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

Salvatore Sirica

Università degli Studi di Salerno

January 8, 2025

Contents

1 Introduzione **3**

1.1 Il framework G-EVAL 3

1.2 Contributi principali dell'articolo 4

1.3 Applicazioni del framework 4

1.4 Limitazioni e prospettive future 4

2 Metriche e Metodologia **5**

2.1 Introduzione 5

2.2 Descrizione delle Metriche 5

2.2.1 Naturalness 5

2.2.2 Coherence 6

2.2.3 Engagingness 6

2.2.4 Groundedness 6

2.3 Processo di Valutazione 7

2.4 Esempio di Valutazione 7

2.5 Calcolo delle Correlazioni e Distanze 8

2.5.1 Calcolo delle Correlazioni 8

3 Implementazione **8**

3.1 Implementazione del Framework G-EVAL 8

3.1.1 Caricamento dei Template 9

3.1.2 Generazione del Prompt 9

3.1.3 Integrazione con il Modello GPT-4 9

3.2 Prompt Utilizzati 10

3.2.1 Esempio di prompt 10

3.3 Ruolo del Chain-of-Thought (CoT) 11

3.4 Conclusione 11

4 I Dataset **12**

4.1 Introduzione 12

4.2 Descrizione dei Dataset 12

4.2.1 FED (Feedback Evaluation Dataset) 12

4.2.2 Topical-Chat 12

4.3 Preprocessing 13

5 Risultati **13**

5.1 Valutazioni per FED 13

5.1.1 Risultati delle Correlazioni 13

5.1.2 Analisi Grafica 14

5.1.3 Discussione dei Risultati 14

5.2 Valutazioni per TC_USR 16

5.2.1 Risultati delle Correlazioni 16

5.2.2 Analisi Grafica 17

5.2.3 Discussione dei Risultati 17

5.3 Prospettive di miglioramento 19

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. Esistono due categorie principali di metriche:

- **Metriche basate su riferimento**: confrontano il testo generato con un riferimento umano.
- **Metriche senza riferimento**: valutano il testo generato senza la necessità di un riferimento.

Nel contesto del progetto, è stato sviluppato un sistema per valutare la qualità dei dialoghi generati dai modelli di *Natural Language Generation* (NLG) utilizzando il framework G-EVAL. Le metriche di valutazione adottate si concentrano su quattro aspetti fondamentali: **Naturalness**, **Coherence**, **Engagingness** e **Groundedness**. Inoltre, per ogni risposta del modello, è stato calcolato un punteggio complessivo (**Overall Score**) basato sulle annotazioni fornite.

2.2 Descrizione delle Metriche

2.2.1 Naturalness

- **Obiettivo**: Valutare quanto la risposta generata suoni naturale, umana e colloquiale.
- **Calcolo**: Il punteggio di naturalness è stato estratto dall'output del modello G-EVAL basandosi sulla qualità linguistica e sul tono della risposta.

- **Limiti**: In alcuni casi, risposte grammaticalmente corrette ma non colloquiali possono ottenere punteggi elevati, nonostante manchino di autenticità.

2.2.2 Coherence

- **Obiettivo**: Misurare la coerenza del dialogo, valutando se la risposta segue logicamente il contesto precedente.
- **Calcolo**: La coerenza è stata determinata confrontando la struttura e il contenuto della risposta con il contesto del dialogo.
- **Limiti**: Il calcolo della coerenza può risentire di risposte ambigue o contesti poco chiari.

2.2.3 Engagingness

- **Obiettivo**: Valutare quanto la risposta generata sia interessante e coinvolgente per l'utente.
- **Calcolo**: Il punteggio di engagingness è stato calcolato basandosi sull'interesse potenziale che la risposta potrebbe suscitare nell'utente.
- **Limiti**: Risposte brevi o molto specifiche possono risultare meno coinvolgenti, anche se appropriate.

2.2.4 Groundedness

- **Obiettivo**: Verificare se la risposta fornita è supportata da fatti o informazioni accurate rispetto al contesto del dialogo.
- **Calcolo**: Il punteggio di groundedness è stato valutato confrontando la risposta con le informazioni fornite nel contesto o in una fonte esterna.
- **Limiti**: In mancanza di contesti dettagliati o fatti espliciti, il punteggio potrebbe essere influenzato negativamente.

