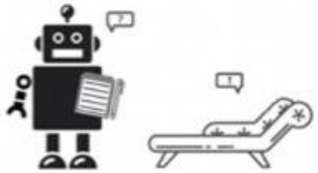
大模型工程化之路

发展历程



2、NLP、AIML标签语言
自动机、Slot-Filling等

心理医师 ELIZA
MIT Artificial Intelligence Laboratory

Loebner Prize设立
每年数千美元的奖金
对首次首次通过图灵测试额外奖励

Jabberwacky
获得Loebner Prize 2005、2006连续2届冠军

IBM Watson
在《危险边缘》Jeopardy!战胜人类

Google Now

Google Transformer

OpenAI ChatGPT

1966 1991 1997 2011 2012 2017 2022

1950 1972 1995 2005 2011 2015 2018

图灵测试
Alan Turing

模拟精神病患者 PARRY
Stanford

A.L.I.C.E
在互联网上可以使用，获得Loebner Prize 2000、2001、2004共3届冠军

Mitsuku
获得Loebner Prize 2014、2016、2017共3届冠军

苹果Siri

微软小冰
亚马逊Alexa
阿里小蜜

OpenAI GPT

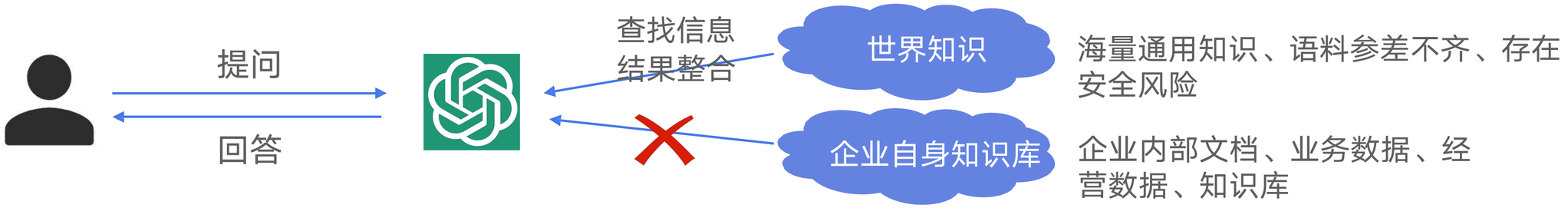
1、专家系统

3、神经网络、强化学习、端到端对话系统

4、生成式AI

如何让大模型更好地为企业服务？

直接使用大模型？



大模型在知识问答场景存在的问题

解决思路

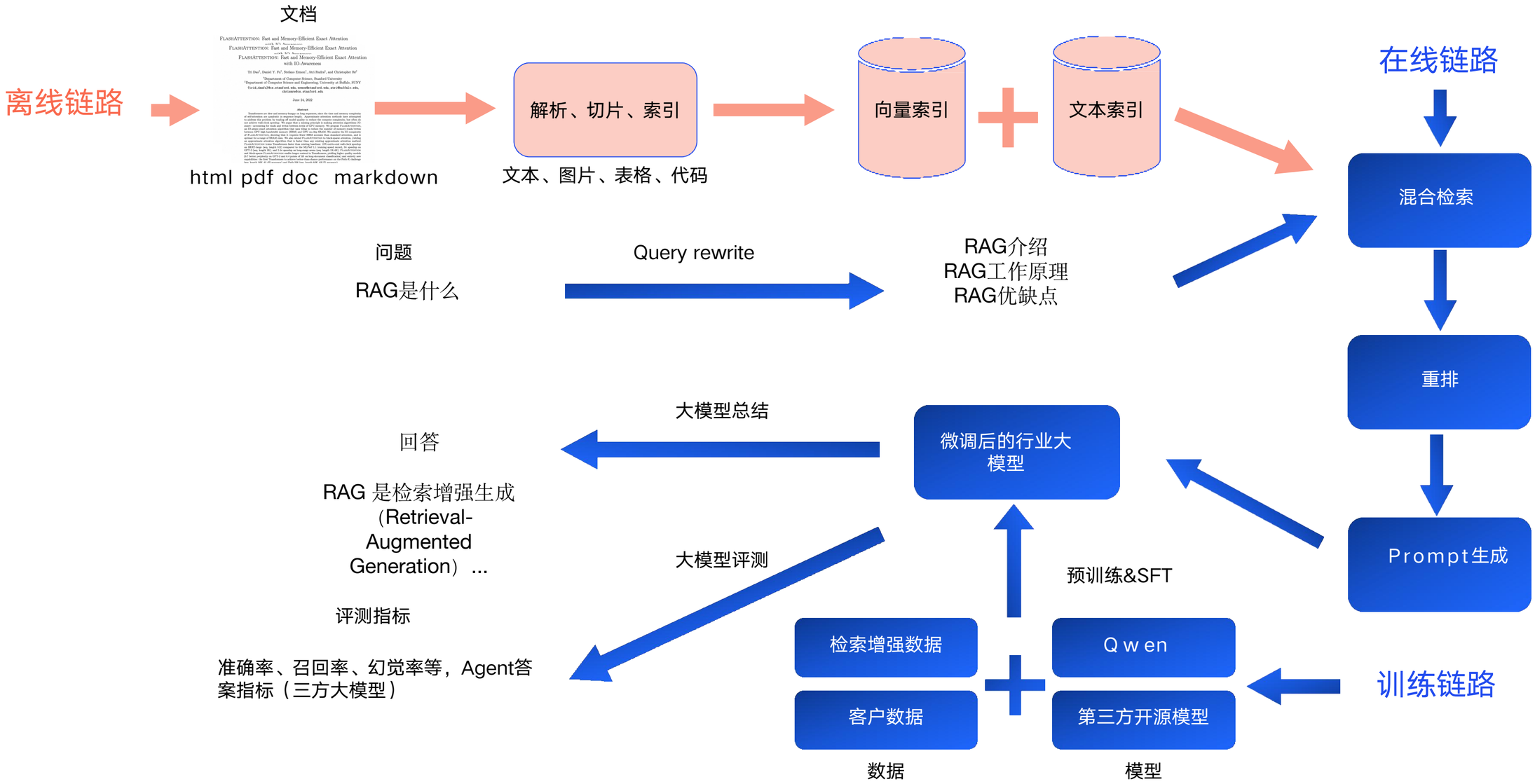
RAG（**知识向量化**）：Retrieval-Augmented Generation，检索增强生成。用搜索结果引导LLM的生成，让LLM在受限的范围内给出答案。

模型微调（**知识参数化**）：Fine-Tuning，指对预训练过的大模型在特定任务或特定领域的数据上进行进一步训练，以提升其在该任务或领域的表现。提升特定场景的问答准确性和业务相关性。

