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

以下是几个适用于测试 GPT-2 推理速度的数据集推荐：

1. **WikiText-2**
   * 描述：一个广泛使用的语言建模数据集，包含维基百科的文本。
   * 用途：适合评估语言模型的生成能力和推理速度。
   * 加载方式：load\_dataset("wikitext", "wikitext-2-raw-v1", split="test")
2. **OpenWebText**
   * 描述：一个开放的网络文本数据集，旨在模仿 OpenAI 的 WebText 数据集。
   * 用途：适合测试模型在更广泛的网络文本上的推理性能。
   * 加载方式：load\_dataset("openwebtext", split="test")
3. **BookCorpus**
   * 描述：一个包含大量书籍文本的数据集。
   * 用途：适合测试模型在长文本上的推理速度。
   * 加载方式：load\_dataset("bookcorpus", split="test")
4. **CNN/DailyMail**
   * 描述：一个新闻摘要数据集，包含新闻文章和摘要。
   * 用途：适合测试模型在新闻领域的生成能力。
   * 加载方式：load\_dataset("cnn\_dailymail", "3.0.0", split="test")
5. **Penn Treebank (PTB)**
   * 描述：一个经典的语言建模数据集，包含新闻文本。
   * 用途：适合测试模型在小型数据集上的推理速度。
   * 加载方式：load\_dataset("ptb\_text\_only", split="test")
6. **Common Crawl**
   * 描述：一个包含大量网络爬取文本的数据集。
   * 用途：适合测试模型在大规模数据上的推理性能。
   * 加载方式：load\_dataset("cc\_news", split="test")

你可以根据需求选择合适的数据集，并通过 datasets 库加载它们