# 茴香豆，RAG智能助理

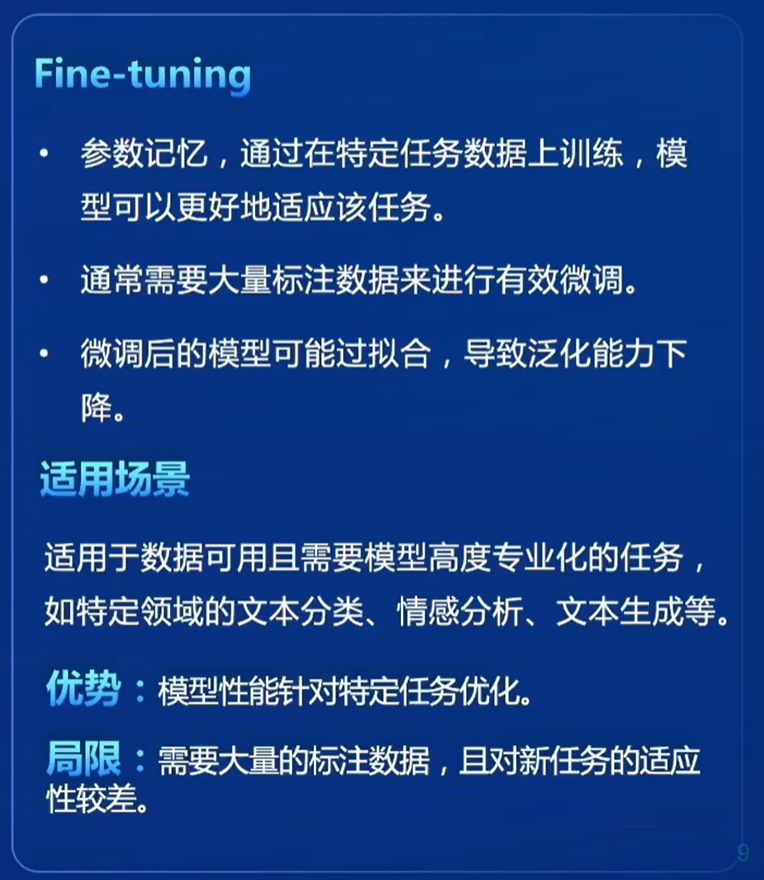
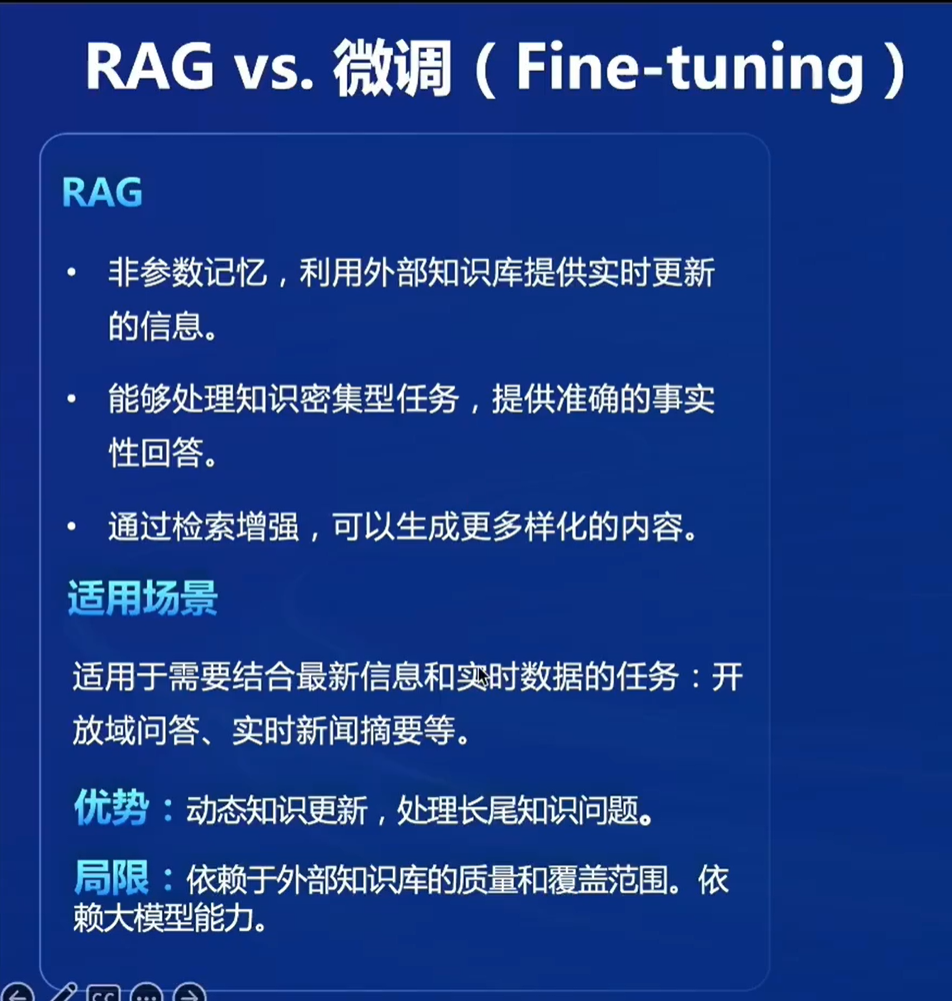
定义：retrieval augmented generation，结合检索和生成技术，通过利用外部知识库增强llms的性能。检索与用户输入相关的信息片段，并结合这些信息来生成更优质的回答。