2.3 Processo di Valutazione

L'implementazione del processo di valutazione è stata effettuata utilizzando il framework G-EVAL e un dataset annotato. I passaggi principali sono i seguenti:

1. **Caricamento del dataset**: I dati sono stati caricati da un file JSON contenente il contesto del dialogo, la risposta generata e le annotazioni fornite.

2. **Generazione dei prompt**: È stato utilizzato un template per generare il prompt specifico per ciascun caso, distinguendo tra valutazioni *turn-level* e *dialog-level*.

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

3. **Valutazione tramite G-EVAL**: Ogni prompt è stato inviato al modello G-EVAL per ottenere i punteggi delle metriche.

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

4. **Parsing dei risultati**: I punteggi di Naturalness, Coherence, Engagingness e Groundedness sono stati estratti dall'output del modello.

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 i punteggi generati automaticamente (*evaluation mean*) e quelli forniti dagli annotatori umani (*overall score*). Inoltre, sono state analizzate le differenze tra questi punteggi per identificare discrepanze significative.

2.5.1 Calcolo delle Correlazioni

Le correlazioni di Pearson, Spearman e Kendall-Tau sono state utilizzate per confrontare i punteggi. Queste metriche permettono di analizzare la relazione lineare, la concordanza dei ranghi e la monotonicità tra i dati.

- **Pearson**: Misura la correlazione lineare tra *evaluation mean* e *overall score*.
- **Spearman**: Analizza la concordanza nei ranghi, utile per dati non lineari.
- **Kendall-Tau**: Valuta la concordanza tra i ranghi, fornendo robustezza in caso di dati discreti.

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 punteggi relativi a quattro metriche principali: **Naturalness**, **Coherence**, **Engagingness** e **Groundedness**. Il framework si basa su una classe Python, `GEvalAPI`, progettata per:

- Caricare template di prompt specifici per ogni modalità di 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`}}, {{`fact`}}) con i dati specifici. Ad esempio:

```
def generate_prompt(self, template, context, response, fact=None):
    prompt = template.replace("{{context}}", context)
                     .replace("{{response}}", response)
    if fact:
        prompt = prompt.replace("{{fact}}", fact)
    return prompt
```

3.1.3 Integrazione con il Modello GPT-4

Il prompt generato viene inviato al modello GPT-4 per ottenere i punteggi. 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 di specifici aspetti del dialogo, 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 dialogue context, a system response, and a related fact.

Your task is to evaluate the response using the following criteria: Naturalness, Coherence, Engagingness, and Groundedness. Use the provided fact to evaluate the Groundedness of the response.

Please respond **ONLY** with the scores in the format provided below. Do not include any additional text, comments, or explanations. If the response cannot be evaluated, assign all scores as `0`.

### Evaluation Criteria:

1. **Naturalness (1-5)**: Is the response human-like, natural, and conversational in tone?
2. **Coherence (1-5)**: Does the response logically follow the dialogue context, maintaining the flow of the conversation?
3. **Engagingness (1-5)**: Is the response interesting and engaging for the user? Does it encourage continued interaction?
4. **Groundedness (1-5)**: Does the response provide accurate and factual information that aligns with the dialogue context and the provided fact? Does it avoid unsupported statements?

### Evaluation Steps:

1. Carefully read the dialogue context to understand its tone, key points, and overall structure.
2. Review the provided fact and use it to verify the accuracy of the response.
3. Analyze the response to determine if it aligns with the dialogue
```

```
    context and maintains engagement based on the evaluation criteria.
4. Assign a score for each criterion on a scale of 1 to 5, where:
   - ‘1‘ indicates the lowest quality,
   - ‘5‘ indicates the highest quality.

