

基于RAG构建生成式 AI应用

最佳实践与“避坑指南”

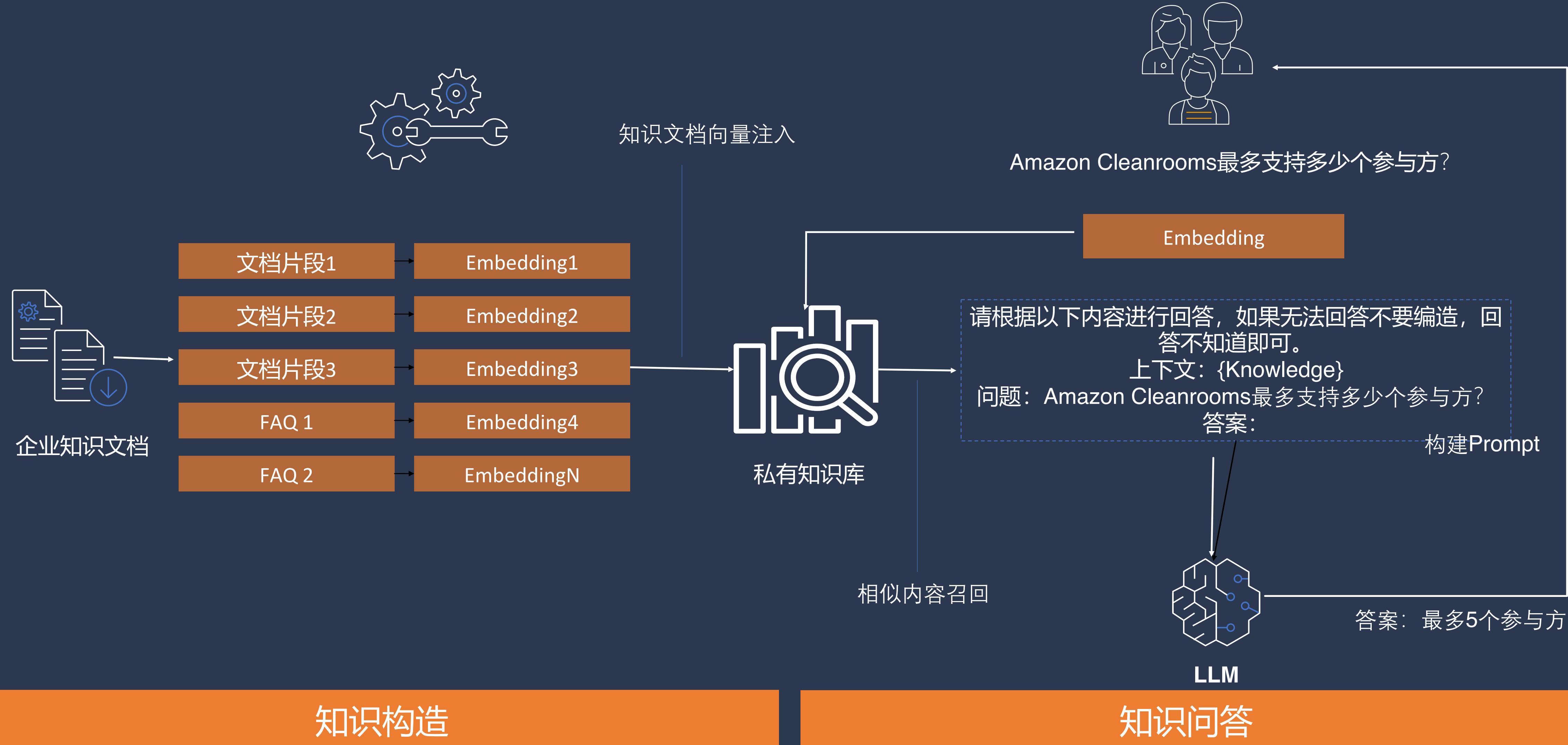
亚马逊云科技人工智能技术专家 / 李元博

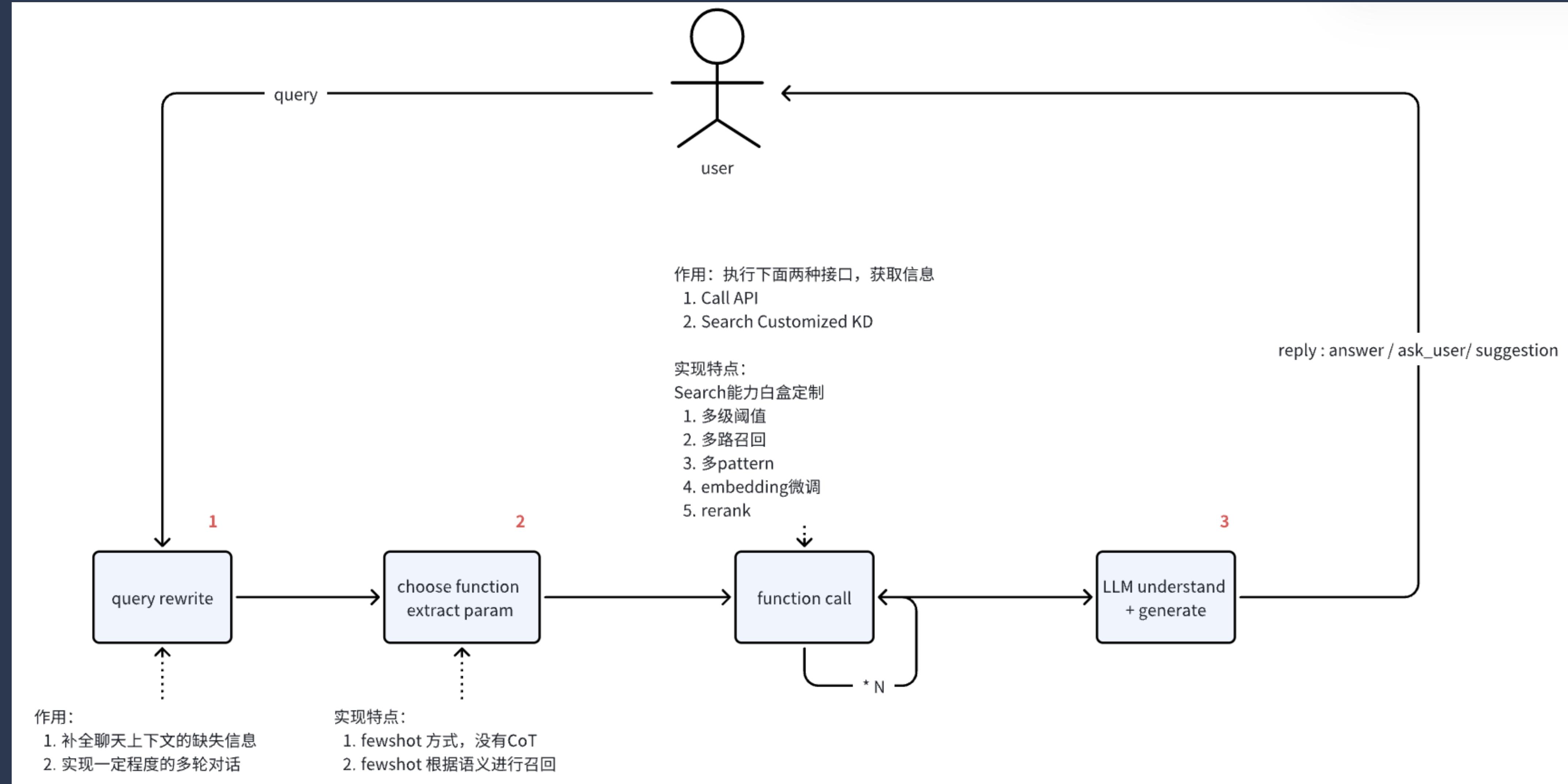
目录

- RAG场景及技术特点
- RAG实践经验总结
- RAG场景中的亚马逊云产品亮点

RAG场景及技术特点

RAG概念流程图

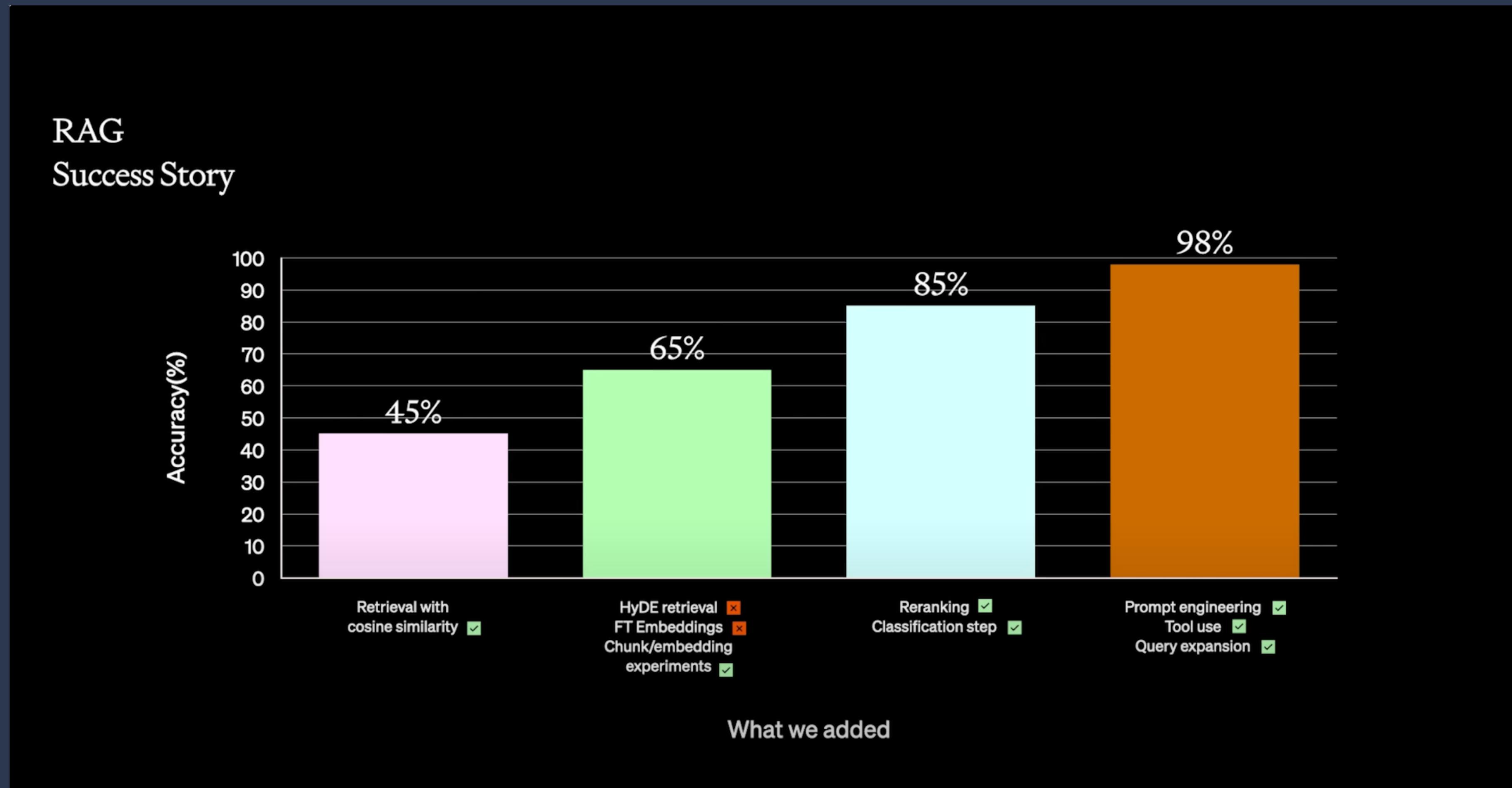




RAG适用典型的木桶理论场景



OpenAI 关于RAG的分享



RAG场景分类

维度	分类	描述	Tips
按使用场景	Chatbot	<ul style="list-style-type: none">➤一问一答聊天交互➤召回topK记录给到LLM归纳总结，返回答案	<ul style="list-style-type: none">➤对于置信度高的召回，可以直接返回top one而不走LLM归纳总结（避免LLM幻觉或自由发挥）。比如某车企的Chatbot通过此方式获得>95%的准确率➤误答容忍度低
	智能检索	<ul style="list-style-type: none">➤知识检索形式交互➤除了返回LLM总结的答案，可选择展示top K的召回记录以及知识源语料	<ul style="list-style-type: none">➤因为返回多条记录，对知识相关性容忍度相对较高➤可使用引导式检索，逐步获取精准答案
按知识类型	FAQ对	<ul style="list-style-type: none">➤按问答对的方式构建知识➤知识信息量完整，知识质量高	<ul style="list-style-type: none">➤按FAQ对进行切片，保证语义完整性➤现有客服场景通常都有语料积累，知识构建相对高效➤上线相对容易
	PDF/word等各类文档	<ul style="list-style-type: none">➤原始文档格式多样，比如各类wiki，产品说明书等，含有表格、图文等信息➤知识质量可能参差不齐，密度低	<ul style="list-style-type: none">➤通常按句子，段落等进行切片，较FAQ对切片方式复杂，需要结合实际文档来保证语义完整性➤表格需要额外处理，暂时不建议对图片处理
按使用对象	服务内部用户	<ul style="list-style-type: none">➤比如企业内部知识库，IT/HR 知识库	<ul style="list-style-type: none">➤内部人员和并发等因素可控，上线相对容易
	服务外部用户	<ul style="list-style-type: none">➤比如各行业对外的智能客服，游戏的NPC	<ul style="list-style-type: none">➤对并发/吞吐等性能指标有较高要求，另外需要考虑内容风控比如屏蔽涉黄/涉恐/涉暴话题

