PROGETTO DI INTELLIGENZA ARTIFICIALE

G-EVAL-Dialogue:
Un nuovo approccio alla valutazione dei modelli NLG

Docenti:

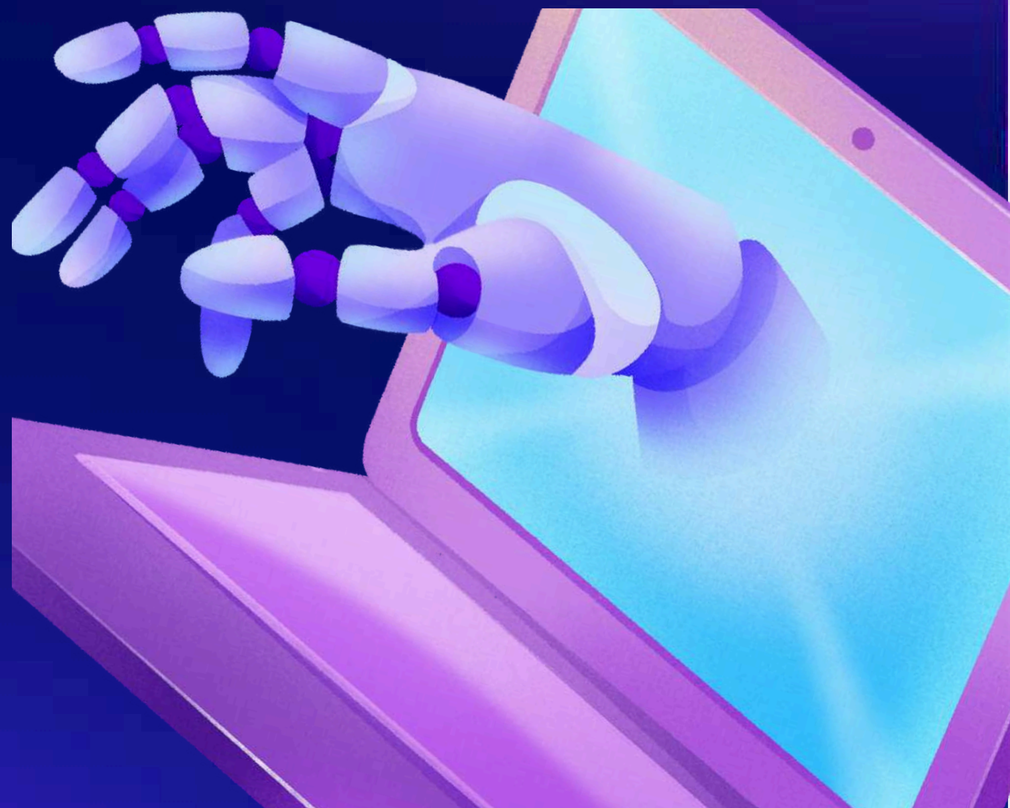
Prof. Vincenzo Deufemia
Dott. Gaetano Cimino

Studenti:
Cavaliere Mattia
Citro Carmine
Nunziata Vincenzo

Obiettivo del progetto

Implementazione del framework G-EVAL per il task di valutazione di dialoghi:

- Comprendere la metrica di valutazione G-EVAL
- Sviluppo di un'implementazione del framework di G-EVAL per la valutazione di dialoghi
- Dimostrare la flessibilità del framework
- Valutazione dei risultati ottenuti da G-EVAL tramite l'utilizzo di metriche di meta-valutazione (Spearman, Pearson e Kendall's-Tau)



Panoramica di G-EVAL

Caratteristiche di G-EVAL:

- Framework basato su LLM per la valutazione della NLG.
- Usa la tecnica Chain-of-Thought (CoT) per generare valutazioni contestualizzate.
- Pondera i punteggi basandosi sulle probabilità generate dall'LLM.

Differenze con i metodi tradizionali:

- Ha una maggiore correlazione con il giudizio umano.
- È adattabile a diversi task senza dover ridefinire metriche specifiche.

Setup del Progetto

LLM utilizzato:

- Llama 3 8B Instruct (open-source, eseguito localmente con GPT4All).

