# 茴香豆：搭建你的RAG智能助理

## RAG是什么、原理

在垂直领域去问大模型的时候，针对新增知识的情况，传统会采集新增预料然后通过微调的方式对大模型再训练，但是这个方案一般也有显著的缺陷，如语料更新太快，语料过大，除了这种方案外，2020年Meta(Facebook)提出了RAG技术，可以再不需要微调训练的基础上就使得大模型实现不错的回答效果。

RAG叫做检索增强生成技术，通过检索与用户输入相关的外部知识库召回信息片段，并结合这些信息来生成更准确丰富的回答，可以有效解决生成幻觉、过时知识、缺乏透明和可追溯的推理过程，提供更准确的回答、降低成本、实现外部记忆。

RAG可以应用在问答系统、文本生成、信息检索、图片描述。

经典RAG分为索引、检索、生成三个部分，索引是将外部数据源分割成chunk并存储在向量数据库中；检索是收到用户问题，将问题编码为向量，并在向量库中找到与之相关的文档块；生成则是将检索到的和用户问题最相关的文档块和原始问题一起作为prompt输入给LLM，生成最终答案。

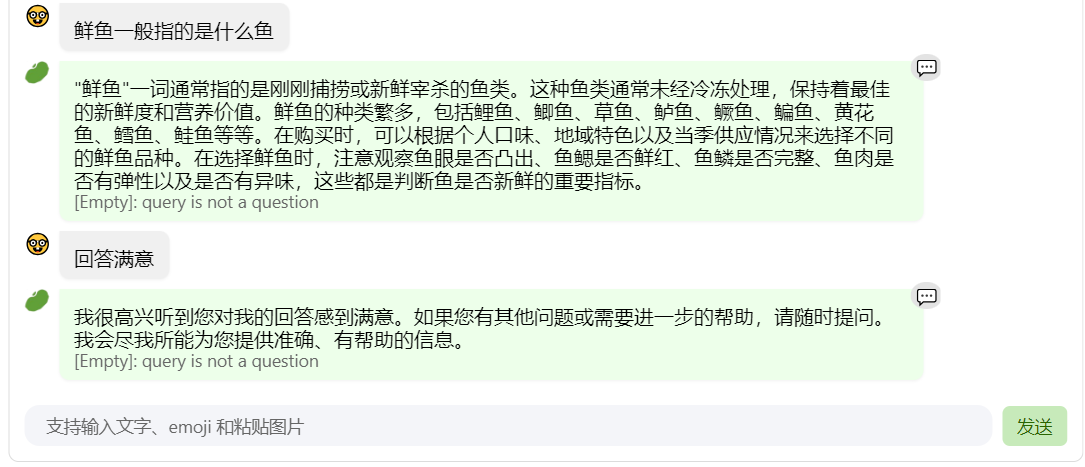
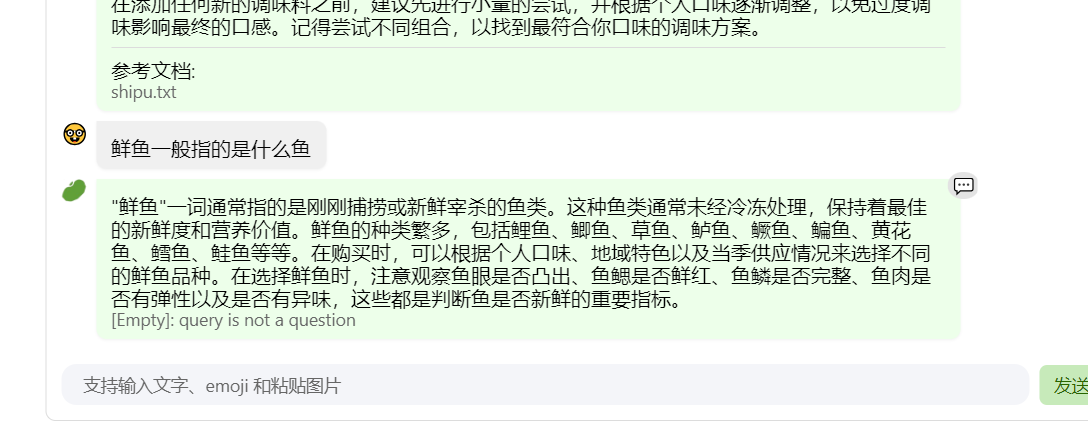
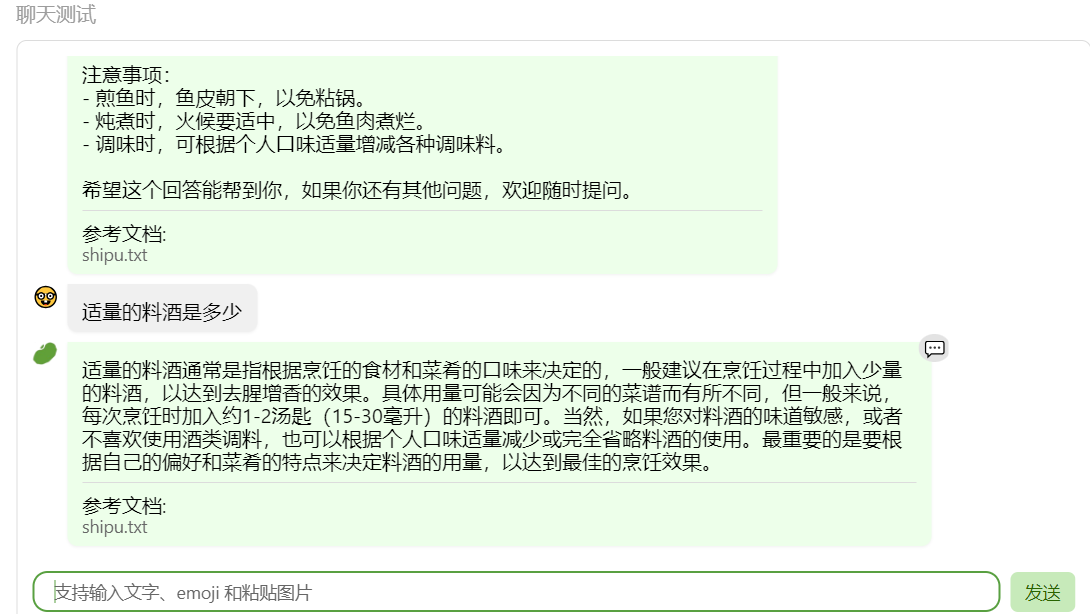
## RAG优化

1. 嵌入优化 结合稀疏和密集检索任务
2. 索引优化 细粒度分割（chunk）添加元数据
3. 查询优化 查询扩展、转换 多查询方法 通过大模型提示工程扩展查询
4. 上下文管理 通过重排 训练小一些的模型进行上下文选择/压缩
5. 迭代检索 递归检索 自适应检索（flare、self-RAG）
6. LLM微调 检索微调 生成微调 双重微调

## 茴香豆是什么

茴香豆是一款基于LLMs开发的领域知识助手，由书生浦语团队开发的开源大模型应用，专为IM工具提供及时准确的技术支持和自动化问答服务，通过RAG技术，茴香豆能够理解和高效准确的回应特定知识领域相关的复杂查询。

## **[茴香豆 Web 版](https://openxlab.org.cn/apps/detail/tpoisonooo/huixiangdou-web)创建**领域知识问答助手截图：



## InternLM Studio 茴香豆技术助手截图：