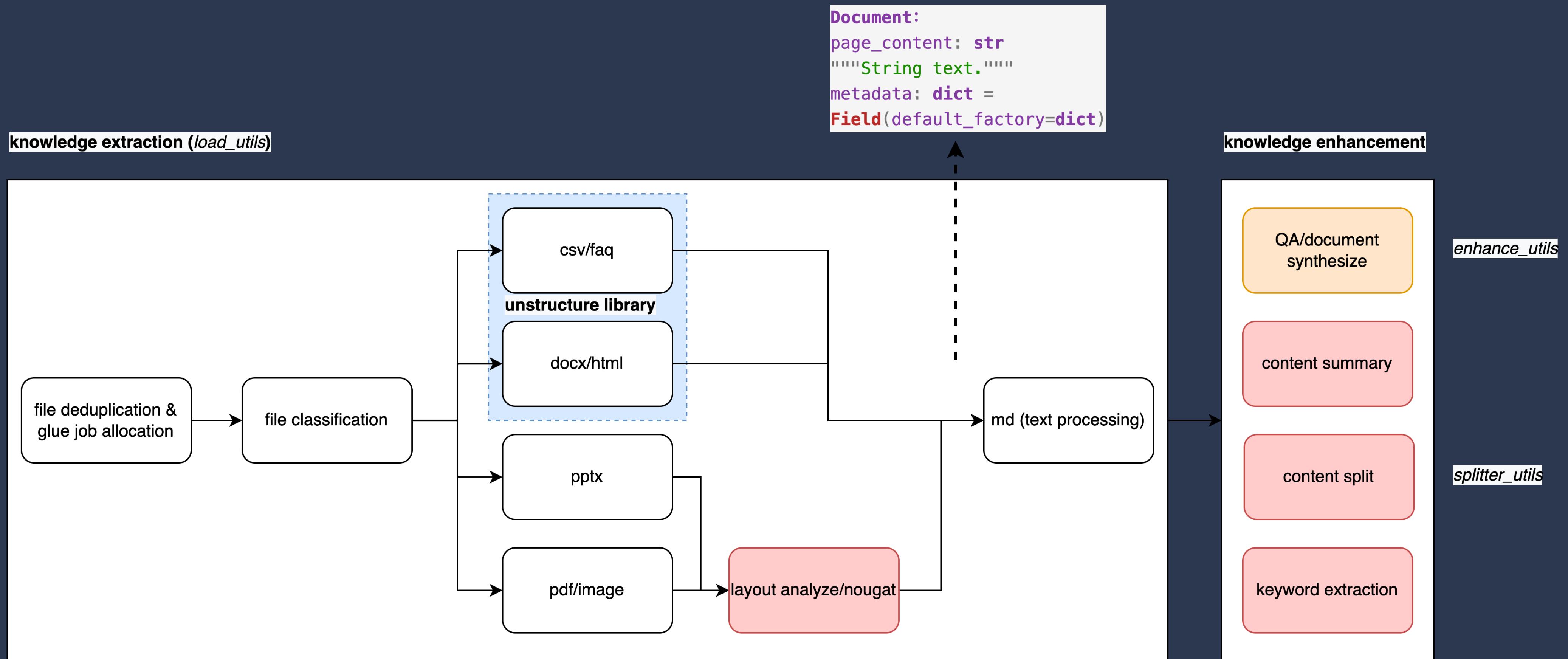
RAG实践经验总结

技术层面经验总结

- 知识构建 – 知识质量不好咋办？格式复杂还有表格咋办？量大注入慢咋办？
- 核心工作 – 知识召回效果差咋办？LLM幻觉咋办？怎么判断超出知识库范围？
- 上线必备 – 效果如何评估和持续监控？Badcase 怎么排查？Streaming 咋支持？
- 经验洞察 – Agent API 设计思考？反问机制设计？Web Search集成方法？飞书的集成与实现？知识共创UGC机制？

知识构建 – 知识质量不好咋办?

✓ 利用商业LLM进行知识增强(OCR Parser + Layout识别 + 信息合成 + LLM视觉能力)



知识构建 – 格式复杂还有表格咋办?

✓ 结合开源代码 或者Amazon AIML SAAS服务(OCR Parser + Layout识别)

□ 实践检验:

解决海外一个金融POC中，针对基金相关PDF的表格信息提问的问题

□ 解决方案

- 利用Langchain开源代码实现PDF转HTML，可保留字号和像素位置信息，按字号进行合并，提高分段质量
- 利用Textract提取表格，实现Textract输出到Json信息的转换
- 利用前两步结果进行位置映射，克服Textract不支持中文的问题
- 其他手段(NSP模型&Layout分析)

□ 使用说明:

- [PDF SPLITER README](#)
- [Workshop \(第三步实验\)](#)

□ 参考效果:

The screenshot shows a PDF page from Allspring's website. At the top, it says "AS OF DECEMBER 31, 2022 | FACTSHEET | ALLSPRINGGLOBAL.COM". Below that is the "Common Stock Fund" logo. The page contains several sections with red boxes highlighting specific parts:

- Asset Class: U.S. Equity**
- Competitive advantages**: A section listing two bullet points about market valuation and core approach.
- Sector allocation (%)¹**: A table comparing the fund's allocation across various sectors against the Russell 2500TM Index.

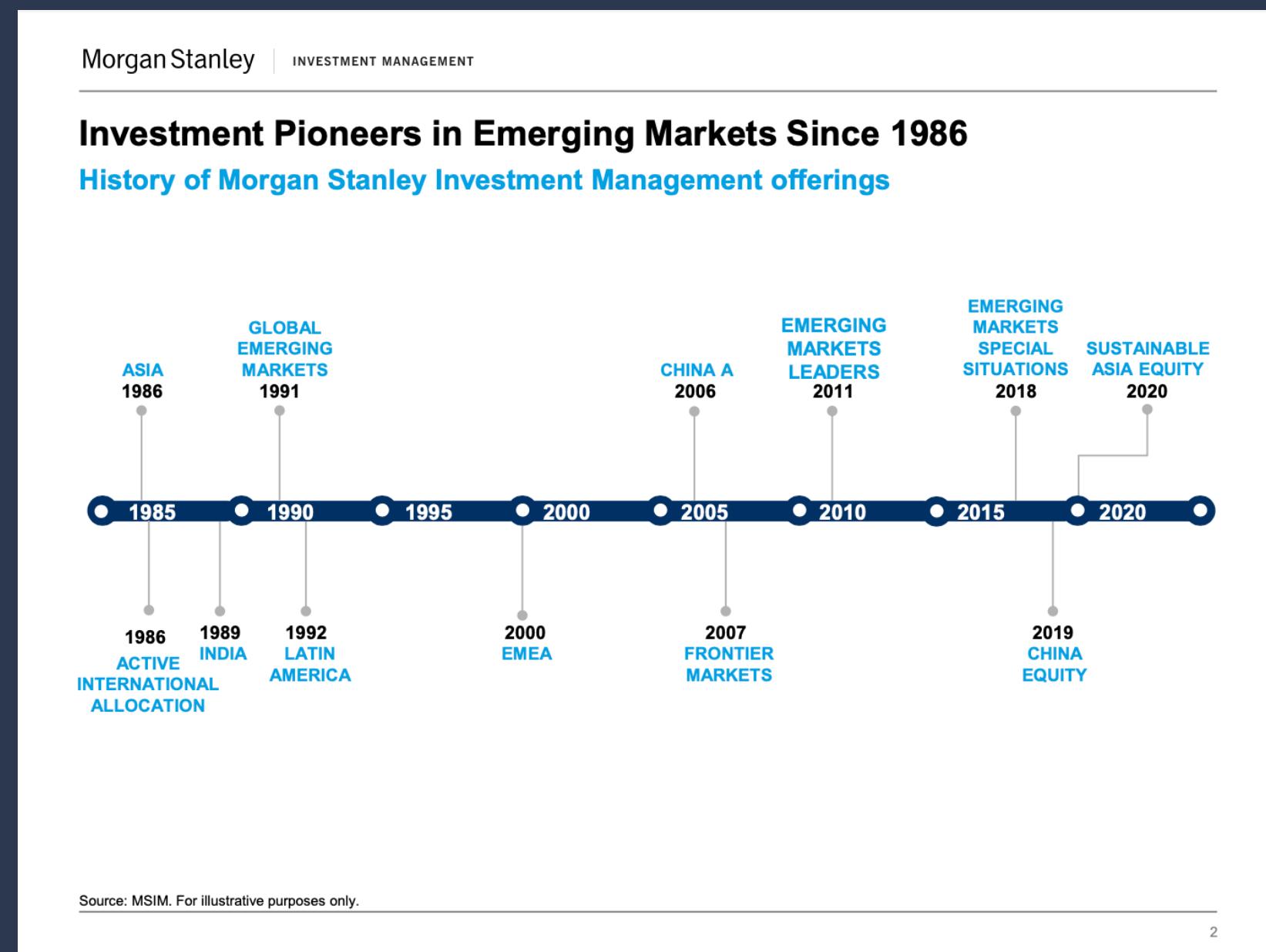
	Fund	Russell 2500 TM Index ²
Industrials	24	18
Information technology	18	14
Consumer discretionary	13	11
Health care	13	13
Financials	11	16
Real estate	9	8
Materials	8	6
Consumer staples	3	3
Communication services	1	3
- Annual Returns**: A table showing monthly, annual, and 10-year returns for different share classes.

Total returns (%)	3 Month	YTD	1Y	3Y	5Y	10Y
Institutional Shares	10.81	-17.21	-17.21	5.39	6.20	9.50
Administrator Shares	10.69	-17.33	-17.33	5.76	6.32	9.43
Class A Shares	10.66	-17.55	-17.55	4.96	5.78	9.07
Including sales charge	4.26	-22.31	-22.31	2.91	4.54	8.42
Class C Shares	10.46	-18.06	-18.06	4.35	5.09	8.44
Including sales charge	9.46	-19.06	-19.06	4.35	5.09	8.44
Class R6 Shares ³	10.81	-17.17	-17.17	5.40	6.22	9.53
Russell 2500 TM Index ²	7.43	-18.37	-18.37	5.00	5.89	10.03
Lipper Mid-Cap Core Funds ⁴	10.10	-12.57	-12.57	5.92	6.04	9.68

```
1 {"content":"FUND STRATEGY • Public equity markets are often driven by emotion, requiring successful investors to have conviction in individual securities and diversification across sectors. • Our team's conviction comes from an in-depth private market valuation (PMV, the price an acquirer would pay to purchase the entire company) process of analyzing the business model, competitive positioning, key trends, management, and other proprietary metrics. • We believe that the PMV of a company is much more stable than its associated public market stock price.", "font_size":9, "doc_title":"Common Stock Fund"}, 2 {"content":"Competitive advantages • Private market valuation (PMV) approach: By constantly measuring a company's \"private market value,\" the team is better able to assess a company's worth and act decisively when \"market emotion\" drives the price of a solid business down to discount levels. Additionally, the PMV investment process helps to discern differences between mispriced stocks and those with cheap valuations, improving the team's likelihood to generate alpha. • Opportunistic core approach: The PMV investment approach is designed to be growth- and value-neutral, with the flexibility to opportunistically invest in the best ideas at either end of the growth and value spectrum.", "font_size":12, "doc_title":"Common Stock Fund"}, 3 {"content": "Sector allocation (%)1, "table":"Sector allocation (%)1", "footer":"Sector allocation is subject to change and may have changed since the date specified. Percent total may not add to 100% due to rounding.", "data": [{"row_key": "Industrials ", "Fund": "124 ", "Russell 2500" Index2 18 :"Russell 2500" Index2 18 "}, {"row_key": "Information technology ", "Fund": "18 ", "Russell 2500" Index2 18 :"14 "}, {"row_key": "Consumer discretionary ", "Fund": "13 ", "Russell 2500" Index2 18 :"11 "}, {"row_key": "Health care ", "Fund": "13 ", "Russell 2500" Index2 18 :"13 "}, {"row_key": "Financials ", "Fund": "11 ", "Russell 2500" Index2 18 :"16 "}, {"row_key": "Real estate ", "Fund": "9 ", "Russell 2500" Index2 18 :"8 "}, {"row_key": "Materials ", "Fund": "8 ", "Russell 2500" Index2 18 :"6 "}], "row_keys": ["Industrials ", "Information technology ", "Consumer discretionary ", "Health care ", "Financials ", "Real estate ", "Materials "], "row_titles": ["Fund", "Russell 2500TM Index2"], "row_descs": ["Sector allocation is subject to change and may have changed since the date specified. Percent total may not add to 100% due to rounding."]}}, 2021-2023 JsonTool.cn, All rights reserved.
```

知识构建 – 知识质量不好咋办？

- ✓ Claude3 sonnet的视觉能力效果



Investment Pioneers in Emerging Markets Since 1986

History of Morgan Stanley Investment Management offerings

↗ ![Investment Timeline Chart](image_url_placeholder)

Year	Event
1986	ASIA
1989	INDIA
1991	GLOBAL EMERGING MARKETS
1992	LATIN AMERICA
1995	-
2000	EMEA, 2000
2006	CHINA A, 2006
2007	FRONTIER MARKETS
2011	EMERGING MARKETS LEADERS
2018	EMERGING MARKETS SPECIAL SITUATIONS
2019	CHINA EQUITY
2020	SUSTAINABLE ASIA EQUITY
1986	ACTIVE INTERNATIONAL ALLOCATION

知识构建 – 知识质量不好咋办？

- ✓ LLM合成数据进行增强

- FAQ 类数据

Question	Answer	Author(optional)	URL(optional)	...
Origin_question	Origin_A			
Synthesized_Q1	Origin_A			
Synthesized_Q2	Origin_A			
...	...			