Dataset usati per la valutazione:

- DSTC9: dataset di dialoghi human-to-chatbot.
- Topical-Chat: dataset usato in G-EVAL per confronto

Motivazioni della scelta:

- Modello open-source senza costi di utilizzo.
- Equilibrio tra performance e risorse computazionali.

Criteri di Valutazione

Task principale: valutare la qualità del dialogo usando G-EVAL.

Metriche usate:

- Overall Quality :
  - Naturalezza
  - Coerenza
  - Coinvolgimento
  - Concretezza

Implementazione

Generazione della Chain-of-Thought (CoT)

- LLM genera step intermedi per migliorare la valutazione.
- Prompt specifico per guidare il modello a produrre CoT chiara.

Prompt Design:

- Formattazione rigorosa per evitare risposte non strutturate
- Prompt differenziati per:
  - Valutazione del dialogo
  - Valutazione della risposta al dialogo

Prompt per la Valutazione del Dialogo

Prompt usato per valutare il dialogo completo

```
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 & Naturalness: The dialogue should feel fluid, dynamic, and free from robotic or
repetitive phrasing.
    - Credibility & Informativeness: Responses should provide meaningful and well-formed contributions.

Evaluation Steps:
{{Steps}}

Conversation:
{{Dialogue}}

Evaluation Form (scores ONLY):

    Overall Quality:
```

Esempio di output atteso:

```
Ovarall: [4, 3, ..., 5]
```

Prompt per la Valutazione della Risposta

Prompt specifico per la valutazione della risposta generata dal sistema

```
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 & 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):
```

Metriche di Meta-Valutazione

Obiettivo: calcolo della correlazione di G-EVAL con il giudizio umano

Metriche usate:

- Spearman's Rank Correlation (ρ)
- Kendall's Tau (τ)
- Pearson's Correlation (r)

Formula per la media pesata dei punteggi:

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

$$\text{Weighted Score} = \sum_{i=1}^{n} P(s_i) \times s_i$$

Esperimenti

Esperimenti condotti:

- **E1:** Valutazione di un dialogo sul dataset DSTC9
- **E2:** Valutazione della risposta ad un dialogo sul dataset DSTC9
- **E3:** Valutazione della risposta ad un dialogo sul dataset Topical-Chat

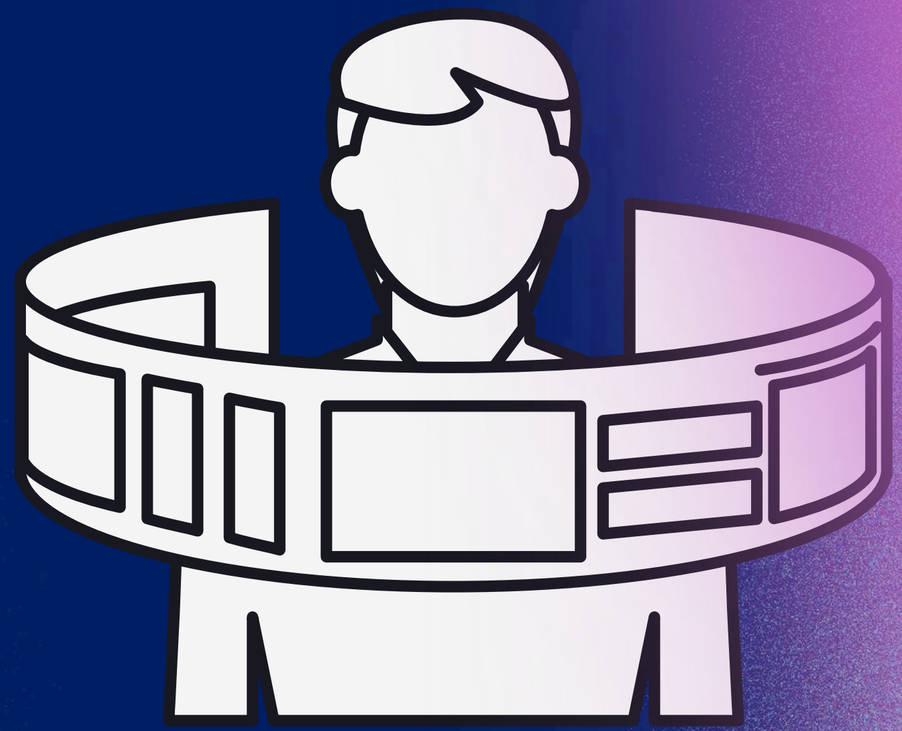
Parametri tecnici:

