G-EVAL-Dialogue

Cavaliere Mattia, Citro Carmine, Nunziata Vincenzo

January 2025

Indice

1 Introduzione

Il presente progetto vuole rappresentare un approfondimento dell'articolo "G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment" [Liu+23], il quale presenta appunto G-EVAL, un nuovo algoritmo di valutazione per i Large Language Model (LLM).

1.1 Contesto del progetto

La valutazione della Natural Language Generation è un task importante ma anche difficile. Il task risulta particolarmente ostico in quanto la valutazione di un NLG non può essere costruita in maniera oggettiva basandosi solamente su regole di sintassi del linguaggio, ma deve tenere conto di correlazioni sintattiche e logiche di un discorso.
Per poter valutare un testo generato automaticamente può essere necessario tenere in considerazione alcuni parametri di valutazione, chiamati **criteri di valutazione**.
I criteri di valutazione dipendono dal task e dalla tipologia di testo che si intende valutare.

Ad esempio, nel caso della valutazione di un riassunto generato automaticamente potrebbe essere utile valutare i seguenti criteri:

- Coerenza
- Consistenza
- Scioltezza
- Rilevanza

Nel contesto della valutazione di un dialogo tra un utente umano e un sistema di generazione automatica, invece, potrebbero essere rilevanti criteri come:

- Naturalezza
- Coerenza
- Coinvolgimento
- Concretezza

Esistono diversi sistemi di valutazione delle NLG, come ad esempio BLEU, ROGUE, USR, UniEval, GPTScore, ecc. Tuttavia, buona parte di questi valutatori ha riportato una scarsa correlazione con il giudizio umano, rendendo così poco attendibili le loro valutazioni, specialmente in un contesto in cui si vuole verificare la human-likeness di un NLG in test aperti come la generazione creativa.
La soluzione proposta da G-EVAL [Liu+23] è quella di utilizzare gli LLM durante il processo di valutazione in modo da analizzare meglio il contesto del testo generato e aumentare la correlazione delle valutazioni generate con quelle assegnate da valutatori umani.

1.2 Obiettivi principali

L'obiettivo di questo progetto è quello di approfondire il funzionamento di G-EVAL come framework di valutazione di NLG, in quanto si ritiene che sistemi di valutazione che sfruttano gli LLM per i task di valutazione possano offrire una maggiore correlazione con il giudizio umano. La valutazione della NLG è un task di fondamentale importanza per la ricerca e sviluppo dei sistemi generativi, in quanto oltre a fornire una metrica che permette di controllare errori (sia di sintassi che di semantica) all'interno di un testo generato, permette anche di distinguere testi qualitativi da testi che presentano errori o incoerenze.
Nello specifico, il presente progetto si inserisce nell'ambito della valutazione di dialoghi tra un utente umano e un sistema NLG.
Si vuole dimostrare la flessibilità di G-EVAL come valutatore di NLG rispetto ad altri sistemi di valutazione, in quanto non sfrutta metriche o meccanismi di valutazione strettamente legati ad un singolo task, bensì è possibile utilizzare lo stesso sistema di valutazione per task diversi applicando piccole variazioni alla struttura del codice e modificando il prompt in input all'LLM.

2 Panoramica di G-EVAL

G-EVAL è un framework che sfrutta gli LLM e la Chain-of-Thoughts (CoT) generata automaticamente per valutare NLG tramite il paradigma del riempimento di form. Tra le tecniche esplorate da G-EVAL è presente anche l'assegnazione di un peso a ogni score basato sulla probabilità del punteggio di essere generato in output dall'LLM, al fine di migliorare e raffinare le metriche finali (Spearman, Kendall-Tau, Pearson, ecc.).

2.1 Descrizione del framework G-EVAL

G-EVAL è un valutatore di NLG prompt-based composto da tre componenti principali:

1. Prompt: un comando scritto in linguaggio naturale che contiene la definizione del task di valutazione e i criteri di valutazione desiderati da dare in input all'LLM
2. CoT: un insieme di istruzioni intermedie generato dall'LLM che descrive dettagliatamente gli step da seguire per la valutazione di NLG
3. Score function: una funzione di score che richiama l'LLM per calcolare lo score in base ai criteri di interesse

I criteri di valutazione inseriti nel prompt possono essere diversi in base al task di valutazione che si vuole eseguire. Gli autori di G-EVAL [Liu+23] hanno proposto l'utilizzo di criteri differenti:

- Task di valutazione di riassunti:
  - Coerenza
  - Consistenza
  - Scioltezza
  - Rilevanza
- Task di valutazione di dialoghi:
  - Naturalezza
  - Coerenza
  - Coinvolgimento
  - Concretezza

Il presente progetto si concentra sulla valutazione di dialoghi, pertanto, si è scelto di utilizzare i criteri proposti all'interno dell'articolo [Liu+23], in modo da poter confrontare in modo diretto i risultati ottenuti.

La CoT è una sequenza di step intermedi generata dall'LLM per fornire maggiore contesto e dettaglio al modello stesso al fine di migliorare la comprensione del task di valutazione. La progettazione manuale di questi step di valutazione può richiedere molto tempo. L'idea è quindi, appunto, di far generare questi step all'LLM stesso.

