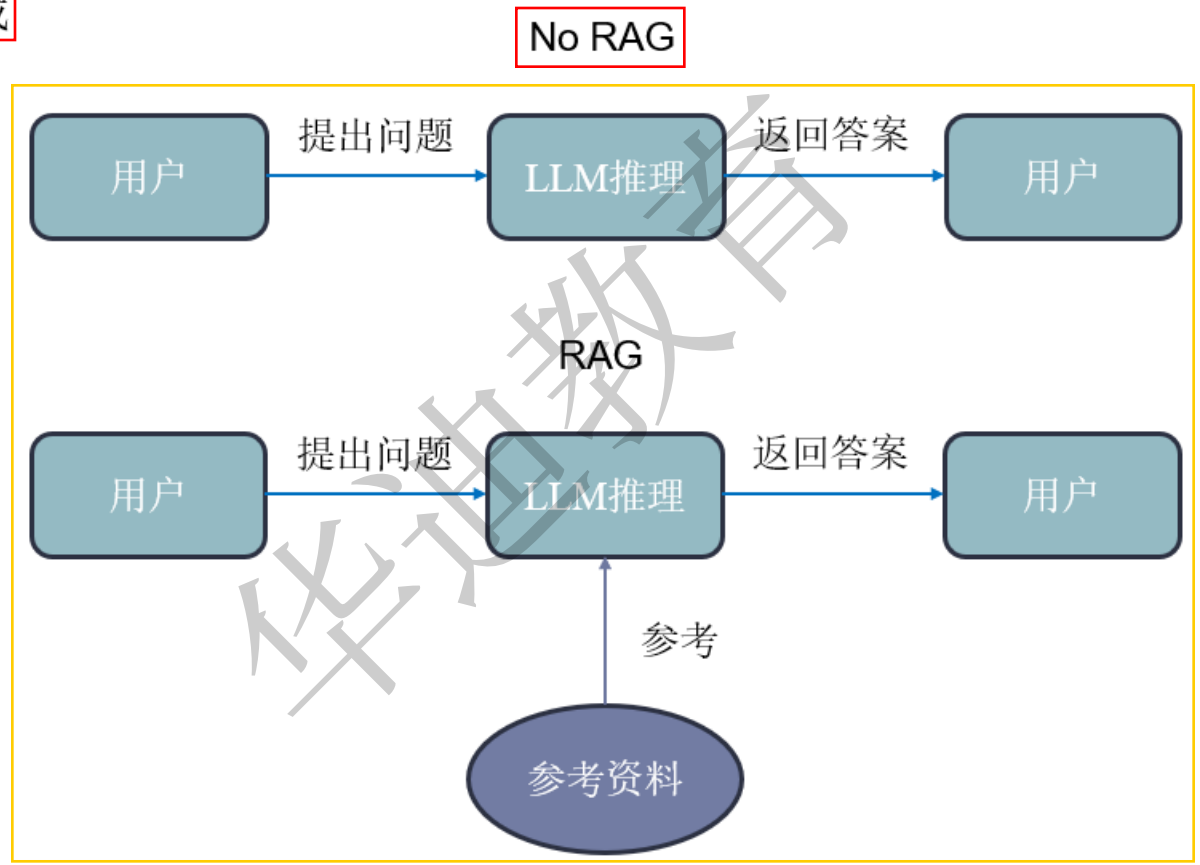
RAG技术

检索增强生成



RAG技术

1：文本分割，将长文档分割为一段一段的小文本。

2：文本嵌入，将每个小文本转化为向量。

3：相似度计算，语义相近的两个文本向量的点积更小。

4：问题和回复拼接，将与问题更相近的文本进行拼接，一起输送给模型。