Correlazioni moderate, con la necessità di ulteriori miglioramenti nelle metriche di valutazione.
- **DSTC9**: Le correlazioni risultano più basse rispetto agli altri dataset, suggerendo una difficoltà nel catturare la qualità delle risposte in contesti strutturati.
- **ConvAI2**: Mostra un buon livello di correlazione, segnalando la capacità del modello di riflettere adeguatamente i giudizi umani.

Questi risultati suggeriscono che le metriche di valutazione automatica necessitano di ulteriori ottimizzazioni per migliorare la loro capacità di riflettere fedelmente i giudizi umani nei diversi contesti di dialogo.