La funzione di score è un metodo matematico utilizzato per assegnare un punteggio alla qualità di un testo generato da un sistema di NLG. Questa funzione combina probabilità assegnate ai punteggi discreti generati dal modello per produrre una valutazione continua e più accurata.
I valori di score s_i vengono generati dall'LLM come richiesto dal prompt dato in input e sulla base di uno o più criteri di valutazione. Il punteggio s è un valore intero tale che: $1 \leq s \leq 5$. In pratica, la funzione di score si occupa quindi di effettuare la chiamata al modello fornendo in input il prompt, la CoT auto generata e il dialogo da valutare, per poi combinare il punteggio ottenuto con la probabilità in modo da generare il valore finale.

2.2 Differenze rispetto ai metodi tradizionali di valutazione

G-EVAl presenta differenze sotto diversi aspetti con le metriche di valutazione tradizionale (come ad esempio BLEU, ROUGE, METEOR ecc.). Tra le principali differenze abbiamo:

- Capacità di catturare creatività e diversità: le metriche tradizionali si concentrano sul calcolo della sovrapposizione lessicale o semantica (n-grams), perdendo informazioni sulla creatività o sulla coerenza globale del testo, mentre G-EVAL utilizza una struttura di CoT che fornisce valutazioni più dettagliate e spiega meglio le ragioni delle valutazioni, migliorando la capacità di catturare aspetti qualitativi come creatività, coerenza e rilevanza

- Allineamento con i giudizi umani: le metriche tradizionali mostrano spesso bassa correlazione con i giudizi umani, specialmente in task aperti come la generazione di dialoghi o testi creativi. G-EVAL dimostra una correlazione molto più alta con i giudizi umani, superando metriche esistenti grazie ad una combinazione di CoT e ponderazione delle probabilità

- Utilizzo dei modelli linguistici: le metriche tradizionali si basano su metodi statici o embedding pre-addestrati. G-EVAL integra direttamente i modelli linguistici generativi non solo per generare valutazioni, ma anche per creare autonomamente criteri e passaggi di valutazione (CoT)

- Flessibilità per nuovi task: le metriche tradizionali richiedono criteri predefiniti o personalizzati per ogni nuovo task. G-EVAL può essere addestrato a nuovi scenari grazie alla sua capacità di generare dinamicamente criteri di valutazione basati sul contesto specifico

3 Setup del progetto

Come anticipato, il presente progetto vuole porre il proprio focus sulla valutazione di dialoghi tra un utente umano e un sistema autonomo (ad esempio un chatbot). L'implementazione creata a tale scopo si basa sulla struttura del framework G-EVAL fornita dagli autori, reperibile sul loro repository GitHub.
All'interno dell'articolo di G-EVAL [Liu+23] sono riportati i risultati su due diversi task di valutazione: valutazione di riassunti e valutazione di dialoghi. Nonostante vengano riportati i risultati di entrambi i task, all'interno dell'implementazione fornita non è stata gestita la possibilità di cambiare il task solamente tramite il prompt in input. È stato necessario, pertanto, implementare da zero la gestione del task di valutazione di dialoghi. Ciò è stato fatto ristrutturando la classe `gpt_4_eval.py` la quale conteneva gli script necessari per l'esecuzione del valutatore.
Si è poi riscontrata la mancanza, dal punto di vista pratico, della gestione degli score pesati e che è stata quindi implementata per poter mantenere una delle caratteristiche chiave di G-EVAL.

3.1 Modello LLM utilizzato: Llama 3 8B Instruct

Come riportato dall'articolo di G-EVAL [Liu+23], i test eseguiti sono stati condotti tramite l'utilizzo di due modelli: GPT-3.5 e GPT-4. Entrambi i modelli, però, possono essere utilizzati tramite delle API che richiedono l'utilizzo di token a pagamento. Per questo motivo, nell'ambito di questo progetto, si è scelto un modello utilizzabile in maniera gratuita tramite l'ambiente GPT4All: Llama 3 8B.
Llama 3 è un LLM sviluppato da Meta AI, progettato per comprendere e generare testo in modo simile a un essere umano. È stato rilasciato nel 2024 ed è disponibile in diverse configurazioni, tra cui modelli con 8 miliardi e 70 miliardi di parametri. Tra i principali casi d'uso di Llama 3 sono presenti:

- Chatbot: automatizzazione del servizio clienti con risposte naturali e contestuali
- Creazione di contenuti: generazione di articoli, report, blog e storie
- Comunicazione via email: assistenza nella stesura e formulazione di email coerenti con il tono del brand
- Analisi dei dati: sintesi di documenti complessi e generazione di rapporti dettagliati

Come intuibile dal nome, il modello scelto per il presente progetto è la versione ad 8 miliardi di parametri. Si è optato per la versione ad 8 miliardi di parametri in quanto era quella che offriva il giusto bilanciamento tra la complessità del modello e le risorse necessarie per eseguirlo localmente. Di fatto, tramite l'ambiente GPT4All, i modelli caricati vengono eseguiti localmente, pertanto la scelta del modello da utilizzare è stata dettata anche dalle risorse a disposizione.

