实验设计指导（可能的一些实验）

实验设置与参数实现

核心参数设置

实验中可能涉及到的一些超参数：

语义熵阈值 (τ): 设定为 0.5。当 $SE(x) > 0.5$ 时，判定模型处于高不确定性状态。

混合检索权重 (α): 设定为 0.7，即 70% 依赖向量相似度，30% 依赖 BM25 关键词匹配。

反思权重 (w_{sup}): 在评分函数中，设定支持度权重 $w_{sup} = 2.0$，以优先保证事实准确性。

一：动态检索策略的有效性验证

目的： 验证提出的混合检索与Cross-Encoder 重排序是否优于单一检索。

对比维度：

Sparse Only: 仅使用 BM25。

Dense Only: 仅使用向量检索。

Hybrid (Ours):加权融合。

Hybrid + Rerank (Ours Pro): 在混合检索基础上增加 Cross-Encoder 重排序。

二：基于语义熵的幻觉检测验证

实验目的： 验证定义的语义熵是否能有效区分“正确回答”和“幻觉回答”。

实验方法：

对测试集中的每个问题进行 M次采样。

利用双向蕴含聚类算法计算 $p(c_k|x)$。

计算每个样本的 $SE(x)$ 值，并与某个baseline模型判定的事实错误率进行相关性分析。

三：自反思机制的修正能力验证

实验目的： 验证检索-生成-批评流程能否修正错误。

实验设计:

Naive RAG: 无反思机制，直接生成。

Self-RAG : 有检索，但移除反思令牌权重（令 $w_k = 0$）。

Self-RAG: 完整模型，启用所有反思令牌。

评价指标：

Citation Precision: 引用准确率（生成的句子是否真的被文档支持）。

FactScore: 事实得分。