### Evaluation Form (scores ONLY):

- Naturalness:
- Coherence:
- Engagingness:
- Groundedness:

Dialogue Context:
{{context}}

Fact:
{{fact}}

Response:
{{response}}
```

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 delle valutazioni, in particolare per:

- **Dialoghi completi**: Mantiene la coerenza tra i turni di dialogo.
- **Singole risposte**: Analizza dettagliatamente il contesto per valutazioni più precise.
- **Fatti forniti**: Supporta la verifica dell'aderenza al contesto.

3.4 Conclusione

L'integrazione di prompt ben progettati e l'uso dell'approccio *Chain-of-Thought* hanno reso il framework flessibile e accurato. La modularità della classe `GEvalAPI` consente di

estendere facilmente il sistema a nuovi scenari o metriche, garantendo una valutazione affidabile dei modelli di *Natural Language Generation*.

4 I Dataset

4.1 Introduzione

Per valutare le prestazioni del framework G-EVAL, sono stati utilizzati due dataset principali: **FED (Feedback Evaluation Dataset)** e **Topical-Chat**. Questi dataset forniscono dialoghi annotati con valutazioni umane, essenziali per analizzare la qualità delle risposte generate dai modelli.

4.2 Descrizione dei Dataset

4.2.1 FED (Feedback Evaluation Dataset)

FED è un dataset progettato per valutare modelli di dialogo open-domain. Contiene dialoghi annotati con punteggi umani relativi a metriche come Coherence, Fluency, Consistency e Relevance.

- **Origine**: Creato per task di valutazione automatica.