3.2 Dataset dei dialoghi

Per poter condurre al meglio gli esperimenti sull'implementazione proposta, si è scelto di utilizzare due dataset diversi per il benchmark, ognuno con un preciso scopo. I dataset in questione sono: DSTC9 e Topical-Chat.
Topical-chat è il dataset utilizzato dagli autori dell'articolo di G-EVAL, ed è stato scelto in modo da poter ottenere dei risultati comparabili con quelli riportati nell'articolo [Liu+23]. In questo modo si possono confrontare le prestazioni ottenute da Llama 3 8B con quelle di GPT-3.5 e GPT-4.
DSTC9, invece, è stato selezionato come dataset di benchmark per verificare le prestazioni di G-EVAL su un dataset completamente diverso.

3.2.1 DSTC9

Il dataset utilizzato come benchmark è DSTC9, ovvero un dataset di dialoghi human-to-chatbot messo a disposizione per la competizione Dialog System Technology Challenge (DSTC9) [Gun+20]. Nella sua versione originale il dataset si presenta come un file JSON che segue la seguente struttura:

```
{
    "contexts":[
        [
        "<dialogue_1_line_1>,
        <dialogue_1_line_2>,
        <...>"
        ]
        [
        "<dialogue_2_line_1>,
        <dialogue_2_line_2>,
        <...>"
        ]
        <[...]>
    ],
    "responses":[
        "<response_1>",
        "<response_2",
        "<...>"
    ],
    "references":[
        "<reference_1>",
        "<reference_2>",
        "<...>"
    ],
    "scores":[
        <score_1>,
        <score_2>,
        <...>
    ],
    "models":[
        "<model_1>",
        "<model_2>,
        "<...>"
    ]
}
```

In questa struttura, ogni dialogo è rappresentato da un array di stringhe contenente le frasi scambiate tra l'utente umano e il modello di LLM. Ad ogni dialogo corrisponde una risposta fornita dal sistema, una reference, un punteggio e il modello che ha generato la risposta. Data la struttura degli oggetti, al primo dialogo corrisponde la prima risposta, la prima reference, il primo punteggio e il primo modello, al secondo dialogo corrisponde la seconda risposta, la seconda reference, il secondo punteggio e il secondo modello, e così via.
Il dataset presentava alcune problematiche. Il parametro `reference`, il quale doveva ipoteticamente contenere una risposta al dialogo fornita da un umano, era "`NOREF`" per tutti i dialoghi, non fornendo così nessuna informazione aggiun-

tiva. Tuttavia, questo parametro potrebbe risultare utile solamente nel task di valutazione della risposta fornita da un processo di NLG, mentre per la valutazione del dialogo nella sua interezza, la reference è poco rilevante.
Per i task di valutazione viene richiesto, comunemente, di assegnare un punteggio intero (ad esempio tra 1 e 3 o tra 1 e 5). Il punteggio riportato per ogni dialogo (`score`), sarebbe dovuto essere quindi un valore intero assegnato da un valutatore umano. Tuttavia, tutti i punteggi sono rappresentati con valori decimali (ad esempio 4.333), quindi sono continui. Ciò ha portato alla conclusione che questo punteggio assegnato ad ogni dialogo rappresenti un punteggio sul criterio di overall quality, assegnato tramite qualche trasformazione (ad esempio una media aritmetica) su punteggi interi assegnati da un valutatore umano.
Il parametro `model` assegnato ad ogni dialogo rappresenta il modello al quale è stato fornito in input il dialogo e che ha generato la risposta. Tuttavia, nella forma originale del dataset, tutti i modelli erano stati identificati con la dicitura "`chatbot_n.json`", dove n è un numero progressivo da 1 a 9. Pertanto non è stato possibile risalire a quale modello avesse generato le singole risposte.
Alla luce delle evidenze riscontrate e della struttura originale del dataset è stato necessario applicare una ristrutturazione, cercando di ottenere una forma di più semplice utilizzo, che cercasse di mantenere solamente le informazioni rilevanti per la valutazione. La forma ottenuta è la seguente:

```
[
    {
        "dialog_id": "<id>",
        "turns": [
            {
                "speaker": "<user|system>",
                "utterance": "<utterance>"
            },
            {
                "speaker": "<user|system>",
                "utterance": "<utterance>"
            },
            <{...}>
        ],
        "score": <score>,
        "system_id": "<id>"
    },
    <{...}>
]
```

In questa forma la risposta ottenuta dal sistema è stata concatenata al dialogo originale in modo da poter eseguire una valutazione sull'intero dialogo. Ogni dialogo è stato suddiviso in turni. Ogni turno è rappresentato dall'interlocutore e dalla frase. È stato, inoltre, aggiunto un identificativo progressivo ai dialoghi.

