CPU型号：AMD Ryzen 7 4800U with Radeon Graphics

GPU：NVIDIA GeForce RTX 4060 Ti

ShortText：“This is a test sentence to measure inference speed.”

模型: bert-base-uncased

设备: cpu

推理次数: 100

总推理时间: 8.4021 秒

平均推理时间: 0.0840 秒/次

模型: roberta-base

设备: cpu

推理次数:100

总推理时间: 9.0764 秒

平均推理时间: 0.0908 秒/次

模型: gpt2

设备: cpu

推理次数: 100

总推理时间: 8.7640 秒

平均推理时间: 0.0876 秒/次

MidText（数据集）:["This is a short sentence.","Here is another example of a slightly longer sentence for testing.","The quick brown fox jumps over the lazy dog.","Artificial intelligence is transforming the world at an unprecedented pace.","Deep learning models like transformers are widely used in natural language processing tasks."

]

模型: bert-base-uncased

设备: cpu

总推理时间: 0.4100 秒

平均推理时间: 0.0820 秒/次

模型: gpt2

设备: cpu

推理次数: 5

总推理时间: 0.4500 秒

平均推理时间: 0.0900 秒/次

Dataset：imdb

模型: bert-base-uncased

设备: cpu

推理次数: 1000

总推理时间: 311.0198 秒

平均推理时间: 0.3110 秒/次

模型: bert-base-uncased

设备: cuda

推理次数: 1000

总推理时间: 7.2724 秒

平均推理时间: 0.0073 秒/次