- **Formato**: Ogni esempio è un file JSON con:
  - `context`: Il contesto del dialogo.
  - `response`: La risposta generata dal modello.
  - `annotations`: Valutazioni umane (es. Overall score).
  - `system`: Il sistema che ha generato la risposta.

4.2.2 Topical-Chat

Topical-Chat è un dataset focalizzato su dialoghi che includono conoscenze fornite. Ogni dialogo è annotato con valutazioni umane relative a metriche come naturalness e engagingness.

- **Origine**: Raccolto da dialoghi umani annotati.

- **Formato**: Ogni dialogo include:
  - `context`: Lista di messaggi.
  - `response`: Risposta del sistema.
  - `annotations`: Valutazioni delle metriche.

4.3 Preprocessing

I dataset sono stati preprocessati per rimuovere dati incompleti e normalizzare i dialoghi. In particolare, i dialoghi sono stati divisi in due categorie:

- **Turn-level**: Ogni risposta è stata analizzata individualmente.
- **Dialog-level**: L'intero dialogo è stato valutato come unità.

5 Risultati

5.1 Valutazioni per FED

Il dataset **FED** è stato utilizzato per analizzare la qualità delle metriche di valutazione applicate ai dialoghi. Le valutazioni sono state effettuate confrontando i punteggi generati automaticamente (*evaluation mean*) con quelli forniti dagli annotatori umani (*overall score*). I risultati sono stati analizzati tramite correlazioni e visualizzazioni grafiche.

5.1.1 Risultati delle Correlazioni