- 文档类数据

Trigger	Content	doc_type	URL(optional)	...
Summary	Origin_Doc	NoChunk		
paragraph_1	Origin_Doc	NoChunk		
paragraph_2	Origin_Doc	NoChunk		
...		
Synthesized_Q1	Origin_Doc	NoChunk		
Synthesized_Q2	Origin_Doc	NoChunk		
...		

核心工作 – 知识召回效果差咋办？

- 有效手段
 - ✓ 多路召回：向量召回 + 倒排召回
 - ✓ BM25打分调优
 - ✓ 更优的向量模型选型
 - ✓ 多种召回范式：对称召回(Query-Question) + 非对称召回(Query-Document)
 - ✓ 微调向量模型
 - ✓ 引入Rerank模型
 - ✓ 添加IUR步骤(Incomplete utterance rewrite)
- 面临问题：
 - 正确的知识没召回，导致回答没有引用到知识
 - 召回了不正确的知识导致LLM产生了误解
- 技术总结
 1. [基于大语言模型知识问答应用落地实践 - 知识召回调优（上）](#)
 2. [基于大语言模型知识问答应用落地实践 - 知识召回调优（下）](#)

核心工作 – 知识召回效果差咋办?

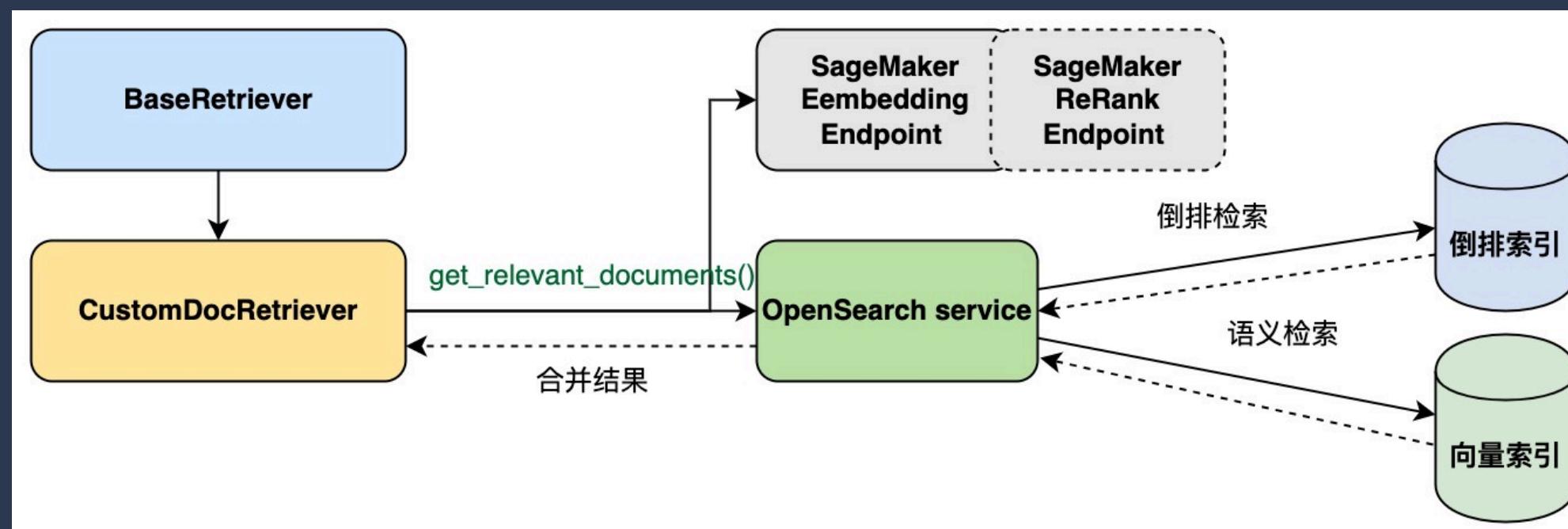
✓ 多路召回：向量召回 + 倒排召回

□ 面临问题：

- 向量模型不理解垂直领域专词
- 向量模型有时出现语义相似但主题不相似情况
- 可解释性弱，不易通过补丁解决bad case
- 倒排召回，缺乏语义信息，仅靠关键词匹配

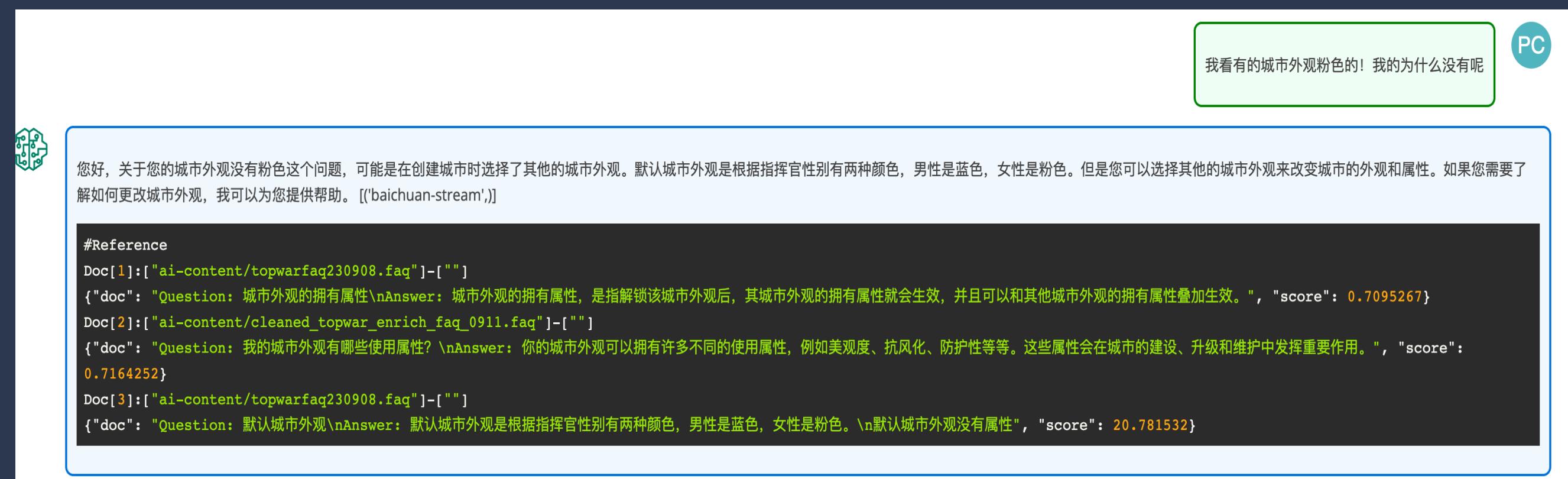
□ 工程实现：

- 重载Langchain Retriaver接口，融合多路召回结果



□ 优势样例：

用户的原始问题中“**城市外观粉色**”的信息，在向量召回的结果中并没有（前面2条 **score < 1** 的结果），向量调优难以解决该 Case，但是在第3条倒排召回（**score > 1**）的知识中，含有相关信息。



核心工作 – 知识召回效果差咋办？

✓ BM25打分调优

□ 面临问题：

- 用户的垂直数据中可能某些专词与停用词词频差不多导致IDF失真，引起得分计算有误
- 有些特定‘黑话’，数据不足时语义向量也无法解决

□ 有效手段：

- 构建停用词表，使得停用词均不参与BM25得分
- 构建同义词表，定向解决‘黑话’问题

□ 技术输出

1. blog 基于大语言模型知识问答应用落地实践 – 知识召回调优（上）

□ 视频Demo

如何调整BM25倒排优化知识召回效果

核心工作 – 知识召回效果差咋办?

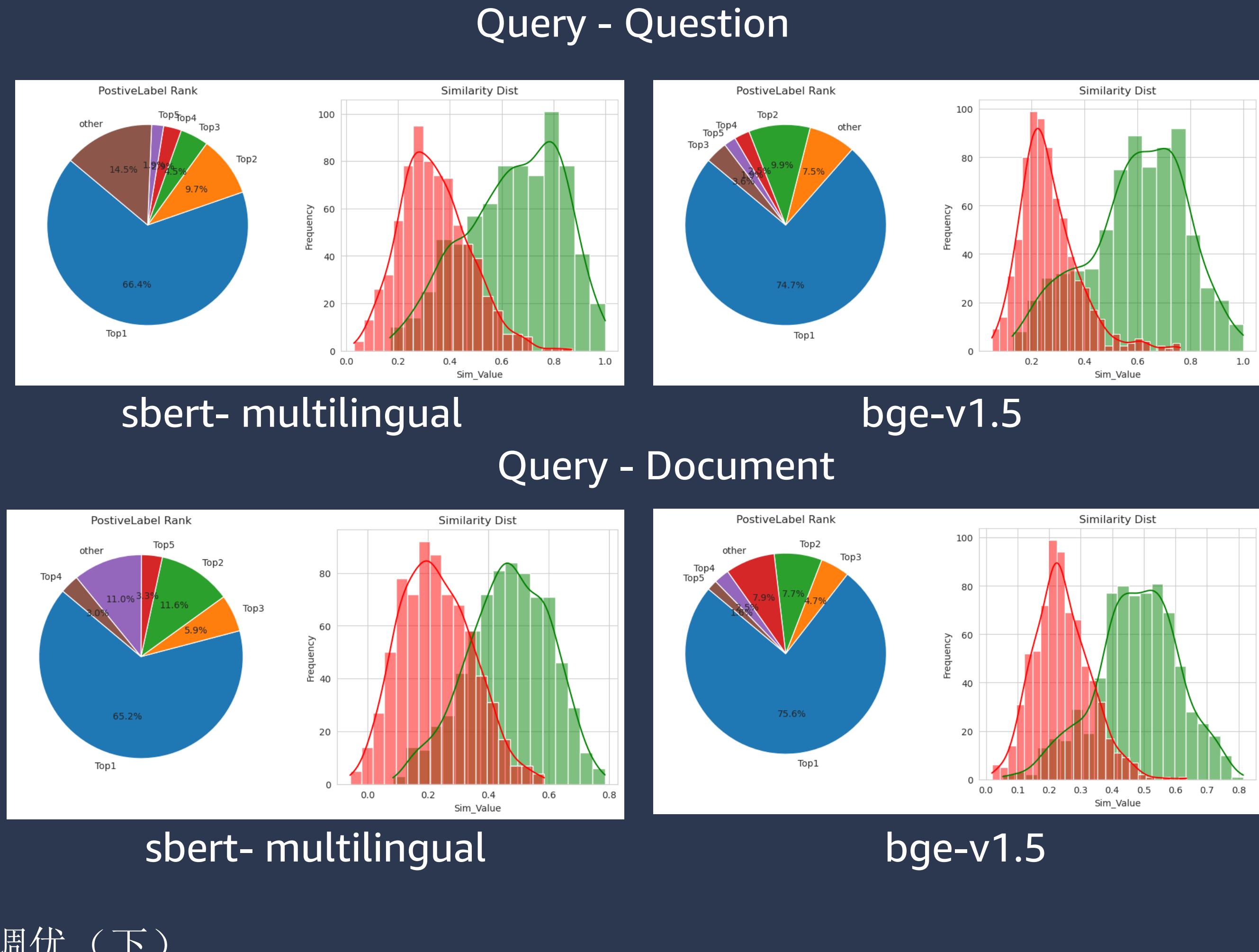
✓ 更好的向量模型选型

□ 面临问题:

- 公开数据集上的表现，在垂直领域没有特别大的参考性
- 在自己数据场景中，通过少量case手工测试无法获取全面信息，难以客观比较

□ 价值提供:

- 给出标准化的评估方式和可视化方法



□ 基本结论:

- 优选bge-large-zh-v1.5
- 优选bge-large-en-v1.5

□ 技术输出

1. 代码实现 [bge_zh_research.ipynb](#)
2. Blog 基于大语言模型知识问答应用落地实践 – 知识召回调优（下）

核心工作 – 知识召回效果差咋办?

✓ 多种召回范式： 对称召回 (Query-Question) + 非对称召回(Query-Document)

□ 面临问题：

○ 两种形式各有弊端

- 对于垂直领域做QD召回需要向量模型具备很强的理解能力，需要用这个领域数据的训练过
- 用户query中的一些信息只出现在知识的Document/Answer中，通过Query-Question匹配难度大

□ 例子：

用户的发问角度，或者query-Question的语义相似性不一定高

□ 视频Demo

知识问答中的对称召回+非对称召回策略

```
1 Question: AWS Clean Rooms的数据源必须在AWS上么?  
2 Answer: 对，目前必须在AWS上，而且必须是同一个region。  
3 ======  
4  
5 user : Clean Rooms的数据源可以不在同一个region么?
```

核心工作 – 知识召回效果差咋办?