3.2.2 Topical-Chat

Topical-Chat è un dataset composto da oltre 10.000 conversazioni. Per poter confrontare direttamente le prestazioni di G-EVAL utilizzato con un LLM diverso, si è scelto di effettuare dei test anche su questo dataset. Tuttavia, la versione di Topical-Chat utilizzata dgli autori di G-EVAL è quella Topical-Chat-USR proposta da [ME20], composta da 60 dialoghi scelti casualmente da Topical-Chat. Ognuno di questi dialoghi è stato sottoposto a 6 modelli diversi i quali hanno fornito una risposta al dialogo, ottenendo così 360 item totali. Il dataset nella sua forma originale è strutturato nel seguente modo:

```
[
    {
        "source": "<dialogue>",
        "system_id": "<id>",
        "system_output": "<response>",
        "context": "<dialogue context>",
        "scores":{
            "understandability": <score>,
            "naturalness": <score>,
            "coherence": <score>,
            "engagingness": <score>,
            "groundedness": <score>,
            "overall": <score>
        }
    }
    <{...}>
]
```

Per adattare il dataset all'implementazione progettata per il benchmark DSTC9 si è scelto di effettuare una ristrutturazione degli oggetti contenenti i dialoghi in modo da poter riutilizzare gli script già progettati. Il dataset ristrutturato ha la seguente forma:

```
[
    {
        "dialogue": "<id>",
        "turns":[
            {
                "speaker": "<A|B>",
                "utterance": "<utterance>"
            }
            <{...}>
        ],
        "score": <score>,
        "system_id": "<id>",
        "system_output": "<response>",
        "context": "<dialogue context>"
```

```
        },
        <{...}>
]
```

Si è scelto di tenere in considerazione solamente lo score di "overall" in modo da poter confrontare direttamente i risultati con quelli ottenuti sul benchmark DSTC9. Inoltre, per far si che i risultati possano essere confrontati con quelli ottenuti dall'articolo di presentazione di G-EVAL, si è scelto di eseguire i test su questo dataset considerando il task di valutazione della risposta ottenuta dal sistema, piuttosto che la valutazione sull'intero dialogo.

3.3 Criteri di valutazione

Tramite il framework G-EVAL è possibile scegliere dinamicamente i criteri di valutazione in base al task da eseguire, modificando solamente il prompt dato in input al LLM. All'interno dell'articolo di presentazione di G-EVAL vengono mostrati i risultati ottenuti da quattro prompt diversi, uno per ognuna delle seguenti metriche:

- Naturalezza
- Coerenza
- Coinvolgimento
- Concretezza

Viene fornita, inoltre, una metrica complessiva chiamata "Overall quality", calcolata come la media aritmetica dei criteri di valutazione.
Siccome il dataset DSTC9 fornisce solamente i punteggi di overall per ogni dialogo, nel contesto di questo progetto si è scelto di progettare i prompt per far si che l'LLM restituisca direttamente il valore di overall, fornendone una definizione dettagliata all'interno del prompt.
Per quanto riguarda il dataset Topical-Chat, esso fornisce i punteggi relativi alle seguenti metriche:

- Comprensibilità
- Naturalezza
- Coerenza
- Coinvolgimento
- Fondatezza
- Overall

Allo stesso modo di DSTC9, i test sono stati condotti con prompt progettati per richiedere all'LLM di fornire il punteggio di overall. Tuttavia, in questo caso, specificando la dimensione (ovvero il criterio desiderato) che si vuole analizzare al tempo di esecuzione, è possibile effettuare test anche sugli altri criteri.

4 Implementazione

La struttura di partenza per l'implementazione del presente progetto è stata reperita sul repository ufficiale dell'articolo di G-EVAL [Liu+23]. Analizzando l'implementazione proposta dagli autori si sono riscontrate diverse problematiche, tra cui:

- Mancanza dell'utilizzo delle probabilità dei token per l'assegnazione di scores pesati
- Mancanza di un modulo per la generazione automatica della CoT

È stato necessario, quindi, riprogettare e aggiungere degli script che permettessero la gestione di queste due componenti chiave di G-EVAL per adattare l'implementazione al contesto del problema.
L'implementazione originale sfrutta le API offerte da OpanAI per l'interrogazione di modelli quali GPT-3.5 e GPT-4. L'ambiente GPT4All offre piena compatibilità con le API ufficiali di OpenAI, pertanto non è stato necessario riprogettare le chiamate al modello per l'interrogazione ma è bastato solamente modificare il modello che si vuole interrogare e fornire l'API endpoint ottenuto una volta eseguito il server di GPT4All.

4.1 Generazione della Chain-of-Thought (CoT)

Una delle caratteristiche principali di G-EVAL è la generazione automatica della CoT. All'interno dell'articolo di G-EVAL [Liu+23] non è specificato in che modo sia stata generata la CoT utilizzata per i test e all'interno del codice sorgente non sono presenti scripts o moduli per gestire questa operazione.
È stato necessario, quindi, implementare un modulo apposito che prima di eseguire G-EVAL interroghi l'LLM tramite un prompt appositamente progettato, in modo da ottenere una CoT generata automaticamente. Una volta generata la CoT, questa viene inserita all'interno del prompt utilizzato per eseguire il task di valutazione e verrà riutilizzata per ogni dialogo sottoposto in esame al modello.
Affinché la CoT venisse generata correttamente è stato necessario fornire, oltre alle informazioni di base sul task da eseguire, indicazioni sulla formattazione della risposta per evitare che il modello restituisse in output un testo che utilizzasse formattazioni particolari o elementi di markdown, i quali avrebbero portato l'LLM ad interpretare incorrettamente il prompt. A seguito di diversi test, il prompt finale utilizzato per la generazione della CoT nel caso del task di valutazione di dialoghi è:

> *Evaluate the quality of a human-machine dialogue based on the following criteria:*
>
> *Criteria:*
>
> *- Overall Quality (1-5): The dialogue should be coherent, engaging, and contextually relevant.*

Your Task:

- Write only the step-by-step evaluation process to assess the overall quality of the entire dialogue. Limit the steps to a maximum of 4.
- Ensure the steps are clear, simple, and strictly focused on the criterion.
- Do not use Markdown or any formatting elements, including headings, or numbered lists.
- Avoid introductions, conclusions, or additional considerations, avoid markdown elements.