I risultati delle correlazioni tra *evaluation mean* e *overall score* sono riportati nella Tabella 1.

Table 1: Correlazioni tra *Evaluation Mean* e *Overall Score* per il dataset FED.

Metrica	**Valore**
Pearson	0.2241
Spearman	0.2148
Kendall-Tau	0.1584

I valori mostrano una correlazione debole tra i punteggi generati dal sistema e quelli umani, suggerendo che le metriche attuali non riflettono pienamente i giudizi umani.

5.1.2 Analisi Grafica

Per analizzare i risultati, sono state generate le seguenti visualizzazioni grafiche:

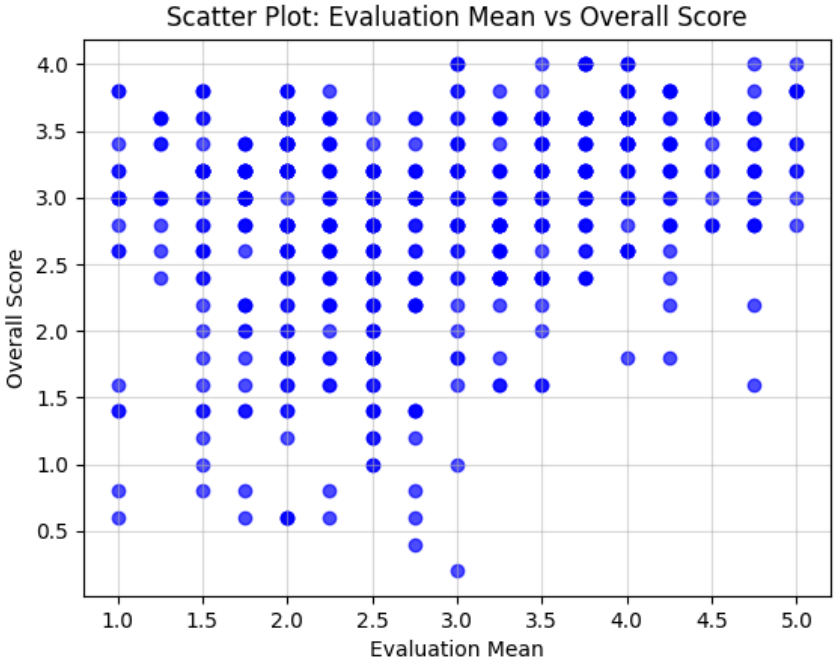

Figure 1: Scatter Plot: Relazione tra *Evaluation Mean* e *Overall Score* per il dataset FED.

5.1.3 Discussione dei Risultati

I risultati mostrano che le metriche di valutazione utilizzate hanno una correlazione modesta con i giudizi umani:

- La correlazione di Pearson (0.2241) indica una relazione lineare debole.
- Le correlazioni basate sui ranghi (Spearman 0.2148 e Kendall-Tau 0.1584) confermano una bassa concordanza con i punteggi umani.
- Le differenze tra *Evaluation Mean* e *Overall Score* (Figura 2) evidenziano discrepanze che richiedono ulteriori miglioramenti.

I risultati suggeriscono che l'integrazione di metriche più sofisticate o personalizzate potrebbe migliorare la capacità del sistema di catturare le valutazioni umane in modo più accurato.

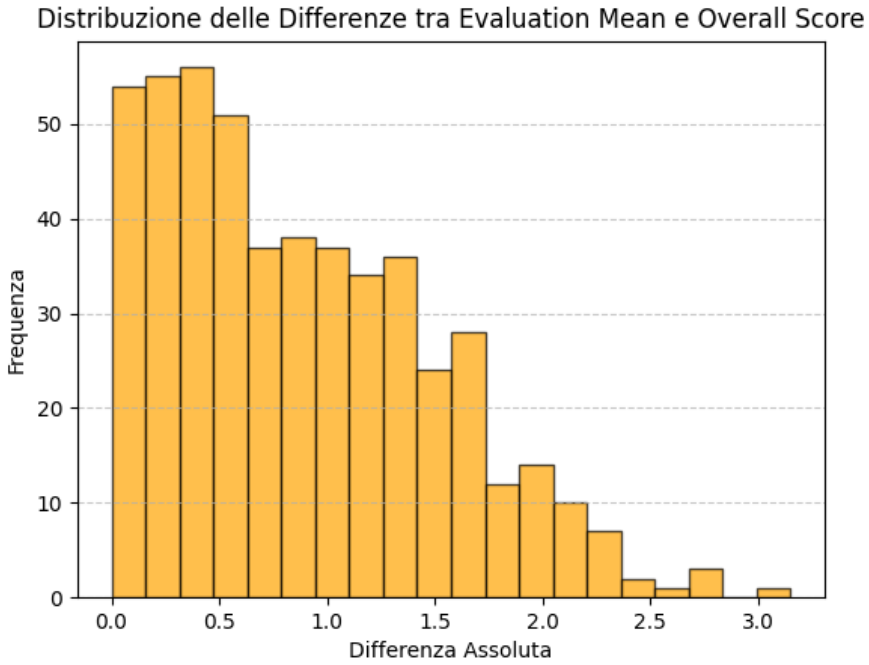

Figure 2: Istogramma delle differenze assolute tra *Evaluation Mean* e *Overall Score* per il dataset FED.

