Titolo del Progetto: Implementazione della Metrica LLM-EVAL per la Valutazione Automatica Multi-Dimensionale di Dialoghi Open-Domain

Obiettivi:

- Comprendere la metrica LLM-EVAL proposta nell'articolo "LLM-EVAL: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models".
- Implementare il framework LLM-EVAL per valutare dialoghi open-domain utilizzando modelli linguistici di grandi dimensioni (LLM).
- Analizzare l'efficacia di LLM-EVAL nel misurare la qualità di risposte generate rispetto a giudizi umani utilizzando dataset di benchmark che vi verranno forniti.

Metodologia di Implementazione:

1. Analisi della Metrica LLM-EVAL:

- Studiare l'articolo per comprendere i seguenti elementi fondamentali:
    - Schema di valutazione unificato per dimensioni come contenuto, grammatica, rilevanza e appropriatezza.
    - Utilizzo di un singolo prompt per coprire più dimensioni con una sola chiamata al modello.
    - Formato dei punteggi (ad esempio, scale da 0-5 o 0-100).
- Valutare le strategie di configurazione suggerite per le sperimentazioni, inclusi i metodi di decodifica (ad esempio, greedy decoding).

2. Preparazione dell'Ambiente:

- Configurare un ambiente Python con librerie essenziali come PyTorch, Transformers e NumPy.
- Integrare API di modelli linguistici, ad esempio OpenAI ChatGPT.

3. Implementazione del Framework:

Schema di Valutazione:

- Definire uno schema di valutazione che includa istruzioni per il modello su come calcolare punteggi per ciascuna dimensione (contenuto, grammatica, rilevanza, appropriatezza).
- Creare un template di prompt che concateni il contesto del dialogo, la risposta generata e, se disponibile, la risposta di riferimento.

Generazione dei Punteggi:

- Implementare il sistema per inviare i prompt al modello linguistico e ricevere in output punteggi multi-dimensionali.
- Normalizzare i punteggi per garantire coerenza tra diverse dimensioni.

4. Valutazione:

Dataset:

- Vi verranno forniti dataset di benchmark con annotazioni umane per testare la metrica.

Analisi delle Performance:

- Calcolare le correlazioni tra i punteggi prodotti da LLM-EVAL e i giudizi umani utilizzando metriche statistiche come:
    - Kappa di Cohen
    - Correlazione di Spearman
    - Correlazione di Pearson
    - Correlazione di Kendall-Tau

Risultati Attesi:

- Implementazione funzionante di LLM-EVAL per la valutazione di dialoghi open-domain.
- Analisi dettagliata delle correlazioni tra i punteggi di LLM-EVAL e i giudizi umani su diverse dimensioni.
- Discussione delle sfide incontrate durante l'implementazione e suggerimenti per miglioramenti futuri.

Risorse:

- Articolo di Riferimento: "LLM-EVAL: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models".
- **API ChatGPT**: https://github.com/selfsff/GPT4ALL-Free-GPT-API