Output:

Il prompt utilizzato, invece per la generazione della CoT nel caso del task di valutazione della risposta è:

Evaluate the quality of the next response given a dialogue between two indiviuals, based on the following criteria:

Criteria:

- Overall Quality (1-5): The response should be coherent, engaging, and contextually relevant.

Your Task:

- Write only the step-by-step evaluation process to assess the overall quality of the response. Limit the steps to a maximum of 4.
- Ensure the steps are clear, simple, and strictly focused on the criterion.
- Do not use Markdown or any formatting elements, including headings, or numbered lists.
- Avoid introductions, conclusions, or additional considerations, avoid markdown elements.

Output:

4.2 Prompt design e criteri di valutazione

Per progettare il prompt si è tenuto in considerazione l'esempio riportato nell'articolo di G-EVAL [Liu+23]. Come anticipato, gli esperimenti sono stati condotti utilizzando l'overall quality come criterio di valutazione. Per questo motivo, è stata inserita all'interno del prompt una definizione dettagliata di tale criterio. Successivamente, la descrizione del task è stata modificata per indicare all'LLM di valutare l'intero dialogo anziché limitarsi alla risposta fornita dal sistema autonomo.
Ogni prompt è quindi composto da una sezione che introduce il task da eseguire, una specifica dei criteri di valutazione e una descrizione dettagliata del task. L'ultima parte di ogni prompt è strutturata in modo da ottenere una

risposta da parte del modello in maniera simile a quanto accade per la compilazione dei form.
Il prompt utilizzato per l'esecuzione di G-EVAL viene poi arricchito con l'inserimento della CoT auto generata dall'LLM.
Siccome si è scelto di eseguire due task leggermente diversi tra loro (valutazione del dialogo e valutazione della risposta ad un dialogo) è stato necessario creare due prompt diversi. I prompt sono molto simili, tuttavia presentano qualche piccola differenza nella specifica del task da eseguire.
Il prompt utilizzato per il task di valutazione di dialoghi è il seguente:

> *You will be given a conversation between a human user and an automatic system. The dialogue consists of alternating turns, one per line, with each line starting with the speaker.*
>
> *Your task is to rate the overall quality of the dialogue, considering both human inputs and system responses. Provide only the final score as an integer—do not include summaries or conclusions.*
>
> *Evaluation Criteria:*
>
> *- Overall Quality (1-5): The dialogue should be coherent, engaging, and contextually relevant.*
> *- Logical Flow: Turns should connect naturally without contradictions or abrupt shifts.*
> *- Context Awareness: Responses must be relevant and appropriately continue the discussion.*
> *- Engagement and Naturalness: The dialogue should feel fluid, dynamic, and free from robotic or repetitive phrasing.*
> *- Credibility and Informativeness: Responses should provide meaningful and well-formed contributions.*
>
> *Evaluation Steps:*
>
> {{*Steps*}}
>
> *Conversation:*
>
> {{*Dialogue*}}
>
> *Evaluation Form (scores ONLY):*
>
> *Overall Quality:*

Durante l'esecuzione di G-EAVL le keywords indicate con {{`Steps`}} e {{`Dialogue`}} vengono sostituiti rispettivamente con la CoT auto generata dal modello e il dialogo da sottoporre in esame all'LLM.
Il prompt utilizzato, invece, per il task di valutazione della singola risposta è:

> *You will be given a conversation between two individuals. The dialogue consists of alternating turns, one per line, with each line starting with the speaker. You will then be given one potential response for the next turn in the conversation.*

Your task is to rate the response on one metric. Provide only the final score as an integer—do not include summaries or conclusions.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

- Overall Quality (1-5): The response should be coherent, engaging, and contextually relevant.
- Logical Flow: Response should connect naturally without contradictions or abrupt shifts.
- Context Awareness: Response must be relevant and appropriately continue the discussion.
- Naturalness: The response should feel fluid and free from robotic or repetitive phrasing.
- Credibility and Informativeness: Response should provide meaningful and well-formed contributions.

Evaluation Steps:

{{*Steps*} *Conversation:*

{{*Dialogue*}}

Response:

{{*System output*}}

Evaluation Form (scores ONLY):

- Overall Quality (1-5):

4.3 Valutazione tramite il framework G-EVAL

La valutazione tramite G-EVAL è un processo relativamente semplice. Una volta ottenuto il prompt finale, compreso di CoT auto generata e di dialogo da valutare, il prompt viene semplicemente sottoposto al LLM il quale si occuperà di analizzare il dialogo valutandolo in base ai criteri descritti e restituirà il punteggio di ciascun criterio.
Nel caso specifico di questa implementazione, è stato richiesto al modello di generare 20 risposte per ogni dialogo, in modo da poter calcolare la probabilità di ciascun punteggio e generare così il punteggio finale pesato rispetto alle probabilità dei singoli punteggi.
I punteggi generati vengono poi assegnati al dialogo corrispondente all'interno del dataset sotto forma di array, in modo da poter essere utilizzati successivamente per il calcolo delle metriche di meta-valutazione.