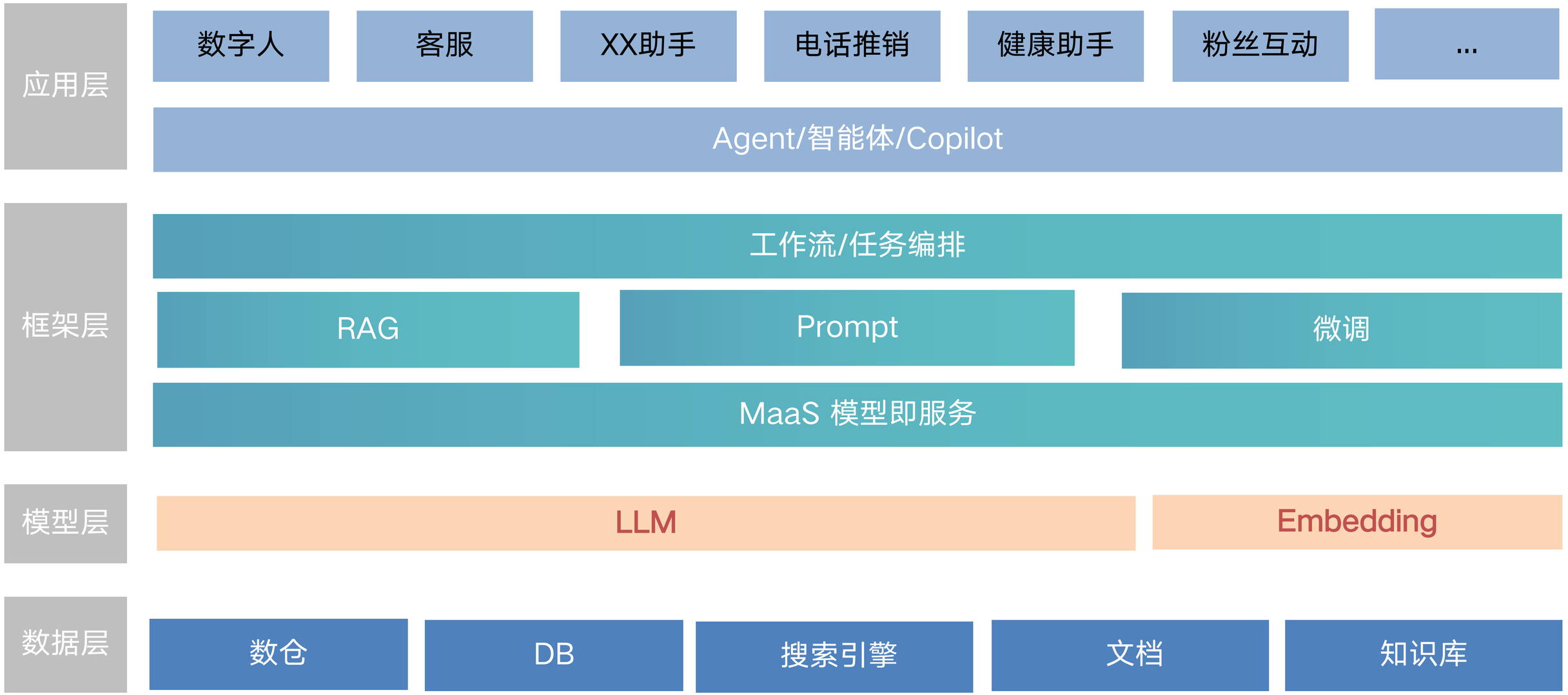
RAG与微调的优劣对比

比较维度	检索增强生成（RAG）	模型微调
优点		
实时性	可以实时获取最新信息，无需更新模型	需要定期更新模型以适应新知识
计算成本	计算成本较低，主要依赖检索系统	计算资源需求高，特别是在大模型上进行微调
适应性	通过调整检索库可以快速适应不同领域	微调后的模型生成内容更定制化，更符合预期
数据需求	无需大量训练数据，仅需构建高质量检索库	需要大量高质量数据来进行模型微调
幻觉问题应对	RAG通过检索真实数据降低幻觉问题，信息更具可溯源性	通过高质量、定期更新的数据微调减少幻觉，但仍可能存在
缺点		
依赖检索质量	依赖检索库内容的覆盖和准确性，检索库不足时效果受限	独立生成答案，不依赖检索库
上下文一致性	容易出现上下文断裂，复杂对话场景中表现较差	能保持上下文一致性，更适合处理复杂多轮对话
适应新问题能力	若检索库无对应内容，可能无法生成满意答案	能泛化处理变动问题，不局限于检索库

