yaml\_config\_template = """

model\_type: llm

base\_model: {model\_path}

input\_features:

  - name: DOMANDE

    type: text

    preprocessing:

      truncation: true

      max\_sequence\_length: 210

      lowercase: true

output\_features:

  - name: RISPOSTE

    type: text

    preprocessing:

      max\_sequence\_length: 1000

      max\_length: 1000

      truncation: true

    metrics:

      - type: bert\_score

        model: dbmdz/bert-base-italian-uncased

      - type: loss

      ### - type: rouge, - type: perplexity

prompt:

  template: >-

    Sei un esperto cardiologo. Rispondi alla seguente domanda in modo empatico e con frasi complete e concise, terminando ogni frase con un punto.

    {DOMANDE}

generation:

  temperature: 0.1    ### per la generazione del testo deve essere molto basso

  max\_new\_tokens: 80

  top\_p: 0.9          ### aiuta a generare testo più coerente e fluido. , il modello avrà un equilibrio tra casualità e coerenza, garantendo che i token generati siano tra quelli con le probabilità più alte.

adapter:

  type: lora

quantization:

  bits: 4

preprocessing:

  split:

    type: random

    probabilities: [0.8, 0.1, 0.1]  ### a gestione di più sottoinsiemi può richiedere risorse aggiuntive.

  global\_max\_sequence\_length: 1210

trainer:

  type: finetune

  epochs: 3

  batch\_size: 2

  eval\_batch\_size: 2

  gradient\_accumulation\_steps: 8  # effective batch size = batch size \* gradient\_accumulation\_steps

  learning\_rate: 1.0e-4  # Prova con un learning rate leggermente più basso

  enable\_gradient\_checkpointing: true

  learning\_rate\_scheduler:

    decay: cosine

    warmup\_fraction: 0.1  # Aumenta il warmup fraction

    reduce\_on\_plateau: 0

### un checkpoint del modello verrà salvato ogni 500 step di addestramento, permettendo il ripristino del modello in caso di interruzioni.

early\_stopping:                             ### monitora la perdita di validazione (val\_loss) e interrompe l'addestramento se la perdita non migliora per 5 epoche consecutive, per prevenire l'overfitting (min: il monitoraggio è configurato per minimizzare la perdita).

  monitor: val\_loss

  patience: 5

  mode: min

logging:                                    ### i log dell'allenamento verranno inviati a TensorBoard.

  log\_to\_tensorboard: true

callbacks:

  - type: progress\_bar

  - type: model\_checkpoint

    save\_best\_only: true

    monitor: val\_loss

    mode: min

  - type: early\_stopping

    monitor: val\_loss

    patience: 5

    mode: min

"""

Immagine che contiene testo, schermata, diagramma, Policromia

Descrizione generata automaticamente

Immagine che contiene testo, linea, diagramma, Diagramma

Descrizione generata automaticamente

Immagine che contiene testo, schermata, linea, Diagramma

Descrizione generata automaticamente

Immagine che contiene testo, linea, Diagramma, schermata

Descrizione generata automaticamente

Immagine che contiene testo, linea, Diagramma, diagramma

Descrizione generata automaticamente