4.4 Metriche per la meta-valutazione

Le metriche di meta-valutazione sono delle metriche che permettono di valutare un modello di valutazione, in questo caso G-EVAL, misurando la correlazione tra i valori predetti dal valutatore e quelli assegnati da un umano.
All'interno dell'articolo di G-EVAL [Liu+23], sono stati riportati i risultati in termini di due metriche:

- Spearman
  - misura la correlazione tra due variabili ordinali basandosi sulla relazione monotona tra esse
  - Confronta i ranghi delle valutazioni invece dei valori assoluti
  - Il coefficiente varia tra -1 (correlazione negativa perfetta) e 1 (correlazione positiva perfetta), con 0 che indica nessuna correlazione
- Kenadall-Tau
  - Valuta la concordanza tra due insiemei di ranghi basandosi sul numero di coppie ordinate concordanti e discordanti
  - È meno sensibile ai legami nei dati rispetto a Spearman
  - Il suo valore varia tra -1 e 1, con interpretazione simile a Spearman

Più alti sono i punteggi ottenuti su queste metriche e maggiore è la correlazione tra il valutatore e il giudizio umano.
Affinché queste metriche possano essere calcolate, è necessario avere due insiemi di dati da comparare.
Siccome originariamente G-EVAL è stato testato su Topical-Chat, le metriche venivano calcolate singolarmente su ogni criterio di valutazione. Questo era possibile in quanto ogni dialogo di Topical-Chat è stato sottoposto a 6 modelli differenti per ottenere una risposta. Ciò implica che una volta eseguito G-EVAL, ad ogni dialogo corrisponderà un insieme di valori di dimensione 6 della metrica selezionata, generati dal modello, e un insieme di valori di dimensione 6 della metrica selezionata, assegnati da un valutatore umano. Tuttavia, essendo DSTC9 sprovvisto di questo livello di dettaglio, è stato necessario utilizzare un approccio differente.
Si è scelto di utilizzare come insiemi per il calcolo delle metriche tutti i valori di overall di tutti i dialoghi.
Per quanto riguarda il dataset Topical-Chat, sotto il punto di vista teorico sarebbe possibile replicare fedelmente i test eseguiti dagli autori di G-EVAL. Tuttavia si è scelto di mantenere lo stesso approccio utilizzato per DSTC9 affinché fosse possibile confrontare in maniera diretta i risultati.

4.5 Approccio per il calcolo delle probabilità pesate

Come riportato nell'articolo di G-EVAL [Liu+23], alcuni modelli, ad esempio GPT-3.5 sono in grado di fornire direttamente in output la probabilità del token

generato di essere restituito. Siccome GPT-4 non offre questa possibilità, per condurre i test, gli autori hanno scelto di calcolare le probabilità del punteggio di essere fornito in output, facendo predire al modello 20 valori di score per ogni dialogo o documento. In questo modo è possibile approssimare il calcolo della probabilità di un token.
L'LLM Llama 3 dovrebbe essere in grado di fornire le probabilità in output. Tuttavia, come anticipato, il modello è stato caricato tramite l'ambiente GPT4All il quale presenta dei problemi di compatibilità su alcuni parametri (come in questo caso il parametro `logprob`) con le API fornite da OpenAI. Pertanto, nonostante in linea teorica Llama 3 dovrebbe essere in grado di restituire la probabilità, al momento dell'implementazione questa opzione non è ancora utilizzabile a meno che non si utilizzi il modello direttamente all'interno del codice sorgente. Si è scelto quindi di procedere seguendo l'approccio descritto dagli autori dell'articolo di G-EVAL.
Tramite il parametro `n` messo a disposizione dalle API di OpenAI si è richiesto all'LLM di generare 20 score per ogni dialogo.
Sia, quindi, l'insieme degli score possibili $\{1, 2, 3, 4, 5\}$ e n il numero di valori generati per ogni dialogo (in questo caso $n = 20$), la probabilità di ciascun valore dello score è calcolata come la sua frequenza nell'insieme di valori predetti diviso n:

$$P(s_i) = \frac{occorrenze(s_i)}{n}$$

Una volta calcolate le probabilità di ogni score per il dialogo corrente, si è calcolata la media pesata nel seguente modo:

$$weight\_score = \sum_{i=1}^{n} p(s_i) \times s_i$$

5 Esperimenti

Come anticipato, gli esperimenti sul dataset DSTC9 sono stati eseguiti sul task di valutazione dell'intero dialogo e sul task di valutazione della risposta fornita dal sistema. È stato necessario condurre entrambi gli esperimenti in quanto l'idea iniziale del presente progetto era quella di valutare l'intero dialogo; tuttavia, i test condotti dagli autori di G-EVAL sono stati effettuati sulla risposta al dialogo fornita dai modelli. Di conseguenza, i risultati ottenuti dalla valutazione dei dialoghi potrebbero essere troppo diversi. Si è voluto quindi mantenere entrambe le tipologie di esperimento per avere un confronto più dettagliato.