典型工程链路

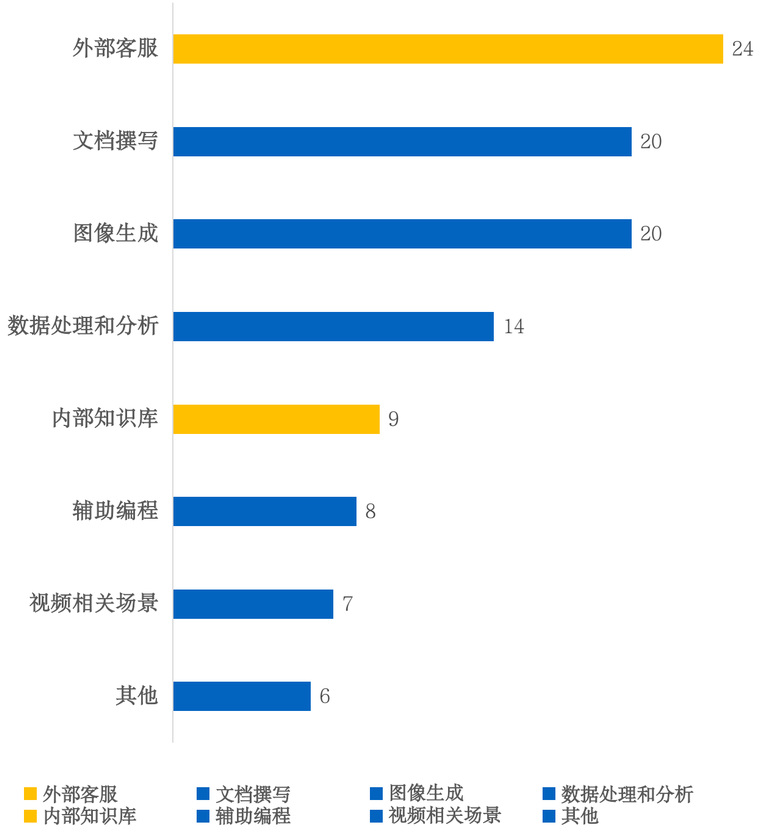

文档
离线链路
FLASHATTENTION: Fast and Memory-Efficient Exact Attention
FLASHATTENTION: Fast and Memory-Efficient Exact Attention
FLASHATTENTION: Fast and Memory-Efficient Exact Attention with IO-Awareness
html pdf doc markdown
解析、切片、索引
文本、图片、表格、代码
向量索引
文本索引
在线链路
混合检索
重排
Prompt生成
问题
RAG是什么
Query rewrite
RAG介绍
RAG工作原理
RAG优缺点
大模型总结
回答
RAG 是检索增强生成（Retrieval-Augmented Generation）...
评测指标
准确率、召回率、幻觉率等，Agent答案指标（三方大模型）
大模型评测
微调后的行业大模型
预训练&SFT
检索增强数据
客户数据
数据
Qwen
第三方开源模型
模型
训练链路

行业产品应用架构


应用层
数字人
客服
XX助手
电话推销
健康助手
粉丝互动
...
Agent/智能体/Copilot
框架层
工作流/任务编排
RAG
Prompt
微调
MaaS 模型即服务
模型层
LLM
Embedding
数据层
数仓
DB
搜索引擎
文档
知识库