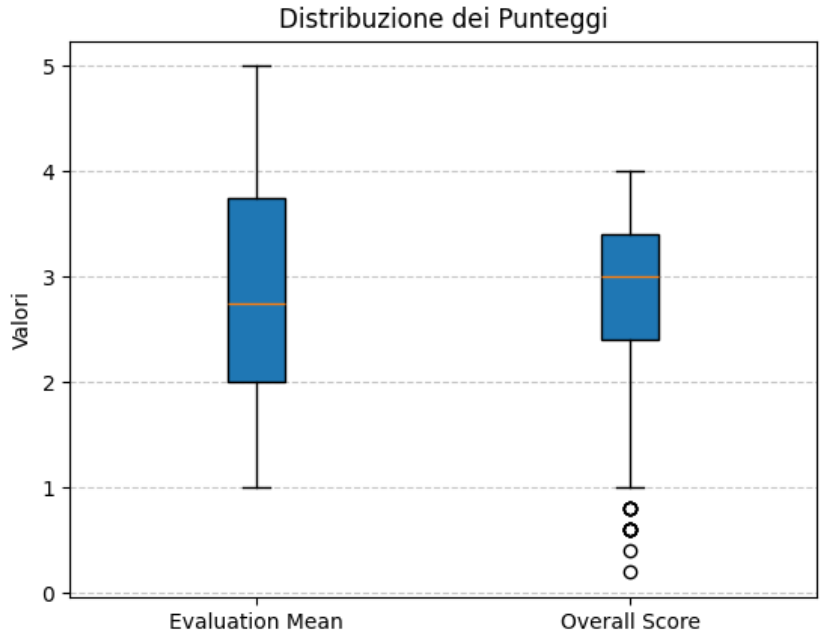

Figure 3: Distribuzione dei punteggi *Evaluation Mean* e *Overall Score* per il dataset FED.

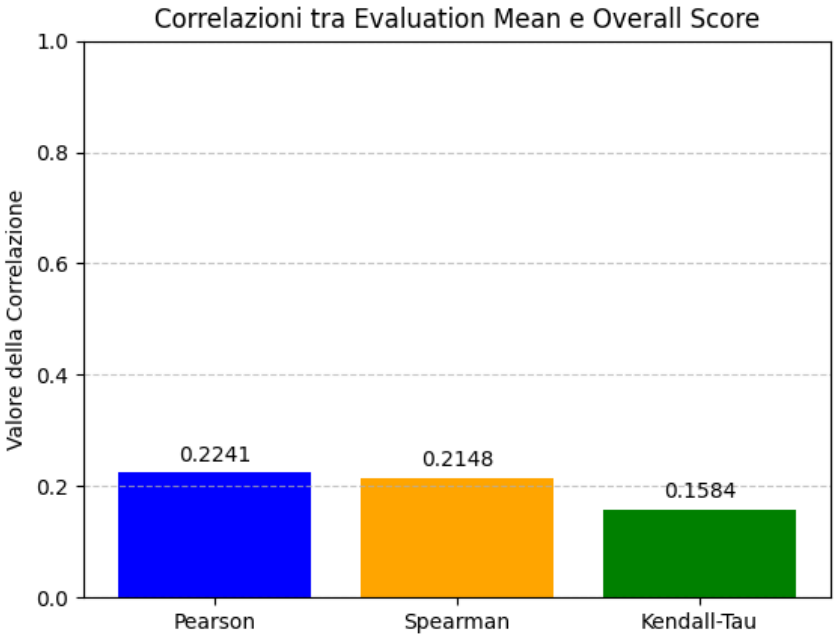

Figure 4: Bar Plot delle correlazioni (Pearson, Spearman, Kendall-Tau) per il dataset FED.

5.2 Valutazioni per TC_USR

Il dataset **TC_USR** è stato utilizzato per valutare la capacità delle metriche di rappresentare accuratamente i giudizi umani in un contesto di dialoghi. I risultati delle analisi sono stati calcolati confrontando i punteggi generati automaticamente (*evaluation mean*) con quelli annotati manualmente (*overall score*).

5.2.1 Risultati delle Correlazioni

Le correlazioni tra *evaluation mean* e *overall score* per il dataset TC_USR sono riportate nella Tabella 2.

Table 2: Correlazioni tra *Evaluation Mean* e *Overall Score* per il dataset TC_USR.

Metrica	**Valore**
Pearson	0.3721
Spearman	0.3547
Kendall-Tau	0.2691

I risultati mostrano una correlazione leggermente superiore rispetto al dataset FED,

ma ancora moderata, indicando che le metriche possono essere ulteriormente migliorate.

5.2.2 Analisi Grafica

Le seguenti visualizzazioni grafiche rappresentano i risultati ottenuti con il dataset TC_USR:

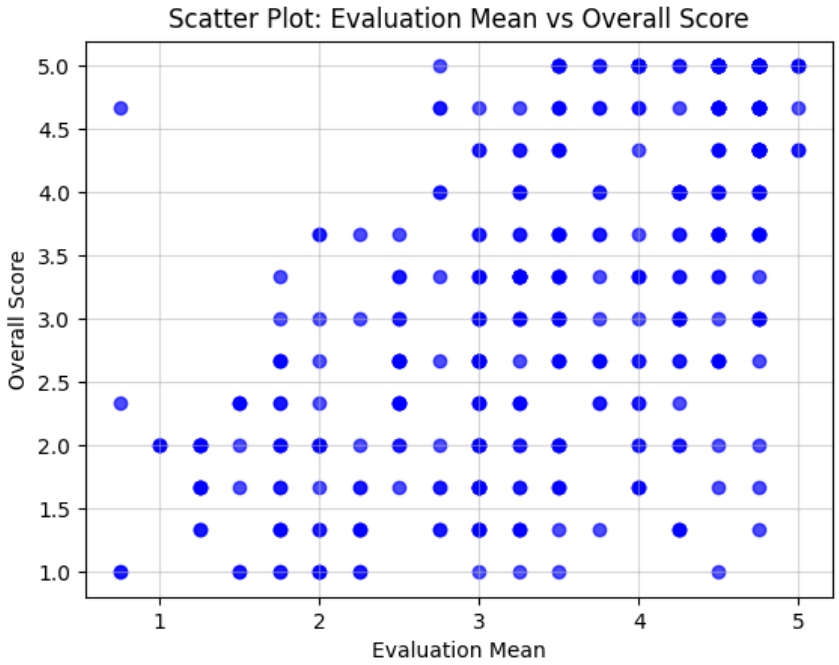

Figure 5: Scatter Plot: Relazione tra *Evaluation Mean* e *Overall Score* per il dataset TC_USR.

5.2.3 Discussione dei Risultati

