Testes:

**1 – Dataset FirstDataset (fusão dos Datasets 1 e 3). 80% train, 10% valid, 10% test. 4 épocas. Batch size de 16 para treino e validação. 500 warmup steps, 0.01 weight decay. MSE loss.**

TrainOutput(global\_step=2928, training\_loss=0.00017847943939065022, metrics={'train\_runtime': 2999.4931, 'train\_samples\_per\_second': 0.976, 'total\_flos': 3761778369557760.0, 'epoch': 4.0, 'init\_mem\_cpu\_alloc\_delta': 4096, 'init\_mem\_cpu\_peaked\_delta': 0, 'train\_mem\_cpu\_alloc\_delta': 302088192, 'train\_mem\_cpu\_peaked\_delta': 1661087744})

TrainOutput(global\_step=2928, training\_loss=0.0001688724602547174, metrics={'train\_runtime': 2975.2714, 'train\_samples\_per\_second': 0.984, 'total\_flos': 3761778369557760.0, 'epoch': 4.0})

{'eval\_loss': 0.026004016399383545,

'eval\_mem\_cpu\_alloc\_delta': 1880064,

'eval\_mem\_cpu\_peaked\_delta': 0,

'eval\_runtime': 729.1011,

'eval\_samples\_per\_second': 2.007,

'init\_mem\_cpu\_alloc\_delta': -4980736,

'init\_mem\_cpu\_peaked\_delta': 4980736}

metrics={'test\_loss': 0.026526669040322304, 'test\_runtime': 706.3228, 'test\_samples\_per\_second': 2.071, 'test\_mem\_cpu\_alloc\_delta': 0, 'test\_mem\_cpu\_peaked\_delta': 0})

outra run do checkpoint dos 2800:

TrainOutput(global\_step=2928, training\_loss=0.00017460240267441084, metrics={'train\_runtime': 3315.3472, 'train\_samples\_per\_second': 0.883, 'total\_flos': 3761778369557760.0, 'epoch': 4.0, 'train\_mem\_cpu\_alloc\_delta': 3455672320, 'train\_mem\_cpu\_peaked\_delta': 161865728})

{'epoch': 4.0,

'eval\_loss': 0.01280676294118166,

'eval\_mem\_cpu\_alloc\_delta': 7299072,

'eval\_mem\_cpu\_peaked\_delta': 0,

'eval\_runtime': 729.0808,

'eval\_samples\_per\_second': 2.007}