✓ 向量模型微调

□ 面临问题:

- 场景数据过于垂直，通用的模型表现不佳

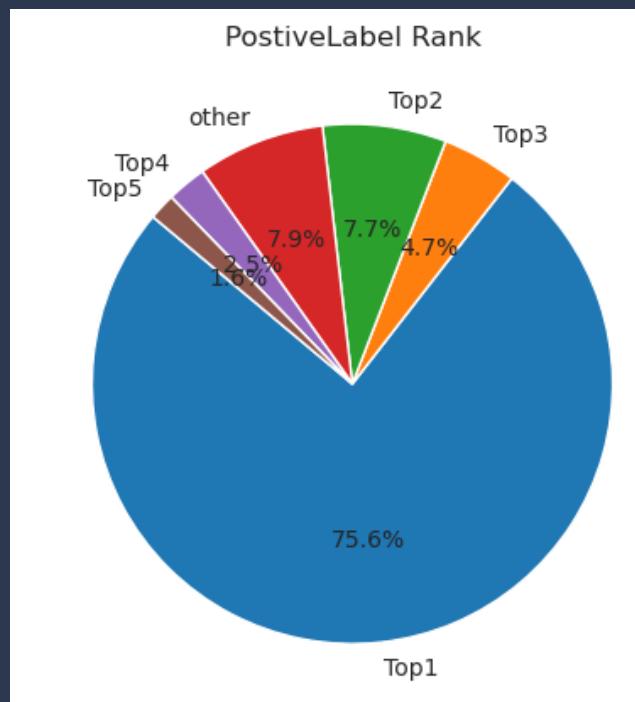
□ 价值解读:

- 在训练集上效果非常好，意味着后续可以通过持续收集用户反馈，并纳入到训练集以更新模型，使得这个效果不断扩大覆盖范围。
- 测试集上效果没有下降，反而有小幅提升，意味着训练没有破坏模型原有语义能力，对于未被训练集覆盖到的场景，模型仍能以优于原模型的性能进行服务

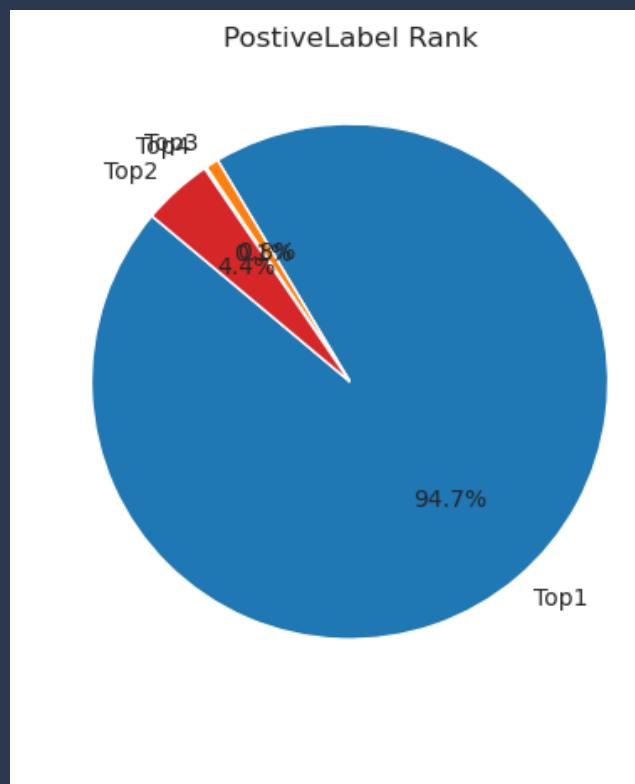
□ 技术输出

1. 代码实现 [bge_zh_research.ipynb](#)，包含训练数据构造，训练部署
2. [blog 基于大语言模型知识问答应用落地实践 – 知识召回调优（下）](#)

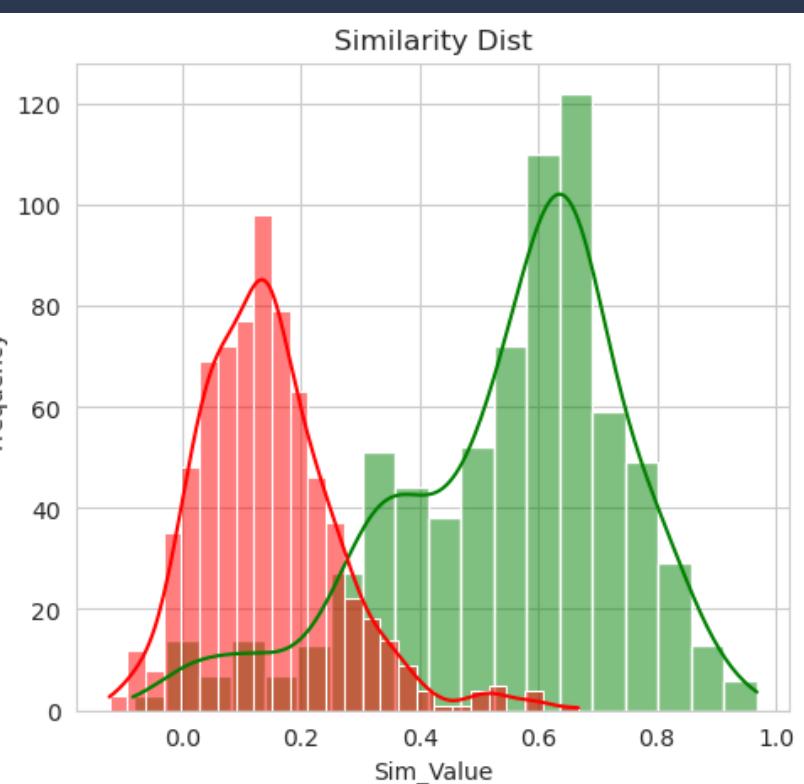
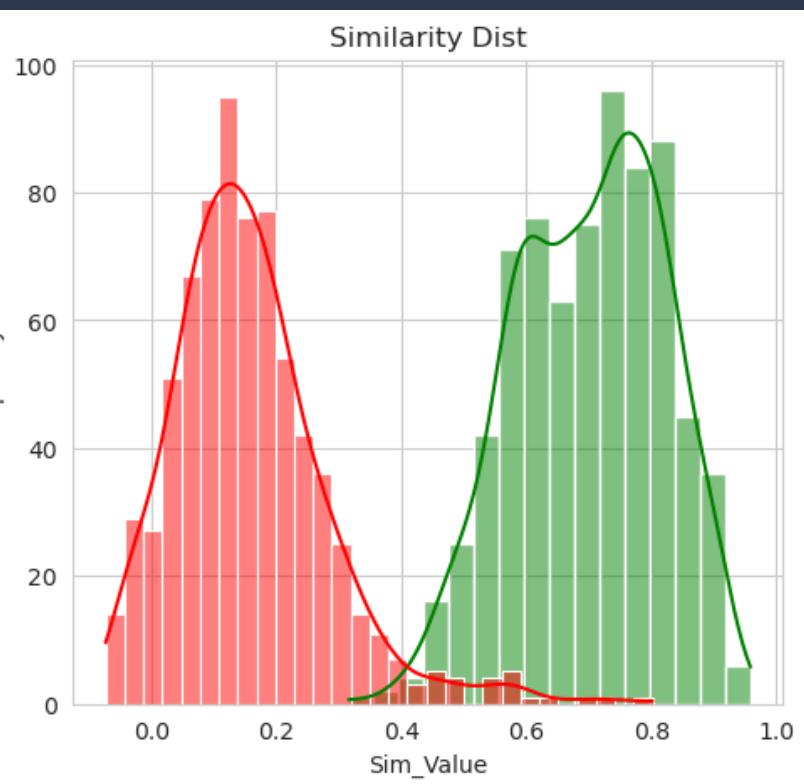
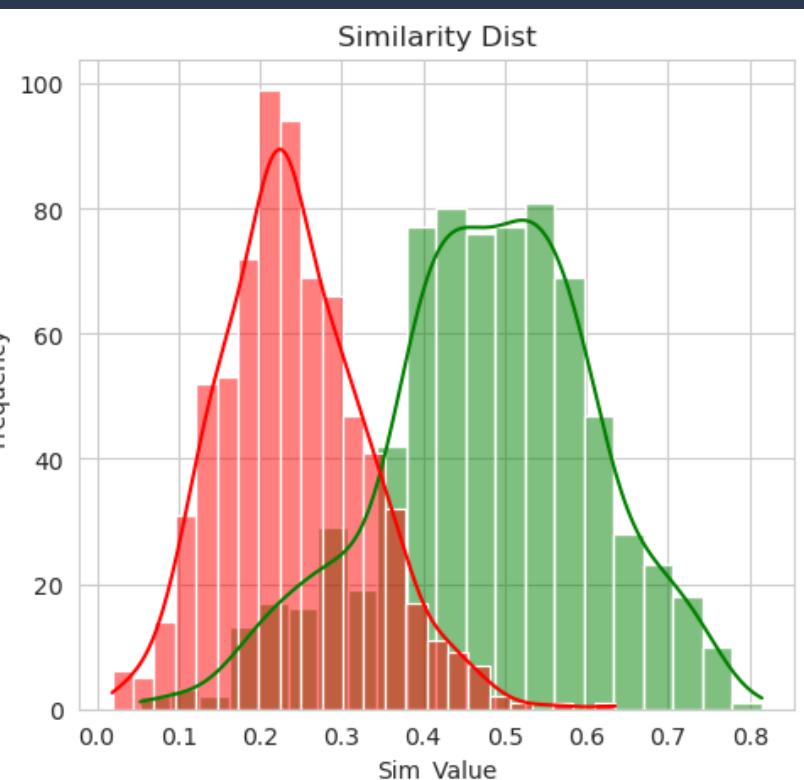
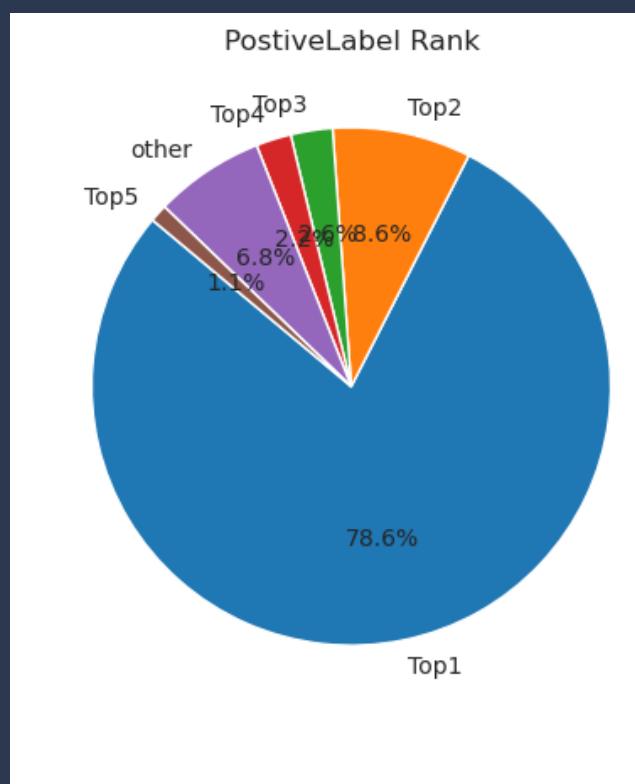
原始



