Linee Guida per la Costruzione del Dataset per il Fine-Tuning di BERT

1 Dimensione del Dataset

1.1 Training Set

- **Obiettivo minimo:** Per compiti di classificazione binaria con fine-tuning di BERT, la letteratura e le esperienze pratiche indicano che è necessario disporre di almeno qualche migliaio di esempi per classe. Nel tuo caso, dato che devi distinguere tra "jailbreak" e "non jailbreak", potresti partire con un minimo di **2.000–5.000 esempi per classe**.
- **Scalabilità:** Se le risorse e la disponibilità dei dati lo permettono, è preferibile ambire a dataset più grandi (ad esempio **5.000–10.000 esempi per classe**). L'aumento della quantità di dati può aiutare a catturare la variabilità linguistica e a ridurre il rischio di overfitting, specialmente in un dominio potenzialmente complesso e "noisy" come quello delle risposte LLM.

1.2 Test (e Dev) Set

- **Proporzione comune:** Un approccio standard è usare circa il **20–30%** del dataset totale per il test (o per sviluppare ulteriori set di validazione, ad esempio suddividendo l'80% in training, il 10% in dev e riservando il restante 10% per il test finale).
- **Esempio numerico:** Se costruisci un dataset complessivo di **10.000 esempi per classe** (cioè 20.000 esempi totali), potresti dividere il tutto in circa **16.000 esempi per l'addestramento** e **4.000 per il test/validazione** (o usare una divisione 80-10-10 se desideri avere un set di sviluppo separato).

2 Considerazioni Specifiche per il Tuo Task

- **Bilanciamento delle classi:** Assicurati che il dataset sia bilanciato per evitare che il modello impari a privilegiare una classe (tipicamente, le risposte "non jailbreak" potrebbero essere più comuni di quelle "jailbreak"). Se il fenomeno di jailbreak è raro, potresti dover ricorrere a tecniche di oversampling o alla generazione di dati sintetici per compensare.
- **Qualità e Variabilità:** La quantità di esempi è importante, ma ancor più fondamentale è la diversità degli esempi. Includi variazioni linguistiche, differenti

livelli di esplicità e diversi contesti. Nel caso dei jailbreak, il modello dovrà essere in grado di riconoscere espressioni sottili e variazioni nel linguaggio che indicano intenti illeciti.

- **Annotazioni Affidabili:** Considera l'adozione di un processo di annotazione accurato, possibilmente con revisione incrociata da parte di esperti, per minimizzare errori e ambiguità nelle etichette. Questo aspetto è cruciale per compiti tanto delicati.

- **Metriche di Valutazione:** Oltre alla semplice accuratezza, valuta il modello utilizzando metriche che forniscano informazioni sulla capacità di distinguere tra false positività e false negatività (ad esempio, precision, recall e F1-score). La scala di confidenza da 0 a 1 può essere ulteriormente calibrata tramite tecniche come il temperature scaling o altre procedure di calibrazione dei modelli.

3 Stima Finale

- **Training set:** Idealmente tra **4.000 e 20.000 esempi totali** (ossia, **2.000–10.000 per classe**), a seconda delle risorse disponibili e della complessità del linguaggio che si vuole coprire.

- **Test/Dev set:** Circa il **20–30%** dei dati totali, con una suddivisione accurata che mantenga il bilanciamento tra le classi.

Queste indicazioni costituiscono un punto di partenza ragionevole. In fase sperimentale, potresti iniziare con un dataset più piccolo per testare il flusso di lavoro e l'efficacia iniziale del modello, per poi espandere progressivamente la raccolta dei dati se i risultati non raggiungono la robustezza desiderata.

4 Conclusioni

Non esiste una regola universale valida per ogni situazione, ma la linea guida generale per un task così delicato e complesso è quella di mirare ad avere un dataset ampio e variegato, idealmente con almeno qualche migliaio di esempi per classe. In contesti di ricerca, dati compresi tra 10.000 e 20.000 esempi totali offrono una base solida per l'addestramento di un modello affidabile. È importante monitorare costantemente le metriche di performance in fase di validazione e test, adeguando la quantità e la qualità dei dati in base ai risultati ottenuti.