# Inference with Reference Lossless Acceleration of Large Language Models

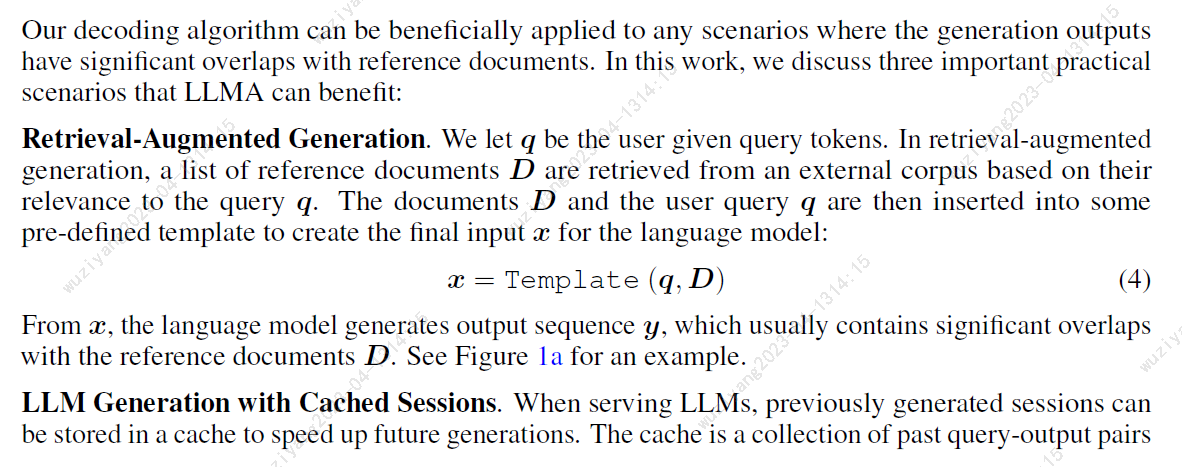
## 一、结论

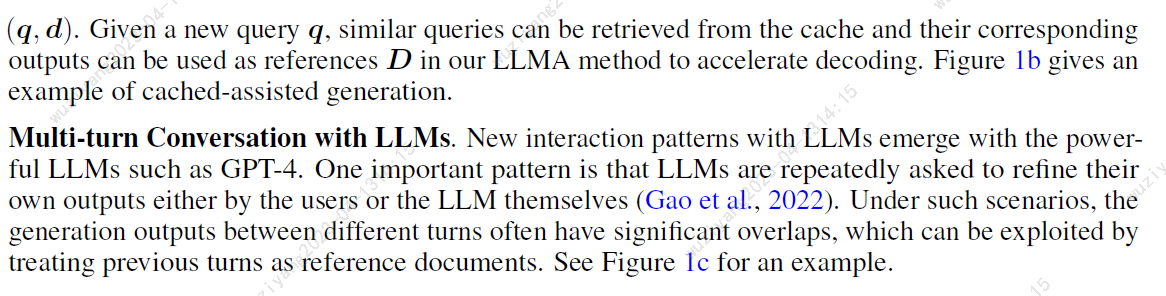
1.针对chatgpt类大模型，性能提升2倍以上；

2.需要大量额外的GPU资源，在不适用的场景下，GPU利用率不高。

## 二、应用场景

适用于输出结果与参考文本具有高度重叠的场景，对于无参考文本或输出结果与参考文本重叠度不高的场景效果不佳且会造成大量资源浪费。