5.1 Descrizione degli esperimenti condotti

In primo luogo gli esperimenti si possono suddividere in due categorie, in base al tipo di task di valutazione:

- Valutazione dell'intero dialogo

- Valutazione della risposta fornita dal sistema

Come anticipato, il dataset di benchmark è DSTC9, mentre Topical-Chat è stato utilizzato al solo scopo di comprovare il corretto funzionamento del modello Llama 3 8B, nonostante la sua dimensione ridotta rispetto a modelli come GPT-3.5 e GPT-4. Tutti i test sul dataset DSTC9 sono stati eseguiti sull'intero dataset (2200 dialoghi). I risultati ottenuti con l'esecuzione di G-EVAL sono poi stati valutati tramite la meta-valutazione, calcolando la correlazione in base alle metriche di Pearson (r), Spearman (ρ) e Kendal-Tau (τ).
Per eseguire gli esperimenti sono stati utilizzati i prompt riportati nella sezione 4.1 e 4.2.

5.2 Parametri tecnici (temperatura, numero di campionamenti)

Per la scelta dei parametri tecnici si è scelto di seguire l'approccio proposto all'interno dell'articolo di G-EVAL, pertanto sono stati configurati i seguenti parametri tramite la chiamata alle API di OpenAi:

- `temperature`: 1. Indica il livello di randomicità delle risposte ottenute dall'LLM (un valore più alto porta ad una maggiore variabilità)
- `top_p`: 1. Indica il fattore di campionamento del nucleo (un valore più basso porta ad un output più prevedibile)
- `frequency penality`: 0. Fattore di panlità per le ripetizioni delle parole
- `presence penality`: 0. Incoraggia il modello ad utilizzare una maggiore varietà di parole per evitare di ripetere uno stesso token più volte

6 Risultati

Di seguito vengono mostrati i risultati ottenuti dagli esperimenti in base al dataset utilizzato e al task di valutazione richiesto.

ID esperimento	Dataset	Task
E1	DSTC9	Valutazione intero dialogo
E2	DSTC9	Valutazione risposta al dialogo
E3	Topical-Chat	Valutazione risposta al dialogo

Per riferirsi ai risultati dei test condotti dagli autori di G-EVAL si è usata la seguente legenda:

ID esperimento	Modello	Dataset	Task
R1	GPT-3.5	Topical-Chat	Valutazione risposta al dialogo
R2	GPT-4	Topical-Chat	Valutazione risposta al dialogo

I valori di Spearman e Kendal-Tau riportati all'interno dell'articolo sono:

	ρ	τ
R1	0.574	0.585
R2	0.575	0.588

I risultati ottenuti, invece, dagli esperimenti condotti nell'ambito del presente progetto sono:

	r	ρ	τ
E1	0.1885	0.1683	0.1212
E2	0.1927	0.1913	0.1392
E3	0.4900	0.4574	0.3804

6.1 Analisi delle performance

Si può notare che i risultati ottenuti sono estremamente diversi. Inizialmente si era ipotizzato un punteggio più basso per i task di valutazione sul dataset DSTC9 a causa di diversi fattori, come ad esempio l'utilizzo di un modello più piccolo e meno complesso quale Llama 3 8B. Nonostante ciò, i risultati ottenuti sono inferiori ben oltre le aspettative. Si tratta, infatti, di valori nettamente inferiori anche ai punteggi di altre metriche quali ROUGE-L, BLEU-L, METEOR, BERTScore, USR e UniEval, riportati all'interno dell'articolo [Liu+23] Per confutare l'ipotesi che il modello fosse troppo poco complesso affinché potesse prevedere dei valori corretti e più vicini ai punteggi assegnati dagli umani, si è eseguito l'esperimento sul dataset Topical-Chat. L'idea è che, ipotizzando il modello sia troppo poco complesso, le performance dovrebbero essere basse anche nel caso dell'esperimento su Topical-Chat, in quanto il modello non dovrebbe essere in grado di prevedere valori plausibili. Tuttavia, è stato possibile notare che in realtà i punteggi ρ e τ ottenuti dall'esperimento E3 sono molto vicini ai valori di riferimento (R1 e R2), il che ha portato alla conclusione che i motivi di punteggi così bassi sugli esperimenti E1 ed E2 sono da ricercare altrove.

6.1.1 Correlazioni con giudizi umani

Le metriche di Pearson, Spearman e Kendall-Tau sono utilizzate per misurare la correlazione tra le valutazioni ottenute dall'LLM e il giudizio umano, rappresentato dai punteggi assegnati da un valutatore umano ai singoli dialoghi.Per garantire l'affidabilità dei risultati, il punteggio di overall, utilizzato per calcolare la correlazione tra i valori ottenuti e il giudizio umano, dovrebbe essere calcolato seguendo gli stessi criteri per entrambi i punteggi (sia quello assegnato dal valutatore umano sia quello assegnato dall'LLM). Tuttavia non è questo il caso, in quanto non è stato possibile stabilire i criteri sui quali si basa il punteggio riportato originariamente all'interno del dataset DSTC9. Inoltre, come descritto precedentemente, nel caso del presente progetto è stato richiesto al modello di valutare il criterio di overall, non delle metriche specifiche. La combinazione di questi due fattori potrebbe aver portato ad ottenere dei valori troppo discordanti, abbassando così i punteggi delle metriche di meta-valutazione.