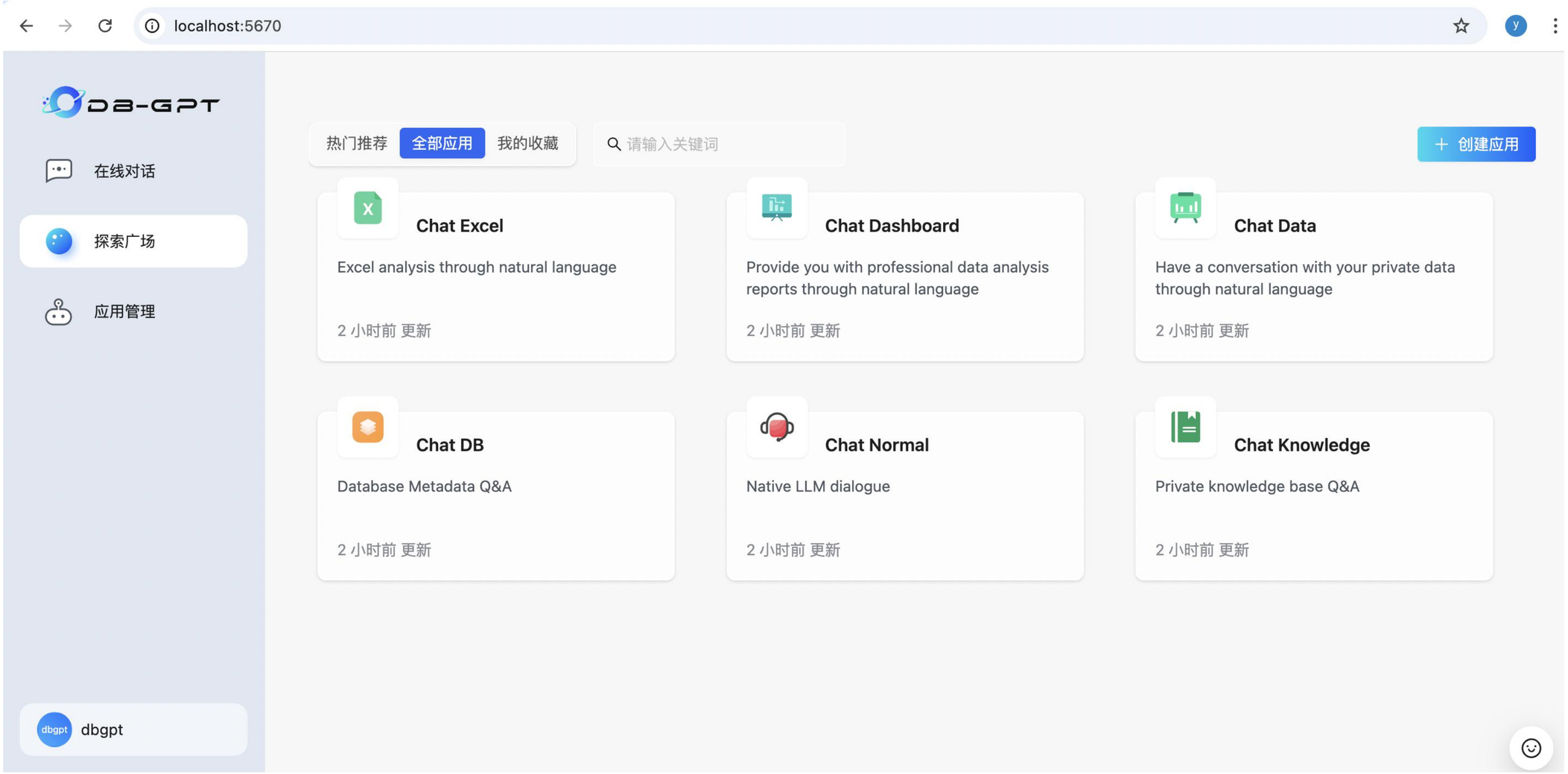
应用场景

大模型应用场景调研



行业	具体应用场景
汽车	• 客户服务与支持：通过RAG，客户可以咨询车辆功能、维护、最新技术或政策法规相关信息。 • 工程师研发：工程师利用RAG结合LLM检索技术文档、研究论文和专利信息，加速新技术研发。 • 销售辅助：销售人员通过RAG检索最新车辆数据和市场趋势，为潜在买家提供个性化车型推荐
零售	• 智能客服：结合RAG的LLM提供24/7客户服务，快速检索产品信息。 • 销售培训：利用RAG搜索相关资料，进行销售培训。 • 自动化产品描述：LLM自动生成吸引人的产品描述和营销文案，RAG检索产品规格和用户评价，确保 信息准确性和吸引力。
文娱	• 音乐与电影推荐：LLM结合用户喜好和历史数据，生成个性化音乐播放列表或电影推荐；RAG实时 检索最新音乐和电影数据库，确保推荐时效性和多样性。 • 互动式娱乐：在游戏或虚拟现实体验中，LLM作为虚拟角色AI与用户自然对话，RAG检索用户与角 色间经历，丰富互动体验。 • 粉丝互动：明星或品牌官方平台使用结合RAG的LLM回答粉丝问题，提供最新新闻和活动信息，提 升粉丝参与度和忠诚度。
金融	• 智能客服：结合RAG的LLM提供24/7客户服务，快速检索金融产品信息，提供个性化服务与支持。 • 保险销售辅助：RAG检索保险理赔条款，助力快速推动产品销售。
医疗	• 文献检索与分析：研究人员利用结合RAG的LLM检索医学文献和研究报告，分析数据，发现新研究线 索和趋势。 • 虚拟健康助手：结合RAG的LLM回答患者健康相关问题，提供个性化健康建议，检索最新医疗研究信 息，教育患者。 • 症状评估：LLM分析患者症状描述，RAG检索医疗知识库，提供诊断和治疗方案。

