# 工作的意义

这篇文章解决了Multi-Vector Retrieval(MVR)与RAG中的一个核心瓶颈——查询分解质量。传统方式要么逐token分解(ColBERT)，颗粒度过细导致语义错配，要么依赖人工提示驱动LLM生成子查询，但往往包含一些干扰信息，从而削弱检索与问答性能。论文通过分析“Hong Kong”被误解成“Hong”+ “Kong”的案例，直观揭示了现有方法的局限性：错误的粒度划分或冗余子查询会使模型把无关文档或图像排到更靠前的位置，直接影响下游QA质量。详细来说，给定一个由m个token构成的查询Q，如何自动生成由这些token组成的k个子查询，使下游检索性能最大化，难点有二：1.下游模型对子查询是不可微分的，无法通过梯度传播改进2.评估候选分解需反复训练下游模型，计算代价太高。

基于上面的难题，这篇论文提出了POQD：通过LLM生成子查询，并借助LLM Optimizer搜索和迭代refine提示；训练时采用交替优化，仅需少量下游模型训练即可逼近最优解。该方法在理论上可得到合理性保证，并在图文QA与文本QA等多类RAG任务上显著优于现有的方法，在检索准确率与问答正确率上均大幅提升。

# 方法论

**POQD Framework**

针对现有多向量检索方法依赖人工启发式分解、难以对齐下游任务需求的不足。本文提出Performance Oriented Query Decomposition(POQD)框架，其核心是以下游问答性能为唯一优化目标，通过LLM主导的Sub-query Generation生成多样候选子查询，再利用Prompt Optimization自动提出语义冗余与噪声子查询，从而保留能够覆盖原始查询意图的关键表达。在此基础上，引入Alternating Training机制：提示优化与下游模型训练交替进行，每轮迭代均以QA正确率反馈为信号，逐步逼近最优分解。整个流程形成生成——优化——训练——反馈的闭环，确保子查询分解真正提升最终答案质量，而非仅仅优化中间检索指标。

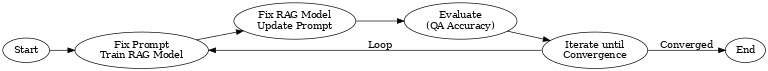
POQD_flowchart

**End to End Training Algorithm**

POQD的优化目标本质上是不可微分的，因为性能反馈来自于离散的QA正确率评估，且在候选空间极大时。直接对每个分解方案进行完整训练代价高昂。因此本篇论文提出了端到端交替训练算法。其基本思想是将优化问题分解为两个互补的子过程：

1. 当prompt fixed，训练下游RAG模型若干轮，使其参数与当前子查询分解保持一致
2. 当下游模型固定时，利用反馈信号更新提示，通过LLM或轻量调优机制生成更优子查询

这种交替式优化避免了对不可微分目标的直接梯度估计，通过显著降低了计算成本。这篇论文基于前人的多种假设上证明了该方法的收敛性，可以在有限迭代下即可获得接近全局最优的分解效果



# 优缺点分析

# 未来的研究方向

Based on this paper, 主要集中在对Query的分解，同时训练中。可以考虑Query-Document 协同分解与多目标优化框架，