实质：通过预先建立知识源作为外部知识，通过类似聚类的方式作为辅助提示词对llms进行提问。

关键：向量数据库—将知识源以固定长度的向量进行语义表示。通过计算余弦相似度、点乘等方式完成检索。

优点：避免幻觉问题，问题回答更确切，信息更新更快、时效性好。

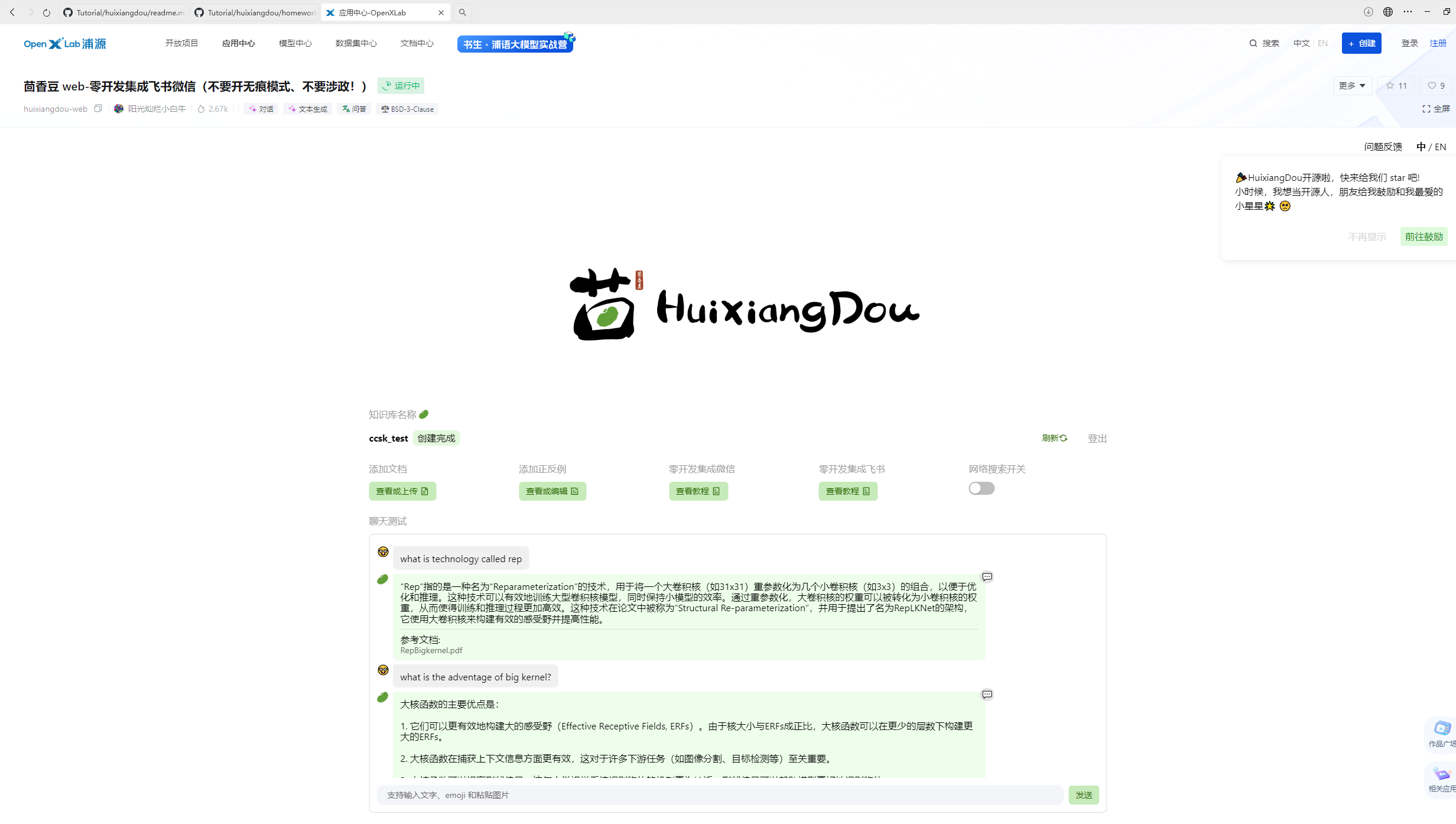
延伸：可以用它泛化到我们的研究领域，作为一种实时的数据增强方式。



模型优化的参考点：对模型适应性的要求，对外部知识的需求。

流程：问答预处理之后通过相似性检验决定是否回答，随后进入问答通道。

运行截图：



作用：可以作为一个论文总结的阅读助手。

