# Linee Guida per l'Annotazione – Dataset di Paragraph-Pairs per Neutralizzazione (HIPP)

### Obiettivo del task

Valutare e correggere (se necessario) tre tipi di testi generati dal modello Qwen a partire da paragrafi originali del dataset HIPP (etichettati come hyperpartisan = 1/BIAS o neutral = 0/NO BIAS). Lo scopo finale è ottenere un dataset pulito di coppie (originale → versione trasformata) utile per addestrare modelli di neutralizzazione.

**DEFINIZIONE HYPERPARTISAN (BIAS)**
Hyperpartisan content is characterized by a strong ideological bias, explicit references to extremist political positions, or a subjective style intended to influence reader opinion. It often uses emotionally charged language, exaggerations, or one-sided arguments, Irony and/or Sarcasm. Depending on the context, the reported speech text could be used as a mean to professionally refer to explicit political leanings and/or ideologies. Please, pay attention to the reported speech, since not always journalists use them to neutrally report discourses.

All of those traits can be implicit or explicit.

For instance, entry 8: Lo stesso Tobia ha definito la guerra della Russia all'Ucraina "un'aggressione criminale". "Noi lo possiamo dire – ha aggiunto – non quelle forze politiche che si riempiono la bocca della parola pace ma poi votano l'invio di soldati e armi in Ucraina , gettando benzina sul fuoco per una possibile terza guerra mondiale".

Nonostante sia un paragrafo costituito per lo più da virgolettato, è chiara l'intenzione comunicativa del giornalista nel trasmettere un messaggio BIAS, facendo riferimento ad 1) un sopruso in maniera iperbolica; 2) ad una posizione privilegiata per cui un determinato gruppo di persone (Noi) ha la possibilità di descrivere fatti in un certo modo piuttosto che in un altro, 3) l'utilizzo di linguaggio figurato "riempirsi la bocca di" per 4) sminuire la posizione dell'opposizione facendo leva sulla sua 5) incoerenza, presentando questo comportamente attraverso una 6) metafora iperbolica (benzina sul fuoco) e 7) un appeal to fear misto a 8) iperbole (terza guerra mondiale).

### Le tre tipologie di augmentazione

Per ogni riga del dataset troverai queste colonne da valutare/correggere:

| Tipo | Colonna | Obiettivo della trasformazione | Deve mantenere il label originale? | Note importanti |
|---|---|---|---|---|
| 1 | {model}_rephrased | Parafrasi del testo originale | SÌ | Stesso tono e stesso label |
| 2 | {model}_it_perturbed | Back-translation + modifica di tratti linguistici | SÌ | Cambiare tratti (lessico carico, figure retoriche, struttura, etc.) ma mantenere significato e label |
| 3 | {model}_rewritten_text | Riscrittura con inversione di polarità (flip del label) | NO (deve invertire) | Neutralizzare se originale è BIAS → NO BIAS; rendere marcatamente hyperpartisan se originale è NO BIAS → BIAS. I fatti devono rimanere identici |

**Procedura di annotazione (da seguire rigorosamente nell'ordine)**

**Step 1 – Analisi del testo originale**

- Leggi con attenzione il testo originale
- Identifica: argomento, claim principali, fatti oggettivi, tono, eventuali tratti di linguaggio carico (caratteristici dell'hyperpartisan)
- Prendi nota del label originale (1 = BIAS / hyperpartisan; 0 = NO BIAS / neutral)

**Step 2 – Valutazione di ogni testo generato (un box alla volta)**

Leggi il testo generato **senza** guardare le predizioni automatiche di GPT-4o-mini o BERT.

**Step 3 – Valutazione del compito svolto dal modello**

Assegna una delle tre opzioni: **Yes / Partially / No**

| Valutazione | Quando assegnarla | Conseguenza |
|---|---|---|
| **Yes** | Il testo generato raggiunge perfettamente l'obiettivo del task (parafrasi corretta, perturbazione riuscita, o flip corretto) e non ha errori grammaticali/semantici significativi | Non toccare il testo a meno che non presenti lievi refusi, o lessico impreciso |

| | | |
|---|---|---|
| **Partially** | Ci sono piccoli errori (qualche errore grammaticale, lieve perdita di significato, aggiunta minima di contesto, o cambiamento parziale del tono) | Dovrai correggere tu il testo |
| **No** | Fallimento evidente: grammatica gravemente rotta, significato alterato, tono sbagliato, aggiunta di informazioni nuove, o flip non avvenuto/cattivo | Dovrai riscrivere tu il testo |

**Step 4 – Correzione del testo (solo se Partially o No)**

- Mantieni **esattamente** i fatti dell'originale
- Rispetta rigorosamente l'obiettivo del task e il label che deve avere (vedi tabella sopra)
- Non aggiungere informazioni o paragrafi nuovi
- Scrivi in italiano naturale e grammaticalmente corretto, rispettando le norme giornalistiche per quanto riguarda la professionalità
- Inserisci il testo corretto nella colonna corrispondente (sovrascrivendo quello generato)