微调后训练集



微调后测试集



核心工作 – 知识召回效果差咋办？

✓ 引入Rerank模型

□ 面临问题：

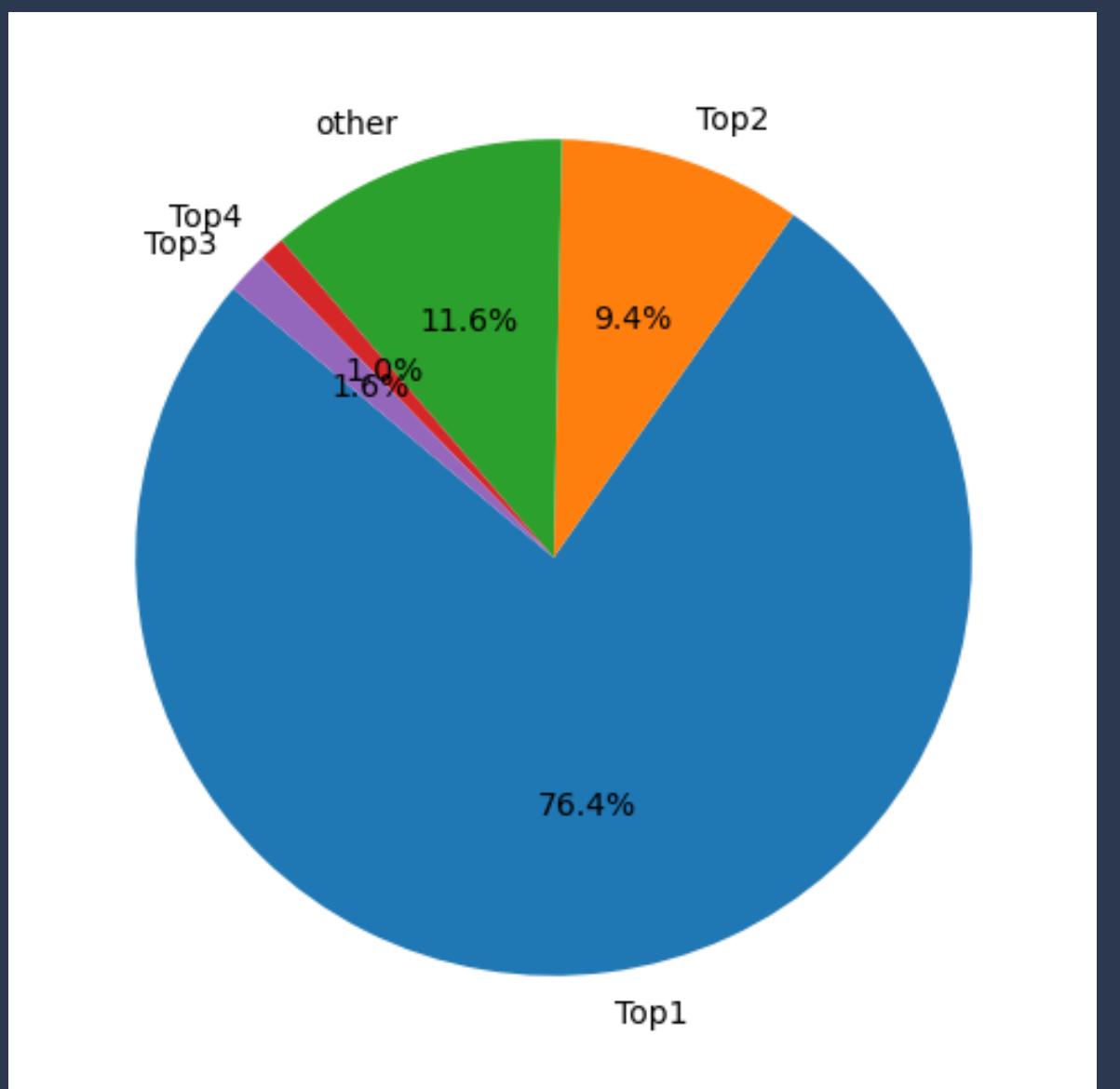
- 向量召回与倒排召回的评分体系不一致，只能随便各取TopK，缺乏依据

□ 价值解读：

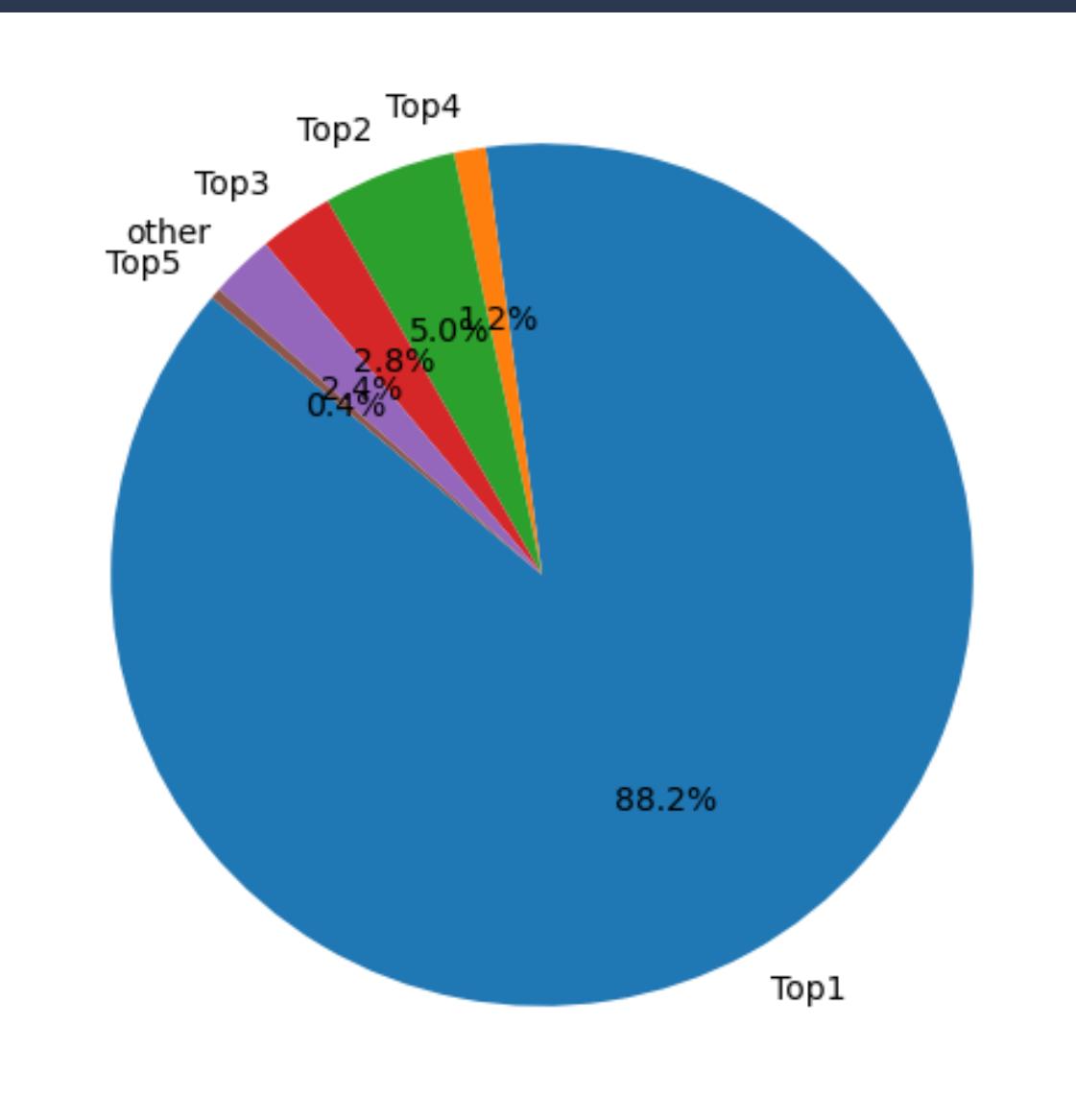
- 解决两路合并各取多少的问题。
- 进一步提高知识召回质量，作为一个独立插拔模块解决向量模型微调解决不了的问题

□ 技术输出

1. 代码实现 [bge_zh_research.ipynb](#)，包含难样本挖掘，训练部署等
2. [blog 基于大语言模型知识问答应用落地实践 – 知识召回优（下）](#)



无Rerank的正例排名



有Rerank的正例排名

核心工作 – 知识召回效果差咋办？

✓ 添加IUR步骤(Incomplete utterance rewrite)

□ 面临问题：

- 在多轮对话情形下，用户的当前输入会存在一些隐含的指代关系和信息省略。缺乏上下文信息的语义缺失严重无法有效召回

□ 解决思路：

- 利用LLM进行当前query的重写，对上下文隐含信息重新纳入到新生成的query中。重写效果好，但多调用一次LLM，会加重全流程latency问题
- 部署一个独立的IUR模型。重写效果没有前者好，但收集到数据后，可以基于采集数据进行微调，更加适应特定场景。

□ 技术输出

1. 参考代码 [query_rewrite.py](#)
2. IUR 可用模型及部署方法 <https://huggingface.co/csdc-atl/dialogue-rewriter>

□ 例子：

<history>

User: "有戴森的吹风机吗？"

Bot: "没有哦亲亲"

User: "戴森都没有"

Bot: "不好意思，看看其他品牌呢"

</history>

<query>"那有松下的吗"</query>

<rewrite>有松下品牌的吹风机吗?</rewrite>

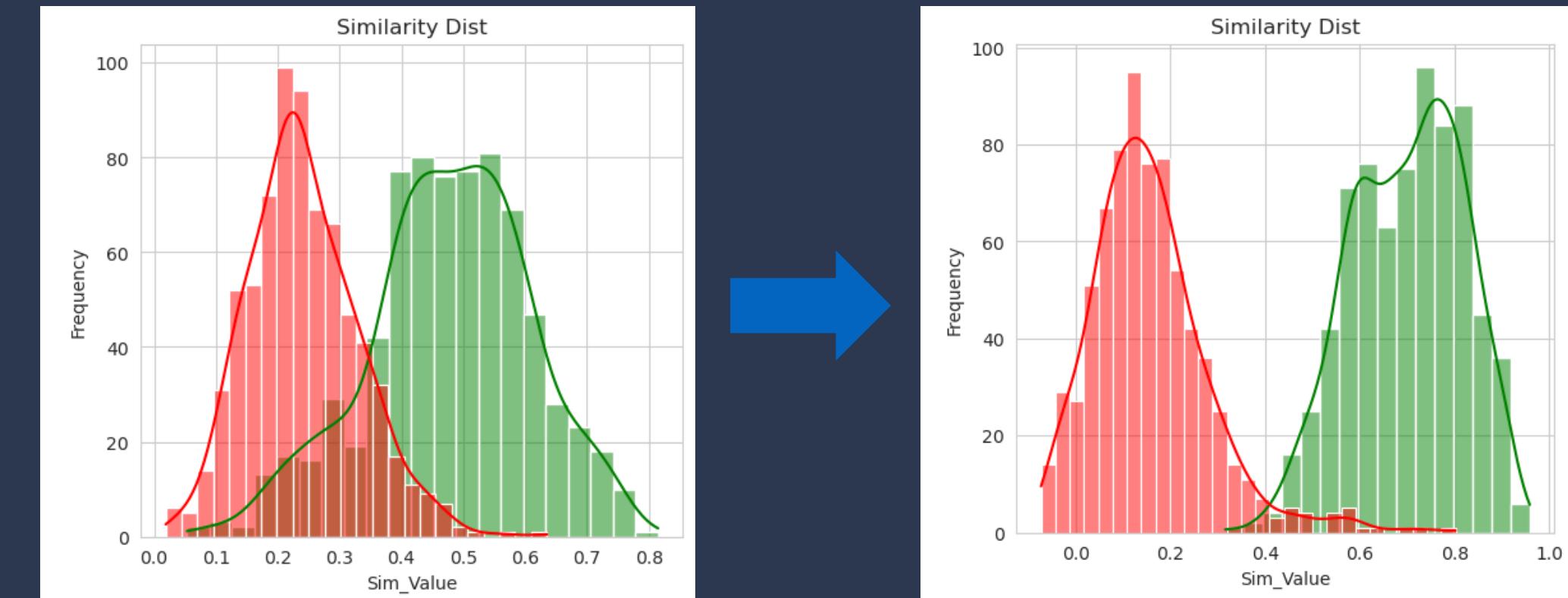
核心工作 – LLM幻觉咋办？怎么判断超出知识库范围？

✓ 根据多级召回阈值采取灵活降级策略并结合意图识别

□ 面临问题：

- 由于LLM幻觉编造一些内容可能误导用户，在某些情况下造成负面影响非常大（跟钱，账单相关的）
- 通过Prompt提示LLM，要它依据知识不要胡说的方法是不靠谱
- 用户可能还会问一些不在服务范围的话题，滥用服务引起GPU浪费。