- **temperatura** = 1
- **top_p** = 1
- **frequency penality** = 0
- **presence penality** = 0
- **n** = 20

Risultati

Risultati principali:

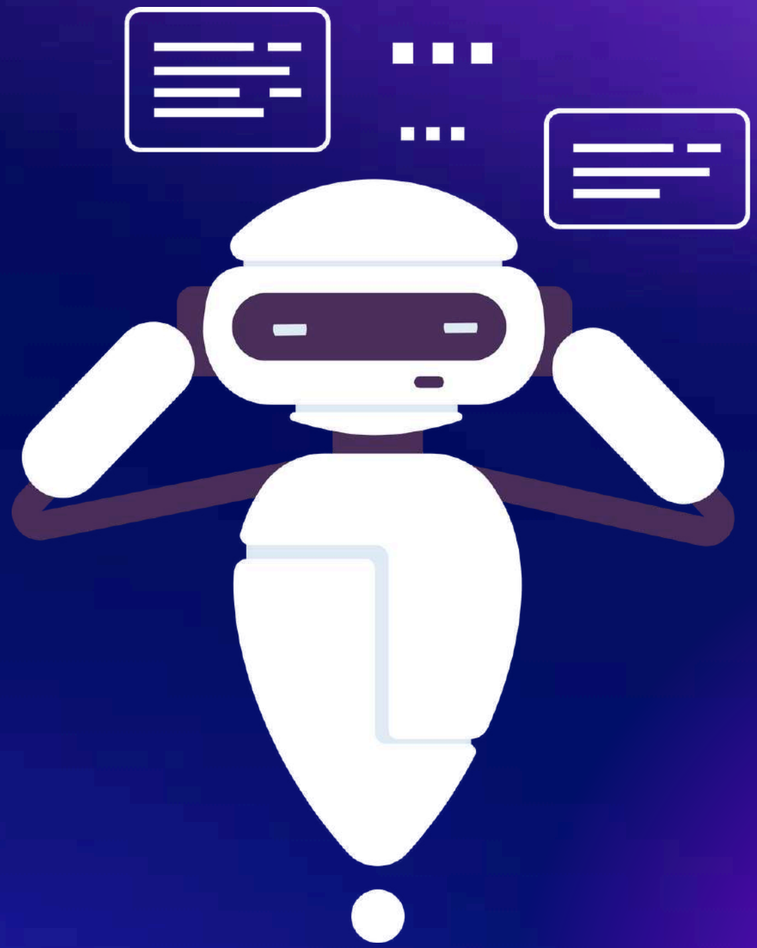
- **DSTC9** (Valutazione intero dialogo):
  - **Spearman's:** $\rho = 0.1683$
  - **Kendall's:** $\tau = 0.1212$
  - **Pearson:** $r=0.1885$
- **DSTC9** (Valutazione risposta):
  - **Spearman's:** $\rho = 0.1913$
  - **Kendall's:** $\tau = 13.92$
  - **Pearson:** $r=0.1927$
- **Topical-Chat:**
  - **Spearman's:** $\rho = 0.4574$
  - **Kendall's:** $\tau = 0.3804$
  - **Pearson:** $r = 0.4900$
- **Confronto con risultati originali:**
  - **GPT-3.5:** $\rho = 0.574$, $\tau = 0.585$
  - **GPT-4:** $\rho = 0.575$, $\tau = 0.588$



Analisi delle Performance

Fattori che hanno influenzato i risultati:

- Modello meno potente rispetto a GPT-3.5 e GPT-4.
- Approssimazione dell'Overall Quality rispetto ai criteri singoli.
- Dimensione ridotta del dataset rispetto agli esperimenti originali.



Conclusioni e Lavori Futuri

Conclusioni:

- G-EVAL è un framework promettente per la valutazione della NLG.
- Le performance dipendono fortemente dal modello LLM utilizzato.

Prossimi passi:

- Usare un LLM più grande (es. Llama 3 40B).
- Testare criteri di valutazione più dettagliati.
- Applicare G-EVAL ad altri dataset.

Grazie mille per l’attenzione