# 阶段性计划

1. 数据集：dureader的部分数据（数据来源：<https://huggingface.co/datasets/THUDM/LongBench/viewer/dureader>，是将dureader部分数据格式化处理后的结果）
   1. 具体信息：Task Type: Multi-doc QA Avg len: 15,768 Sample: 200 Eval Metric: ROUGE-L
2. 基线模型：langchain-chatchat（原langchain-chatglm） v0.1.17，代码源 <https://github.com/chatchat-space/Langchain-Chatchat/blob/v0.1.17>
3. 时间安排
   1. 8.25-8.29 对拆分方法进行测试，包括对chunk尺寸的设计以及一些小思路（比如强化标题的作用、上下文加强等）
   2. 8.29-8.31 对不同的embedding进行测试
   3. 8.31-9.3 借鉴chatlaw的keyword llm等思路，将关键词信息嵌入，进行相关测试