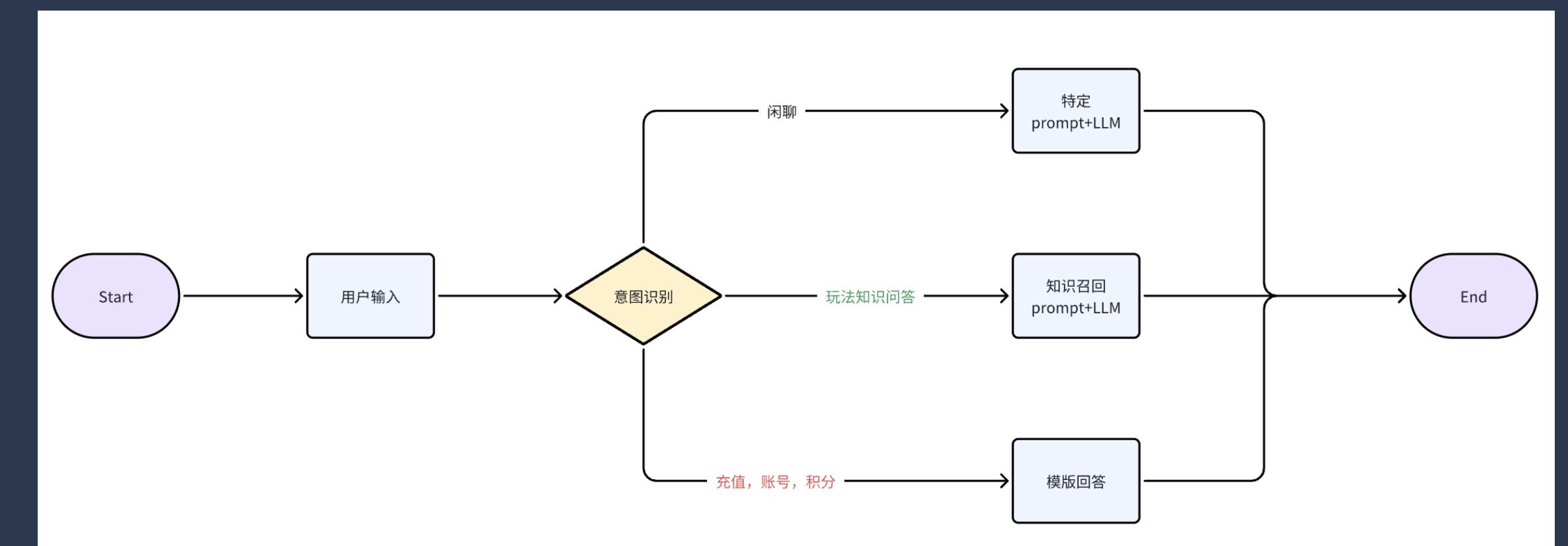
拉开值域分布



□ 有效手段：

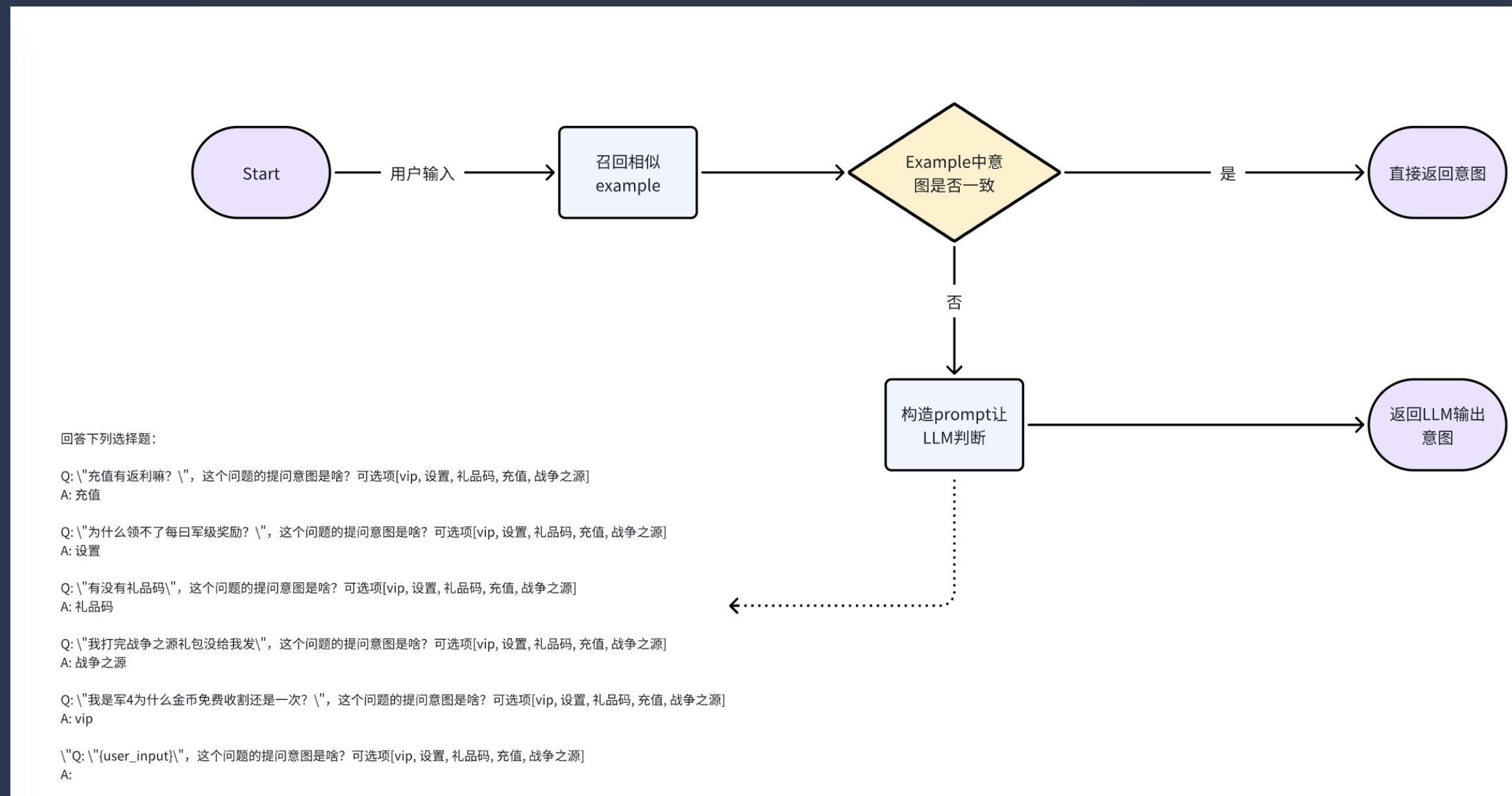
- 通过收集反馈数据，微调向量模型叠加微调Rerank模型，拉开相关召回和非相关召回的得分的值域分布。得分分为多级采用不同策略
 - 最置信的走LLM
 - 次置信的提示LLM如果不相关进行拒答
 - 不置信的仅返回召回TopK或直接拒答
- 意图识别进行场景分流，敏感场景避免LLM介入直接走预制答案。

意图识别避免敏感场景幻觉的示意图



核心工作 – 意图识别模块

□ 实现方式



□ 优势特点

- 简单易用，不太需要太多算法能力
- 随着积累example越多越准，越不需要LLM参与，性能越好
- 生活场景的用户输入比较易用，非生活场景可能需要向量微调

□ 实践检验

采用bge_large_en向量模型 + Bedrock Claude大模型结合这个技术方案。

在一个IOT领域的的意图识别场景的POC中，把之前基于OpenAI的准确率从80%+提升到90%+

□ 技术输出

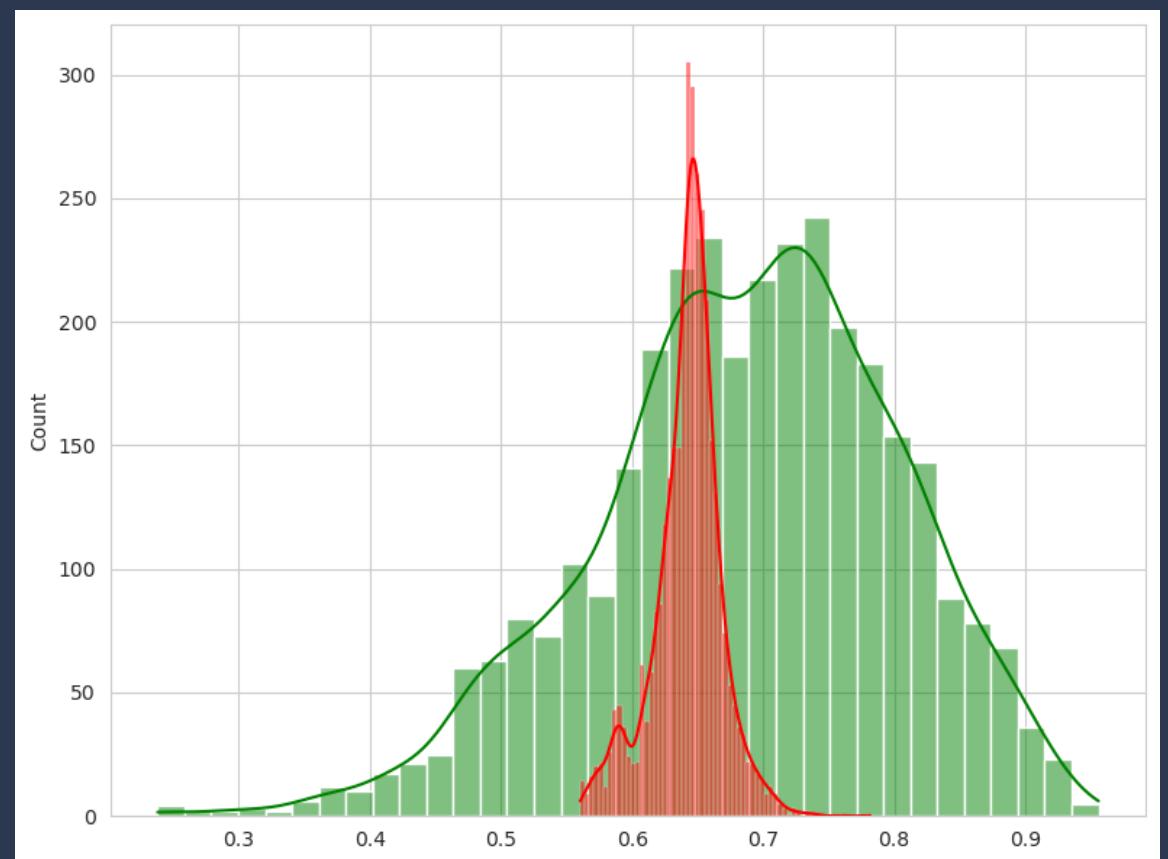
1. 使用说明 [README.md](#)
2. 参考代码 [intention.py](#)

□ 视频Demo

[知识问答中的意图识别方案](#)

核心工作 – 如何找出那个阈值?

- 实现步骤
 - 收集所有的用户正负反馈，找出应该拒答的query集合
 - 计算这些query和知识库内所有知识的相似分，统计其分布，比如可参考下图
设定多级阈值(红色为拒答query的相似分分布)



- 可见区分度不足，如果上图是Rerank分数则微调Rerank模型，如果是向量模型分则微调向量模型

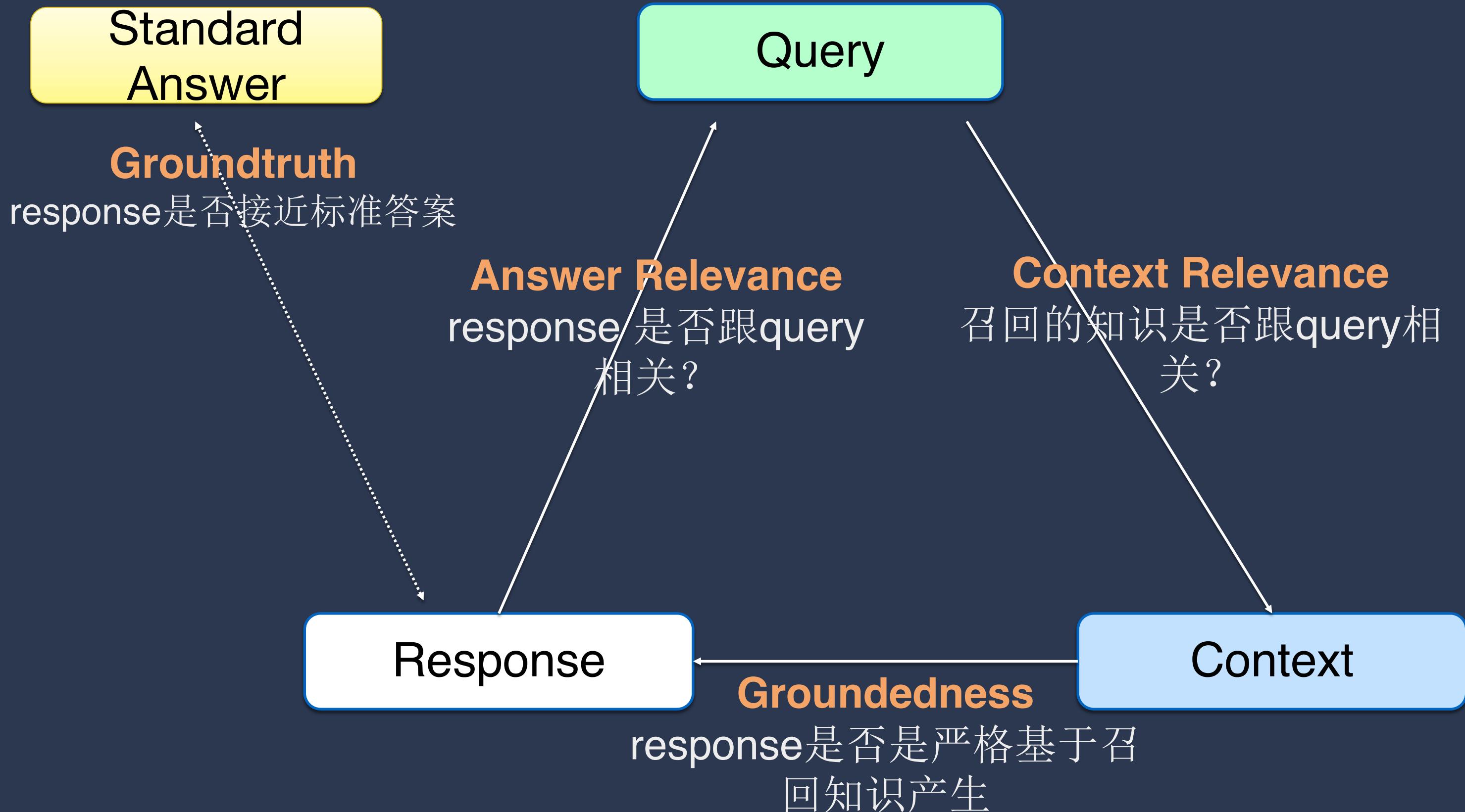
- 技术输出
 1. 参考代码 [HowToFindThreshold.ipynb](#)
 2. Blog 基于大语言模型知识问答应用落地实践 – 知识召回调优（下）

□ 视频Demo
[知识问答中拒答策略及实现方式](#)

上线必备 – 效果如何评估和持续监控

✓ 维护评估集 + 量化评估框架

测试另一个版本的提示词模板是否有改进?



Template v1 vs v2

Template v1 vs v2

Ground Truth	Groundedness	Context Relevance	Answer Relevance
0.9	0.5	0.2	1
0.9	1	1	1
1	1	1	1
0.7	0.525	1	1
0.9	1	1	1
0.9	0.9	1	1
0.9	1	1	1
0.9	1	1	1
0.9	1	1	1
0.9	0.5	1	1
0.9	0.75	1	1
0.9	0.8333333333	0.8	1
0.9	1	1	1
0.7	0.75	1	1
0.2	0.75	1	1
1	1	1	1
1	1	1	1
1	1	1	1
0.9	1	0.4	1
0.9	1	1	1
1	1	1	1
0.9	1	1	1
0.3	1	1	1
0.9	0.75	1	1
1	1	1	1
0.9	0.2	1	1
0.9	1	1	1
1	1	1	1
1	1	1	1
0.9	1	1	1
0.9	0.8571428571	1	1
0.9	1	1	1
0.9	0.8	1	1

Ground Truth Groundedness Context Relevance Answer Relevance

Template v1 Template v2

结论: v2的提示词模板反而使回答质量下降了

上线必备 – BadCase 怎么排查?