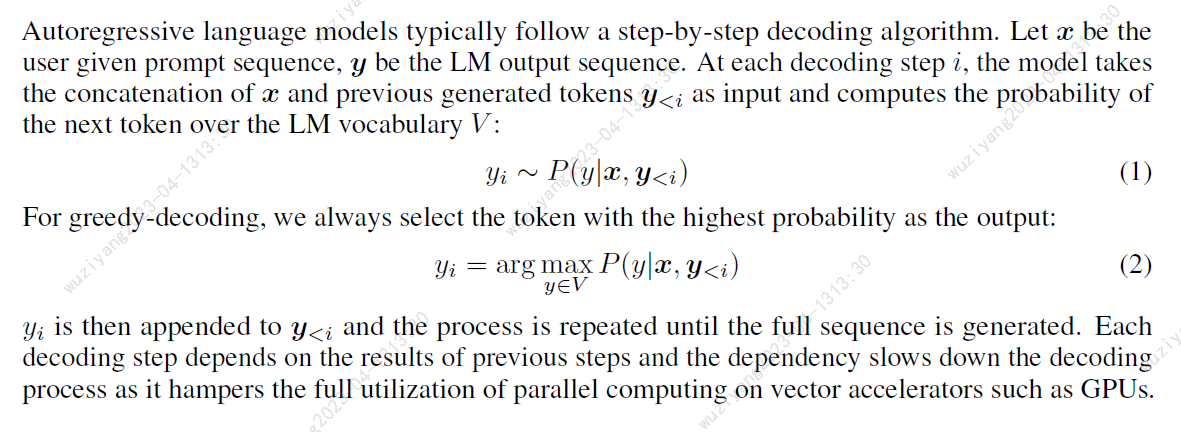




## 三、理论

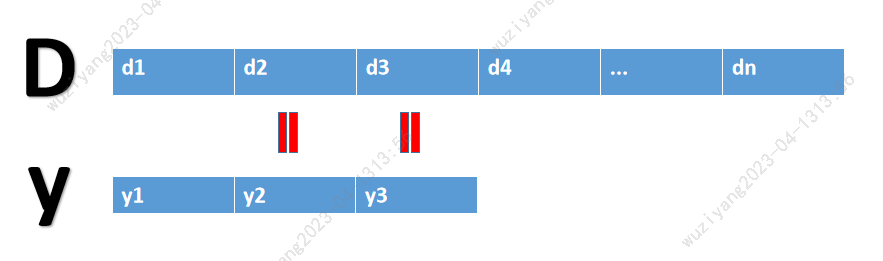
### 1.原方法

单字推理，逐步迭代，得到最终序列，无法充分利用多GPU并行能力

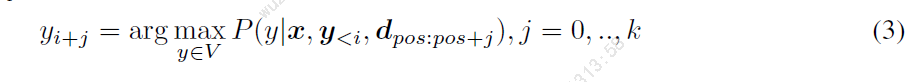


### 2.改进方法

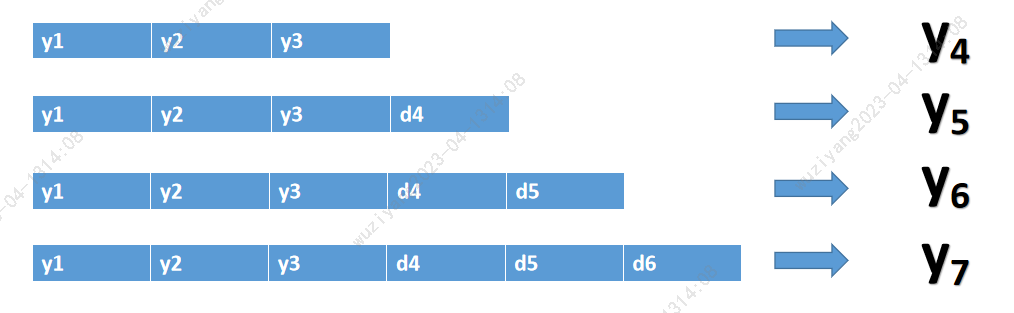
准备一组参考文本D，当D中存在与当前生成序列的最后n个token匹配的序列时，将D中匹配序列后续部分作为模型输入；如果存在多个匹配段，选择与y匹配长度最长的一段，如果存在多个最长匹配段，随机选择其中一个。