6.1.2 Approssimazioni

Un altro fattore che potrebbe aver portato a ottenere punteggi estremamente bassi è l'approssimazione effettuata sui criteri di valutazione. Difatti, come riportato nell'articolo di G-EVAL [Liu+23], gli esperimenti sono stati eseguiti sui singoli criteri di valutazione, aspetto reso possibile dalla presenza, all'interno di Topical-Chat, dei punteggi assegnati da un valutatore umano per ognuno dei criteri scelti.
Nel caso del presente progetto, tale livello di granularità non è stato possibile in quanto i dati a nostra disposizione non erano così dettagliati. È stato necessario, quindi, richiedere all'LLM una valutazione sul solo criterio di overall, che è il più generico e, forse, il più difficile da interpretare per un LLM. Di fatto, nei risultati riportati dagli autori di G-EVAL, l'overall era calcolato come media aritmetica dei punteggi ottenuti per gli altri criteri, rendendo quindi l'overall un valore rappresentativo dell'aggregazione di altre metriche.

6.1.3 Dimensioni dei dataset

Un altro fattore da tenere in considerazione come possibile causa dei risultati ottenuti è la dimensione dei dataset utilizzati. Come specificato nella sezione 3.2.2, gli autori di G-EVAL hanno utilizzato un dataset di dimensioni inferiori a DSTC9. Durante lo sviluppo del presente progetto sono stati effettuati dei test per verificare il corretto funzionamento degli script e dei prompt progettati. Per facilitare questa verifica, i test sono stati condotti su un sottoinsieme di DSTC9 composto dalle prime 100 istanze. Si è notato che, in questi test, i punteggi ottenuti in fase di meta-valutazione fossero nettamente superiori rispetto a quelli ottenuti negli esperimenti sul dataset completo (E1 e E2). Si è quindi constatato che gli esperimenti condotti su dataset di dimensioni ridotte tendono a produrre risultati migliori. Ciò potrebbe dipendere dalla maggiore dispersione delle valutazioni fatte dal LLM. Infatti, eseguendo gli esperimenti su un numero maggiore di dialoghi, si ottengono risultati più sparsi e, di conseguenza, i punteggi delle metriche di meta-valutazione tendono a essere più bassi.
Alla luce dei risultati ottenuti, si può concludere che l'implementazione presentata in questo progetto è valida, poiché i punteggi delle metriche dell'esperimento E3 sono molto simili a quelli riportati dagli autori di G-EVAL. Potrebbe quindi essere utile condurre esperimenti sull'intero dataset di Topical-Chat per verificare l'impatto di dati più sparsi sulle prestazioni di G-EVAL.

7 Conclusioni e lavori futuri

Osservando i risultati ottenuti tramite gli esperimenti si è potuto concludere che G-EVAL, in termini di valutatore NLG, può rappresentare una soluzione alternativa alle metriche più comuni. Tuttavia, G-EVAl implica l'utilizzo di LLM i quali rappresentano il vero limite di questo framework. Le performance di G-EVAL sono strettamente correlate al modello utilizzato, oltre che al prompt progettato per il task di valutazione. Tuttavia, mentre il prompt può essere

migliorato e strutturato diversamente in base alle capacità di comprensione dell'LLM, il modello alla base di questo metodo di valutazione può inficiare fortemente sulle prestazioni. Modelli più piccoli e poco complessi, come ad esempio Llama 3 8B, possono portare G-EVAL ad ottenere punteggi nettamente inferiori in termini di meta-valutazione.
Tuttavia, gli esperimenti condotti aprono le porte a diversi altri test che è possibile effettuare. Un possibile sviluppo futuro, data la disponibilità di risorse adeguate, potrebbe prevedere l'utilizzo di un LLM più grande, come ad esempio Llama 3 40B. In questo modo si potrebbe verificare se, con maggiore complessità, i risultati ottenuti tramite G-EVAL sono più uniformi.
Una seconda ipotesi è sicuramente l'implementazione del task utilizzando criteri di valutazione più granulari come riportati nella sezione 3.3. In questo modo si potrebbe esprimere il punteggio di overall come un valore direttamente correlato a metriche più specifiche, ottenendo così un risultato più preciso.
Ancora, si potrebbero ripetere gli esperimenti su altri dataset o su versioni più ampie (nel caso di Topical-Chat) o più dettagliate (nel caso di DSTC9). Lo sviluppo degli esperimenti su due soli dataset può portare a risultati limitati in quanto non permette di valutare a pieno la versatilità del framework.

References

[Gun+20] Chulaka Gunasekara et al. *Overview of the Ninth Dialog System Technology Challenge: DSTC9*. 2020. arXiv: 2011.06486 [cs.CL]. URL: https://arxiv.org/abs/2011.06486.

[ME20] Shikib Mehri and Maxine Eskenazi. *USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation*. 2020. arXiv: 2005.00456 [cs.CL]. URL: https://arxiv.org/abs/2005.00456.

[Liu+23] Yang Liu et al. *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. 2023. arXiv: 2303.16634 [cs.CL]. URL: https://arxiv.org/abs/2303.16634.