产品演示



https://github.com/eosphoros-ai/DB-GPT

选择哪个基座模型

底座	包含模型	模型参数大小	训练token数	训练最大长度	是否可商用
ChatGLM	ChatGLM/2/3/4 Base&Chat	6B	1T/1.4	2K/32K	可商用
LLaMA	LLaMA/2/3 Base&Chat	7B/8B/13B/33B/70B	1T/2T	2k/4k	部分可商用
Baichuan	Baichuan/2 Base&Chat	7B/13B	1.2T/1.4T	4k	可商用
Qwen	Qwen/1.5/2/2.5 Base&Chat&VL	7B/14B/32B/72B/110B	2.2T/3T/18T	8k/32k	可商用
BLOOM	BLOOM	1B/7B/176B-MT	1.5T	2k	可商用
Aquila	Aquila/2 Base/Chat	7B/34B	-	2k	可商用
InternLM	InternLM/2/2.5 Base/Chat/VL	7B/20B	-	200k	可商用
Mixtral	Base&Chat	8x7B	-	32k	可商用
Yi	Base&Chat	6B/9B/34B	3T	200k	可商用
DeepSeek	Base&Chat	1.3B/7B/33B/67B	-	4k	可商用
XVERSE	Base&Chat	7B/13B/65B/A4.2B	2.6T/3.2T	8k/16k/256k	可商用

分组讨论 ME-WE-US

Ques：
这门课程需要掌握多门编程语言，这种情况在未来工作场景中是常态，我们如何快速面向课程目标按需掌握新的编程语言？

每位同学可以先2分钟时间自己思考一下：
最大的困难是什么？
最大的信心来自于哪里？
。。。

然后分小组讨论，每组5人左右，10分钟充分交流观点，选一个组长上来分享一下你们组内的idea~

Project1

准备工作：python环境、IDE（PyCharm /IDEA /VS Code..），python项目初始化

产出：实现控制台简单的多轮对话交互（可选：流式输出）

```
用户：我想买一双鞋，有什么建议吗？请20字以内回答我
助手：确定用途，试穿确保舒适，选择合适材质与尺码，关注品牌与评价。

用户：有哪些尺码呢？请20字以内回答我
助手：常见尺码有中国35-44码，美国5-12码，欧洲35-48码。请根据需要选择。

用户：颜色有什么建议吗
助手：选择经典色如黑、白、灰，或根据个人喜好和服装搭配决定。

用户：怎么搭配衣服呢
助手：根据鞋子类型搭配：运动鞋配休闲装，皮鞋配正装，帆布鞋配牛仔裤。颜色上保持协调即可。

用户：好的我知道了
助手：感谢您的咨询，再见！
```

Project1-需求

1. 支持多轮对话

2. 当用户表达对你表达了希望对话结束的语义，请给用户返回:“感谢您的咨询，再见”

3.（可选）对话支持流式输出

Project1：环境安装

1. python安装

2. 尝试使用通义千问API，完成控制台的输出

【参考文档】
环境确认：https://help.aliyun.com/zh/model-studio/getting-started/first-api-call-to-qwen

API参考：https://help.aliyun.com/zh/model-studio/developer-reference/dashscopellm