将D的后j个token输入到模型中，原方法计算公式变为：



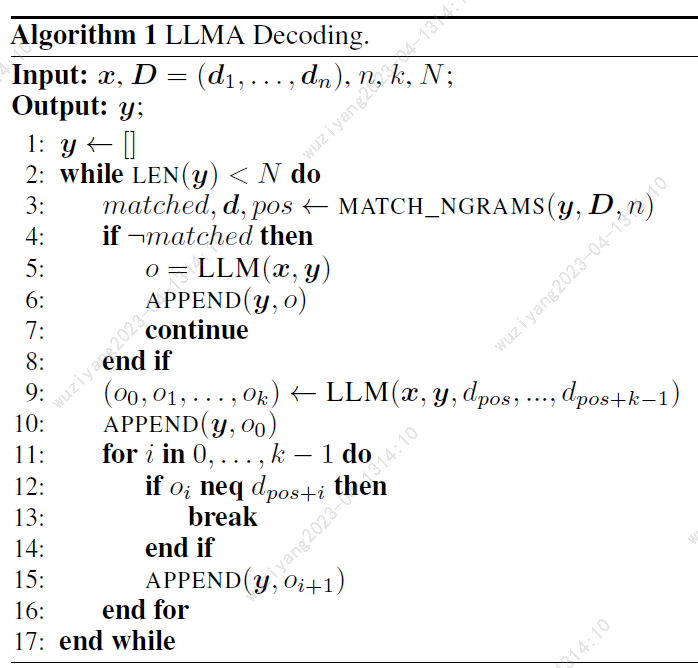
分别送到不同GPU做并行单步推理，一次推理得到多步结果：



依次判断生成结果yi : j是否与dpos : pos+j匹配，取最长匹配序列作为输出结果，随后进行下一轮推理。

如果没有找到匹配序列，则按照原方法进行单步推理。

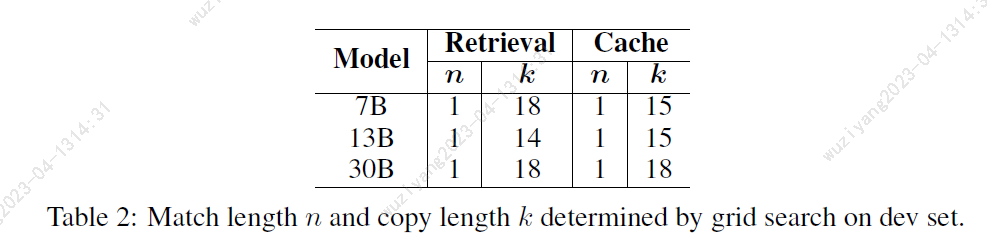
#### 伪代码



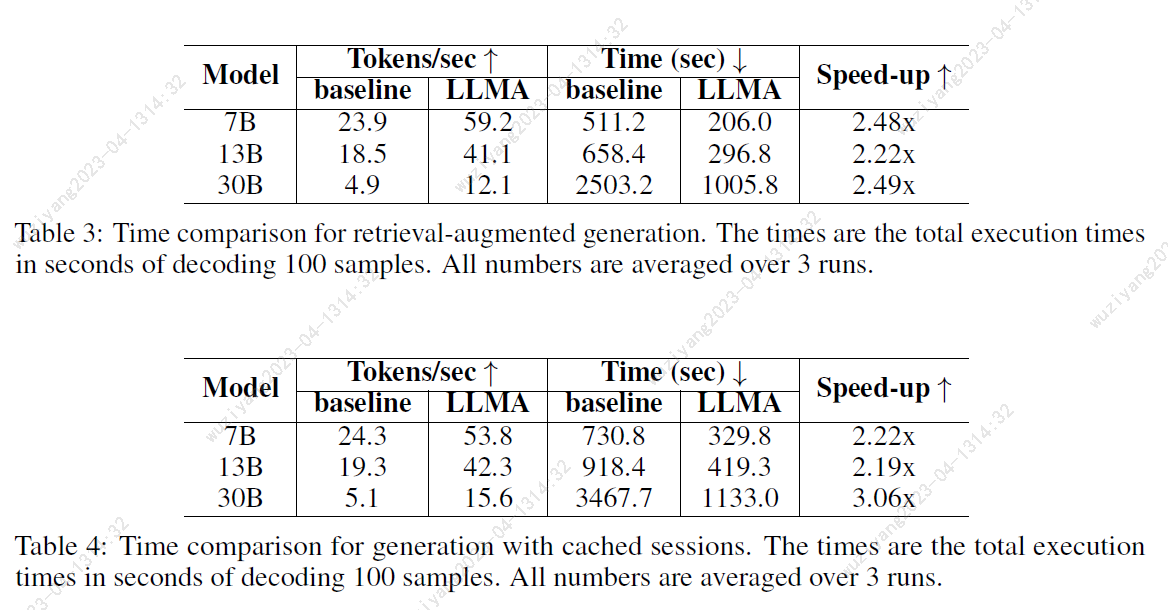
## 四、实验

### 1.配置

对于匹配长度n和并行长度k，参数配置如下：



### 2.结果



### 3.结论

较小的n和较大的k可以充分利用GPU并行能力，减少计算量，提高计算速度，降低数据传输损失。

