Titolo del Progetto: Implementazione della Metrica G-EVAL per la Valutazione di Sistemi NLG

Obiettivi:

- Comprendere la metrica G-EVAL proposta nell'articolo "G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment".
- Implementare il framework G-EVAL.
- Valutare l'efficacia di G-EVAL nella misurazione della qualità di output generati da sistemi di Natural Language Generation (NLG) sul task di generazione di dialoghi.

Metodologia di Implementazione:

1. **Analisi della Metrica G-EVAL**:
    - Studiare l'articolo per comprendere i componenti principali di G-EVAL: Prompt per la valutazione, Chain-of-Thoughts (CoT), e funzione di scoring basata su probabilità.
    - Valutare le performance della metrica su dataset di benchmark che vi verranno forniti.
2. **Implementazione del Framework**:
    - **Prompt Design**: Creare prompt per il task di valutazione, includendo i criteri personalizzati (ad esempio, coerenza, fluidità, rilevanza).
    - **Chain-of-Thoughts**: Implementare un modulo che generi automaticamente i passaggi intermedi per l'analisi del testo.
    - **Funzione di Scoring**: Integrare la normalizzazione delle probabilità per ottenere un punteggio continuo e dettagliato.
3. **Valutazione**:
    - Selezionare output generati da sistemi NLG e i relativi giudizi umani disponibili nei benchmark.
    - Calcolare la correlazione tra i punteggi di G-EVAL e i giudizi umani utilizzando metriche statistiche come Kappa di Cohen, Spearman, Pearson e Kendall-Tau.

Risultati Attesi:

- Implementazione funzionante di G-EVAL per il task di generazione di dialoghi.
- Analisi dettagliata delle correlazioni tra G-EVAL e giudizi umani.
- Discussione delle sfide incontrate durante l'implementazione e delle potenziali migliorie per future applicazioni.

Risorse:

- **Articolo Base**: "G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment".
- **Codice Open Source**: G-EVAL su GitHub.
- **API Modelli**: https://github.com/selfsff/GPT4ALL-Free-GPT-API