I risultati delle analisi condotte sul dataset TC_USR indicano una correlazione più forte rispetto al dataset FED:

- La correlazione di Pearson (0.3721) evidenzia una relazione lineare più marcata.
- Le correlazioni basate sui ranghi (Spearman 0.3547 e Kendall-Tau 0.2691) mostrano una maggiore concordanza nei punteggi.
- Le differenze tra *evaluation mean* e *overall score* (Figura 6) rivelano una maggiore distribuzione, ma con discrepanze ridotte rispetto a FED.

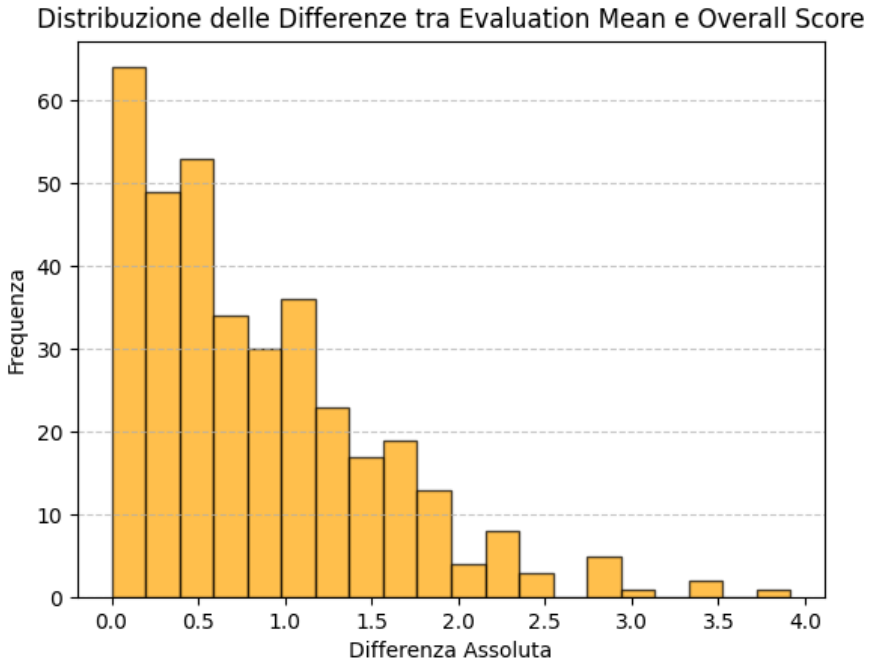

Figure 6: Istogramma delle differenze assolute tra *Evaluation Mean* e *Overall Score* per il dataset TC_USR.

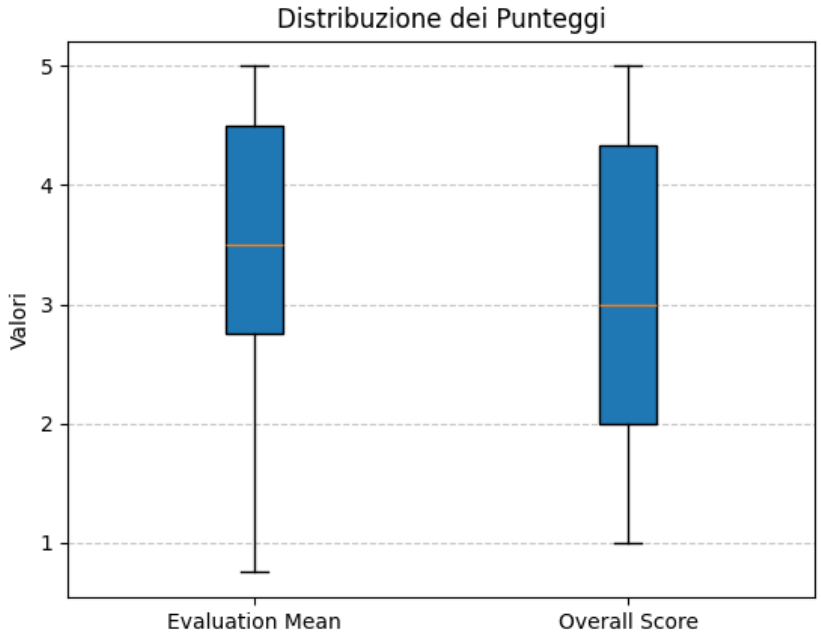

Figure 7: Distribuzione dei punteggi *Evaluation Mean* e *Overall Score* per il dataset TC_USR.

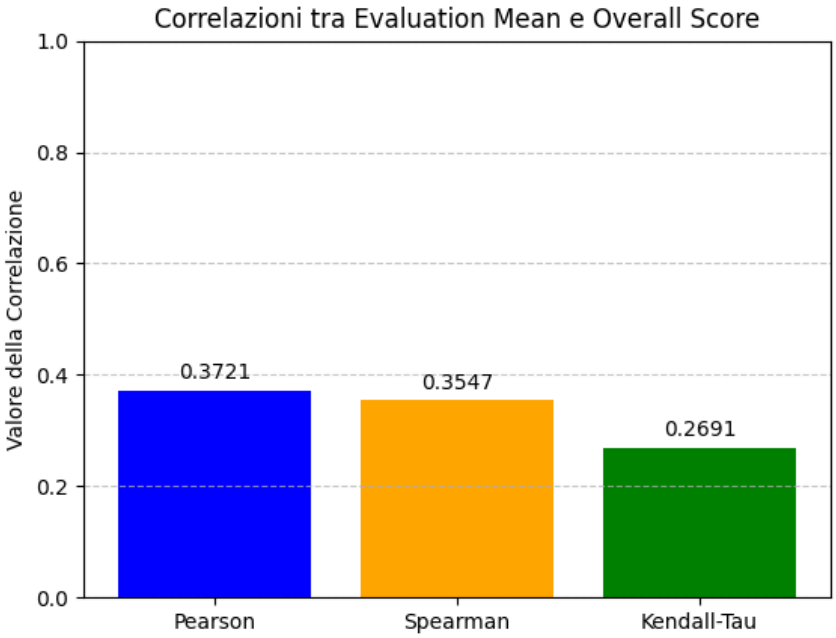

Figure 8: Bar Plot delle correlazioni (Pearson, Spearman, Kendall-Tau) per il dataset TC_USR.

Questi risultati suggeriscono che, sebbene il sistema di valutazione riesca a catturare meglio alcuni aspetti dei giudizi umani rispetto a FED, è necessaria un'ulteriore ottimizzazione per ridurre ulteriormente le discrepanze e migliorare la correlazione.

5.3 Prospettive di miglioramento

- Introduzione di modelli di valutazione basati su reti neurali per migliorare la comprensione semantica.
- Calibrazione delle metriche di valutazione per un maggiore allineamento con i giudizi umani.
- Integrazione di ulteriori dataset con annotazioni diversificate per ampliare la generalizzabilità del sistema.