**Step 5 – Assegnazione del tuo label personale**

Dopo aver eventualmente corretto il testo, assegna **tu** il label finale del testo (alla tua versione finale):

- Colonne "rephrased" e "it_perturbed" → deve essere **uguale** al label originale
- Colonna "rewritten_text" → deve essere **opposto** al label originale Usa: **BIAS** (1) o **NO BIAS** (0)

**IMPORTANTE: se il testo originale viene annotato in questa fase come NO BIAS, procedere alla entry successiva e comunicare la entry. Il dataset finale deve contenere solo casi Hyperpartisan-Neutral e non Neutral-Hyperpartisan. Comunque si richiede di selezionare il Best Model (vd. Step 6 di seguito)**

**Step 6 – Best Model Prediction**
Guarda le due predizioni automatiche (GPT-4o-mini e BERT fine-tuned).
Scegli quale delle due descrive **meglio il testo generato dal modello Qwen** (NON il tuo testo corretto!).

Se nessuna delle due è corretta → scegli "None".

**Casi particolari**

1. **Il testo generato aggiunge informazioni non presenti nell'originale** → Valutazione massima = **Partially** (anche se il tono è corretto). Correggi eliminando le aggiunte.
2. **Il testo originale è borderline (es. etichettato BIAS ma sembra quasi neutrale, o viceversa)** → Per il Rewritten: accetta il label originale come "vero" e forza il flip comunque. → Se il modello non è riuscito a fare un flip evidente perché l'originale

era già ambiguo → valuta “No” e correggi tu rendendo il tono chiaramente opposto. → Segnala l’entry_id nel campo note: sono casi marginali utili per analisi successive.

3. **Il modello Qwen non ha generato nulla o ha generato testo vuoto** → Valutazione = **No** → Scrivi tu il testo corretto partendo dall’originale e rispettando l’obiettivo del task.
4. **Errori di coerenza tra task** Esempio: il Rephrased ha perso il tono hyperpartisan → anche se il modello ha solo parafrasato, per te deve rimanere BIAS → correggi aggiungendo tratti di linguaggio carico compatibili con i fatti.
5. **Presenza dei placeholder dei linguistic traits nel testo generato:** In questo caso assegnare Task Accomplishment: No, Best Model: Neither
6. Il contenuto razzista, complottista, ecc va eliminato? No, va preservato. Eliminarlo significherebbe cambiare il task di neutralizzazione, considerandolo ANCHE stance. Noi vogliamo preservare il punto di vista dell’autore qualora possibile (es casi difficili: Ironia eccessiva che mascheri tutto il contenuto).

Esempio pratico:
La entry n.7: “Il presidente argentino cita “La grande bugia verde” durante l’intervista a Quarta Repubblica” è stata inizialmente annotata come BIAS. Durante la rilettura, l’annotatore converte la label originale a Neutrale, poiché il testo non presenta alcun segno di hyperpartisan. E’ vero che il testo contiene un virgolettato, ma fa riferimento al nome di un libro il cui titolo sminuisce la veridicità delle politiche green attraverso l’utilizzo della fallacia logica del dubbio. Ad esempio, se il verbo ‘citare’ fosse stato originalmente ‘osannare’ o un altro dall’omonimo significato iperbolico, in quel caso si sarebbe potuto considerare il testo BIAS. Tornando all’annotazione, come ci si deve comportare? In questo caso, molto probabilmente, Qwen avrà generato del testo NO BIAS. Avendo compiuto il task correttamente, il task accomplishment è Yes. Si prega però di notare che il testo originale e quello prodotto sarebbero dovuti essere BIAS. In questo caso è richiesto di comunicare la entry al responsabile. Per quanto riguarda il testo Rewritten, anche questo probabilmente non conterrà bias. Si prega di procedere alla prossima entry, dopo aver selezionato i best model in riferimento al testo generato. Sempre in questo caso, il Perturbed è diventato hyperpartisan: Task accomplishment: No, Best Model Both perché entrambi hanno predetto BIAS, si procede a riscrivere il testo in rifeferimento alla label riassegnata al testo originale (NO BIAS) e si seleziona la label finale (NO BIAS).

Esempi per la riscrittura di paragrafi da HP -> Neutral, ovvero parafrasare il testo modificandone esclusivamente il tono, preservandone l’epistemologia:

La **parafrasi** agisce rispetto al periodo linguistico mediante **modifiche del testo** per finalità interpretative, quindi si differenzia dalla epifrasi (parola affine) che indica **aggiunte di testo**.

Alleggerire il peso semantico degli epiteti come “notevole senatore”, “il pluri-acclamato tizio caio”.

Rimozione domande retoriche. Laddove possibile parafrasare il testo, ovvero utilizzare la terza persona e forme verbali attenuate, preferendo l’uso del si
**ESEMPI di annotazioni bordeline e di riscrittura.**