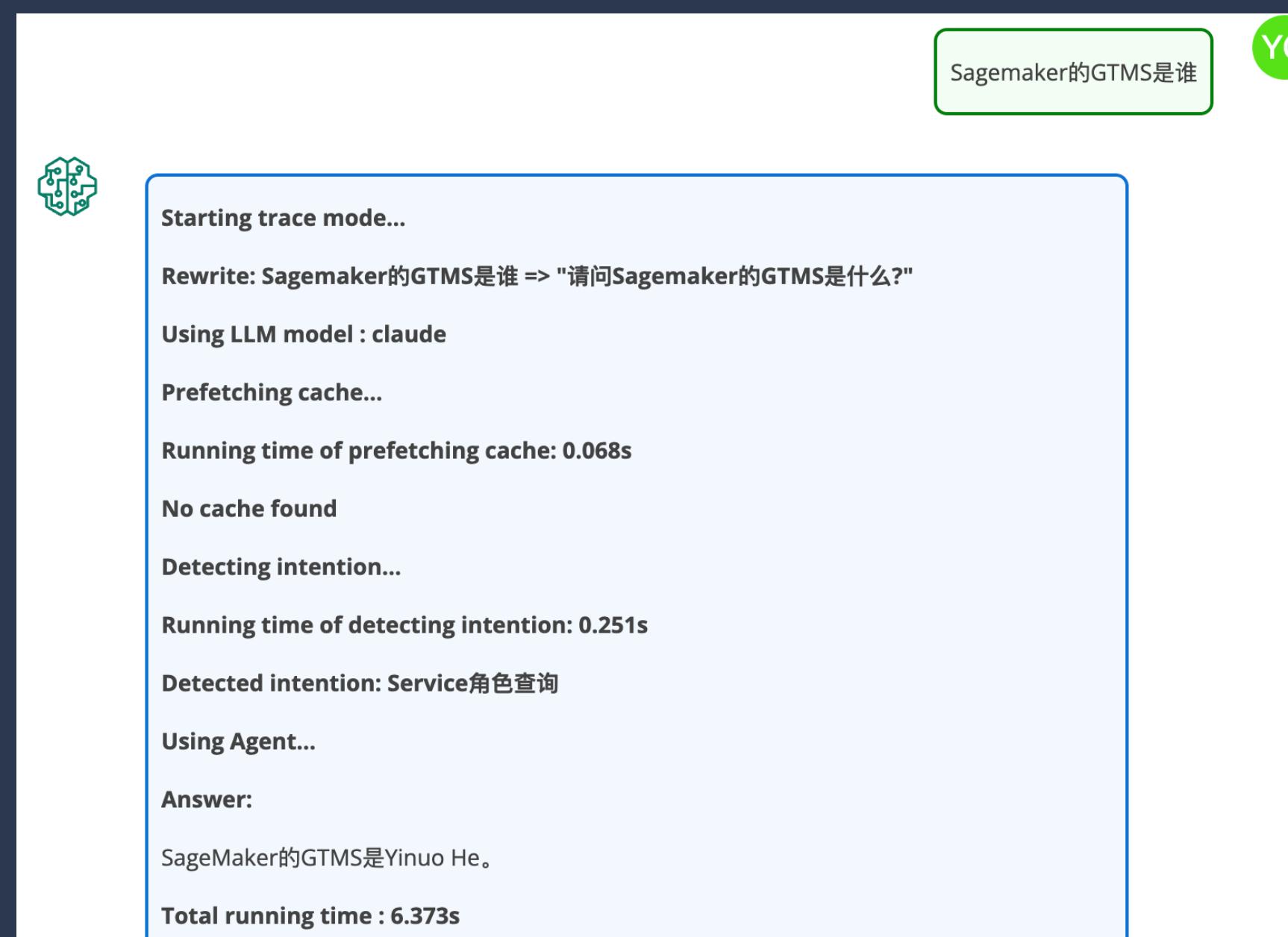
✓ 全流程日志输出并自动回流 + AOS 做日志检索 + 前端信息显示

□ 面临问题:

- RAG项目不是一蹴而就的，需要长期的优化调优
- 链路中多阶段的中间结果需要独立调优

□ 前端信息显示:

- 主要用于快速了解召回质量，提升排查速度



LLM 版本
用户的query

识别的意图 构造的Prompt 各路召回的知识&打分

上线必备 – BadCase 怎么排查?

✓ 通过用户在线反馈，对知识库进行修正，补充

□ 提供用户反馈搜集接口，对反馈问题进行排查。用户纠正过的答案可以作为新的FAQ知识，补充进知识库

对话

GPU 内存限制会导致什么瓶颈 AN

根据context内容可以知道,当训练模型时,如果模型过大无法适应GPU的内存限制,会导致以下几个瓶颈:

- 1. 无法使用更大的模型,模型规模受限。
- 2. 无法增加批次大小利用GPU计算能力,训练效率受限。

为了解决这个问题,SageMaker提供了几种技术:

- 1. 分片数据并行技术,可以将模型状态分片到多个GPU上,降低每个GPU的内存占用。
- 2. 激活检查点技术,通过清除和重新计算部分层的激活来减少内存使用。
- 3. 激活卸载技术,可以进一步减少内存使用。
- 4. 张量并行技术,将模型拆分到多个GPU上执行,降低每个GPU的计算负载。

利用这些技术可以解决GPU内存限制带来的模型规模和批次大小限制问题,实现更高效的大模型训练。

帮我纠正

AWS Chat Portal > 反馈管理

反馈管理 (1/5+)

Filter by text, property or value

Question	Original Answer	New Answer	Feedback
怎么看现有的 Capacity?	根据context内容,可以使用 https://ec2-baywatch-prod-iad.iad.proxy.amazon.com/pages/poolViewer 查看现有的Capacity。 Tips:通常情况下使用默认值搜索即可。	sss	thumbs-up

根据context内容

训练模型时, GPU 内存限制会有成为哪些瓶颈

My customer is requesting current generation capacity in BJS. Can you help to providing a formal response?

根据context内容,您可以使用以下响应为您的客户提供正式答复: "Regarding your inquiry about current generation capacity in BJS: Unfortunately, there are rare occasions when demand in certain AZs outpaces our ability to provide capacity for every instance family. When this occurs, we restrict new instance types in these locations so that we can ensure the highest possible standards of continuity and scalability for all of our customers currently operating in the Availability Zone. We hope that this policy provides assurance that your existing workloads in BJS should expect the same elasticity and availability that EC2 has historically provided. If you are interested in current generation capacity, we encourage you to review options for launching instances in another one of the Region's AZs."

根据context内容,申请GPU FOOB的流程如下: 1. FOOB ticket提交模板为: "" GPU实例需求

帮助纠正

感谢您的反馈

Your answer

取消 提交

所以总的来说, GPU内存限制是训练大型模型的一个重要瓶颈,上述一些技术可以很好地解决这个问题。

经验洞察 – 反问机制设计?

✓ 通过知识的元数据来提示LLM

□ 面临问题:

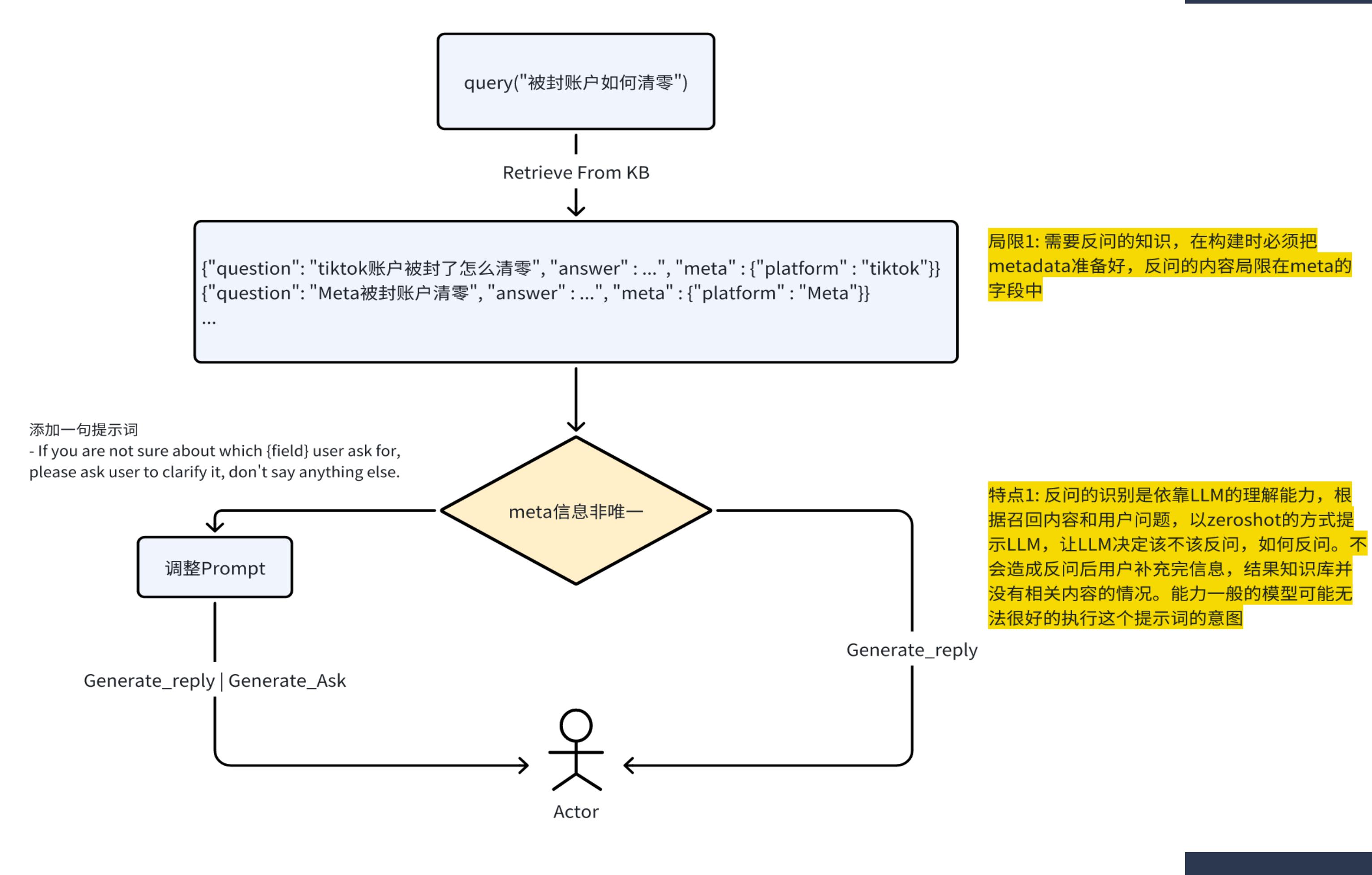
○ 如何判断是否需要反问?

LLM缺乏特定业务的理解，LLM在训练时的默认状态是给出一个泛泛的回答或拒答，反问较少。需要明确的提示词引导他进行反问。

○ 如何知道该反问什么?

只有明确的反问方向，用户才能知道下一步如何做

目 愚路1



经验洞察 – Agent API设计思考

✓ API 提供LLM易解读的报错信息

□ 面临问题：

- 用户的自然语言不可能很精确
- 结构化查询的字段值可能会存在错误

□ 例子：

- 用户的问题一般是：“美西2的g5.2x什么价格？”

但接口的schema如下表，存在多个字段才能定位price

region	price	instance_type	term	purchase_option
us-west-2	1.052\$	g5.2xlarge	On_Demand	All Upfront
ap-southeast-1	1.052\$	g4dn.2xlarge	Reserved	All Upfront

- 用户的问题可能有输入错误，比如“lex的产品经理是谁？”，注意“lex != Lex”

employee	role	domain	scope
Sofia	Product Manager	AIML	Lex
Jason	Tech	AIML	Lex

□ 手段：

- 查询接口的字段约束尽可能宽松, 单条件>复合条件
- 返回条数过多再叠加条件，或者则提示反问，并给出反问方向
- 返回条数过少，则进行近似查询(比如按照编辑距离/同义词映射)，把可能的取值通过接口返回。

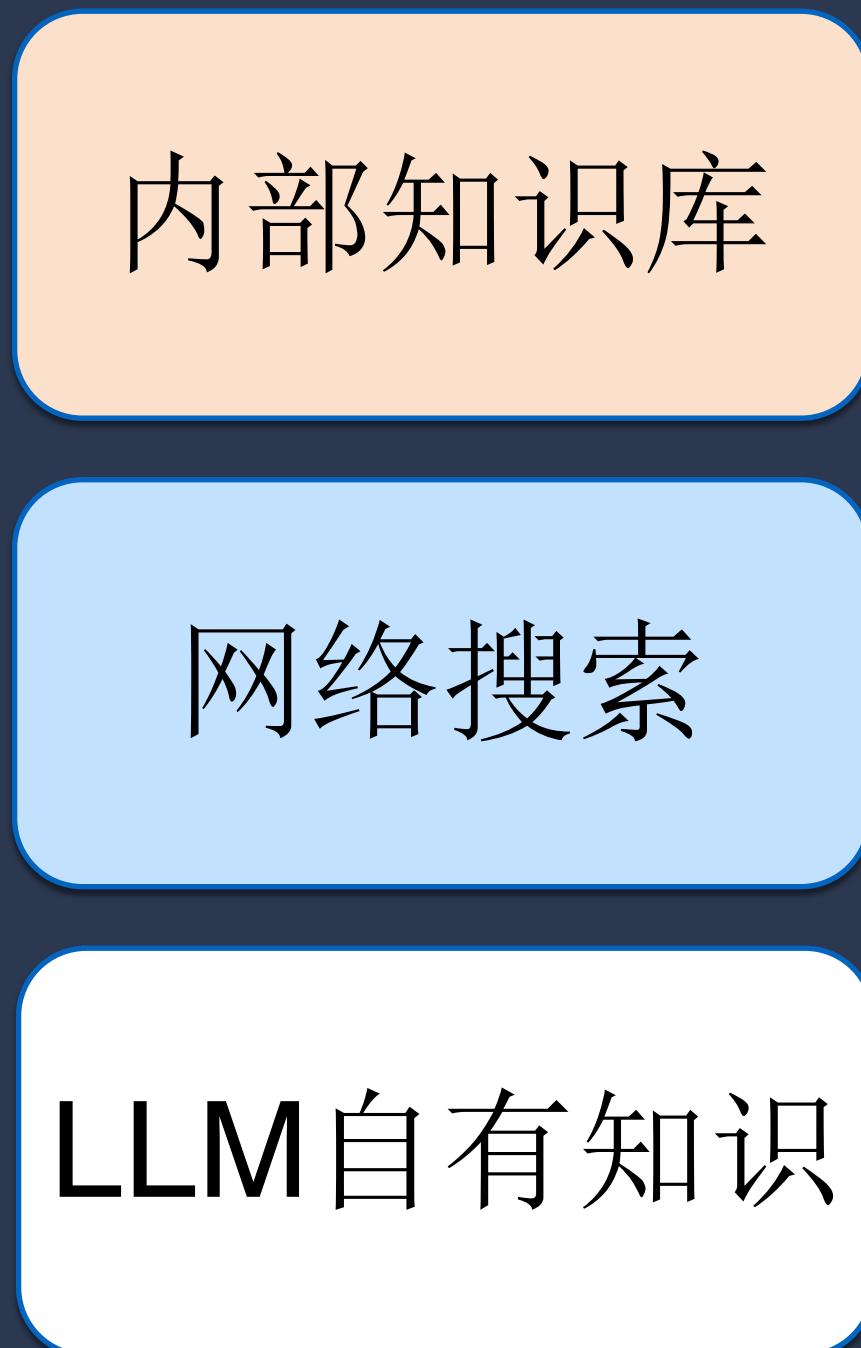
The screenshot shows a tracing interface for an AI system. At the top right, there is a green button labeled 'AN' and a text input field containing 'lex的问题该找谁?'. The main area displays a log of the system's operations:

- Starting trace mode...
- Using LLM model : claude-v2
- Prefetching cache...
- Running time of prefetching cache: 0.130s
- No cache found
- Detecting intention...
- Running time of detecting: 1.588s
- Detected intention: get_contact
- Using Agent...
- Function call result:
- 本次回答基于使用工具[get_contact]为您查询到结果:
- {"statusCode": 404, "message": "Can't find relevant information by - lex.", "suggested_question": "Lex的问题该找谁?"}
- Answer:
- 很抱歉,我无法根据您的问题"lex的问题该找谁?"找到任何相关信息。API建议您可以尝试询问"Lex的问题该找谁?"。
- Total running time : 3.674s

经验洞察 – 自建知识库与Web Search知识的集成

✓ 强调Rerank模型的作用，利用它的score来判断是否要引入网络信息

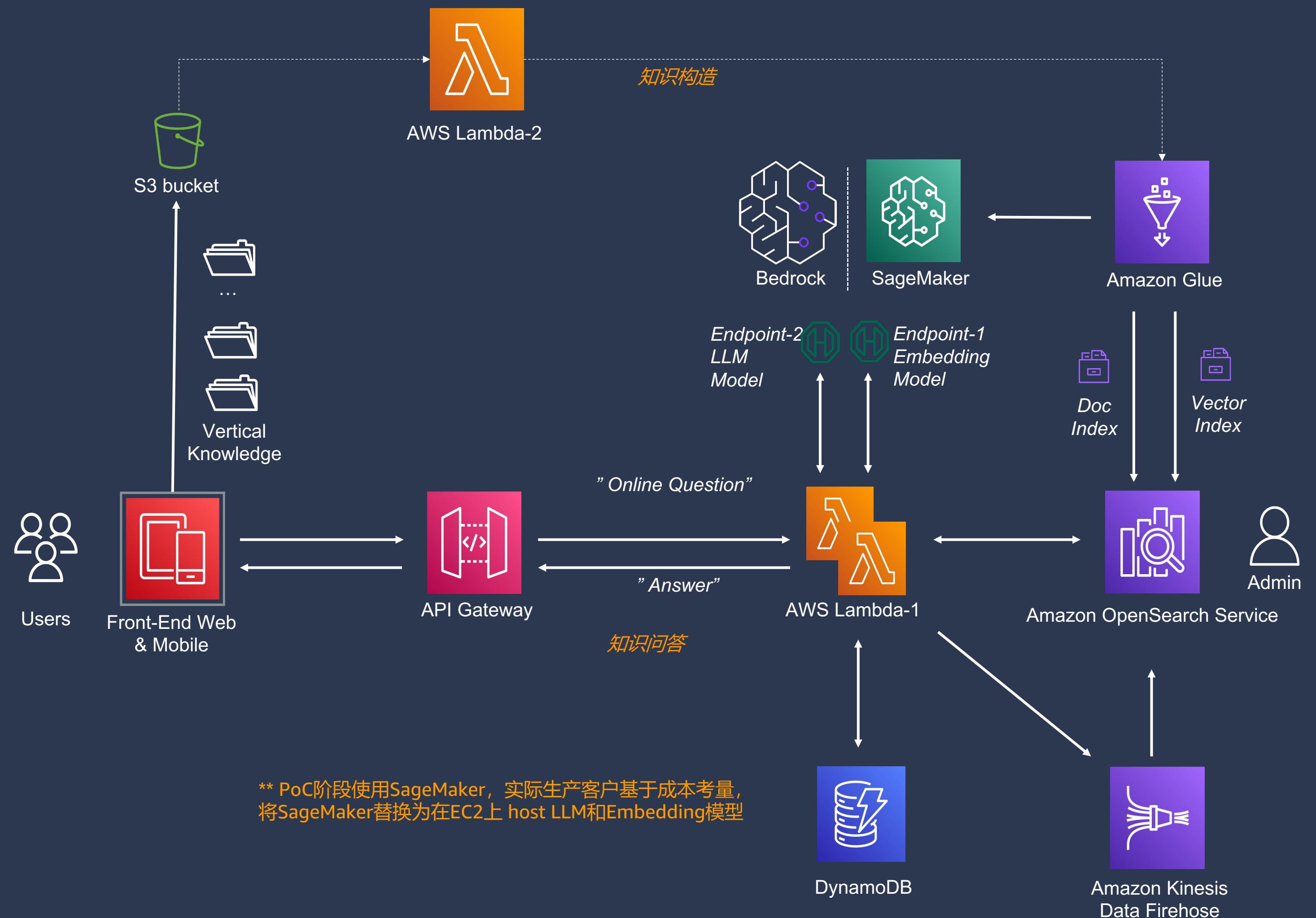
Rerank Score



RAG场景中的亚马逊云产品亮点

建议的整体技术方案

- LLM结合私有知识库，保证知识内容的准确性和时效性
- 方案中的组件可灵活插拔，可根据业务需求可选择不同的大语言和embedding模型，例如Bedrock
- 采用全托管服务或serverless，有效降低用户管理和运维成本
- 基于CDK快速部署



Bedrock 在RAG上的优势

- ✓ Claude-V3 在RAG的外部评测中表现出色

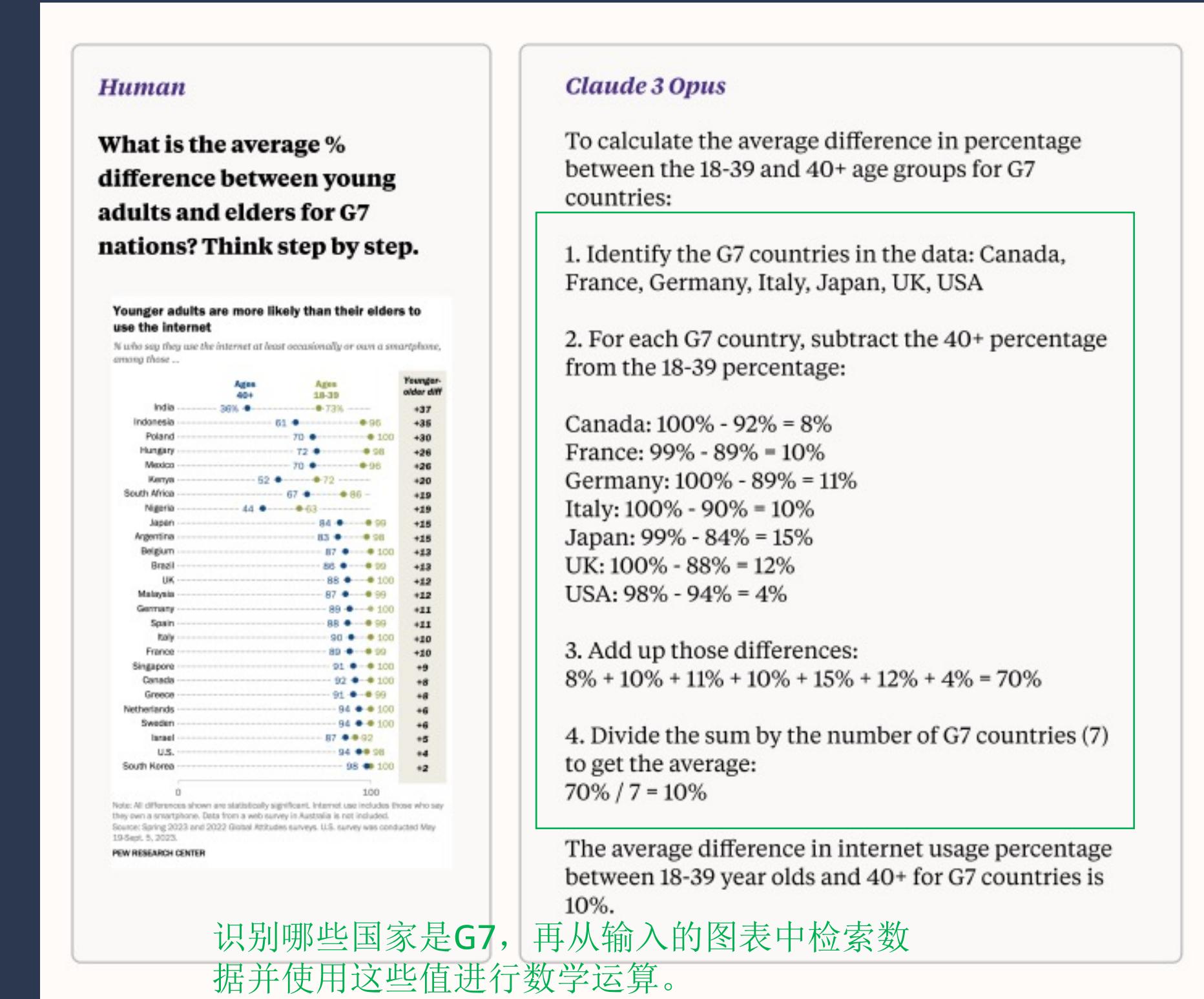
Results:						
RANK	LLM	PASS	FAIL	ACCURACY [%]	COST	TIME
1	claude-3-opus-20240229	124	0	100	18.5161	10407.1
2	gpt-4-1106-preview	122	2	98.3871	9.26881	10017.5
3	claude-2.1	120	4	96.7742	7.57311	10544.2
4	mistral-large-latest	119	5	95.9677	8.7707	8799.88
5	claude-3-sonnet-20240229	119	5	95.9677	3.67315	9823.09
6	mistral-small-latest	118	6	95.1613	2.3335	8960.92
7	mistral-medium	117	7	94.3548	2.91945	9782.42
8	mistralai/Mixtral-8x7B-Instruct-v0.1	115	9	92.7419	0.877247	9431.38
9	claude-3-haiku-20240307	115	9	92.7419	1.20891	9839.28
10	gpt-35-turbo-1106	113	11	91.129	0.917839	8856.65
11	gpt-3.5-turbo-16k-0613	111	13	89.5161	2.74251	8789.65
12	mistralai/Mistral-7B-Instruct-v0.2	111	13	89.5161	0.175633	9010.14
13	NousResearch/Nous-Capybara-34B	110	14	88.7097	0.314923	10063.2
14	gpt-3.5-turbo-0613	109	15	87.9032	0.57664	8294.3
15	openchat/openchat-3.5-1210	109	15	87.9032	0.124055	8399.51
16	mixtral-8x7b-32768	109	15	87.9032	0.312924	8985.34
17	gemini-pro	107	17	86.2903	0.887542	9309.42
18	h2oai/h2ogpt-4096-llama2-70b-chat	105	19	84.6774	1.9228	8600.32
19	mistral-tiny	105	19	84.6774	0.289585	8823.53
20	h2oai/h2ogpt-4096-llama2-13b-chat	99	25	79.8387	0.478241	8458.29
21	h2oai/h2ogpt-32k-codellama-34b-instruct	99	25	79.8387	0.907811	9248.13
22	google/gemma-7b-it	86	38	69.3548	0.123531	8517.83
23	h2oai/h2o-danube-1.8b-chat	72	52	58.0645	0.0627042	8652.49

Bedrock 在RAG上的优势

✓ Claude-V3 视觉能力 - 知识的整理和解析

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4V	Gemini 1.0 Ultra	Gemini 1.0 Pro
Math & reasoning MMMU (val)	59.4%	53.1%	50.2%	56.8%	59.4%	47.9%
Document visual Q&A ANLS score, test	89.3%	89.5%	88.8%	88.4%	90.9%	88.1%
Math MathVista (testmini)	50.5% CoT	47.9% CoT	46.4% CoT	49.9%	53.0%	45.2%
Science diagrams AI2D, test	88.1%	88.7%	86.7%	78.2%	79.5%	73.9%
Chart Q&A Relaxed accuracy (test)	80.8% 0-shot CoT	81.1% 0-shot CoT	81.7% 0-shot CoT	78.5% 4-shot CoT	80.8%	74.1%

Claude 3系列模型具备与其他领先模型相媲美的高级视觉识别能力。支持处理各种格式(JPEG, PNG, GIF, WebP)多种类型类型(照片, 图表, 技术绘图等)的图片。



识别哪些国家是**G7**, 再从输入的图表中检索数据并使用这些值进行数学运算。

Prompt: 输入一张来自皮尤研究中心图表"年轻人比长辈更有可能使用互联网", 然后询问"G7国家的年轻人和老年人之间的平均差异百分比是多少? 请一步步思考"

Bedrock 在RAG上的优势

✓ Claude-V3 多语言的理解能力

多语言版GSM(基础数学), MMLU(本科专业知识), 均取得优秀成绩。 Claude3 Opus 在 MGSM 达到SOTA

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4 ³	Gemini Ultra ⁴	Gemini Pro 1.5 ⁴	Gemini Pro 1 ⁴	
MGSM (Multilingual Math)	8-shot 0-shot	90.5% 90.7%	83.7% 83.5%	76.5% 75.1%	74.5% —	79% —	88.7% —	63.5% —

Table 4 This table shows evaluation results on the multilingual math reasoning benchmark MGSM.

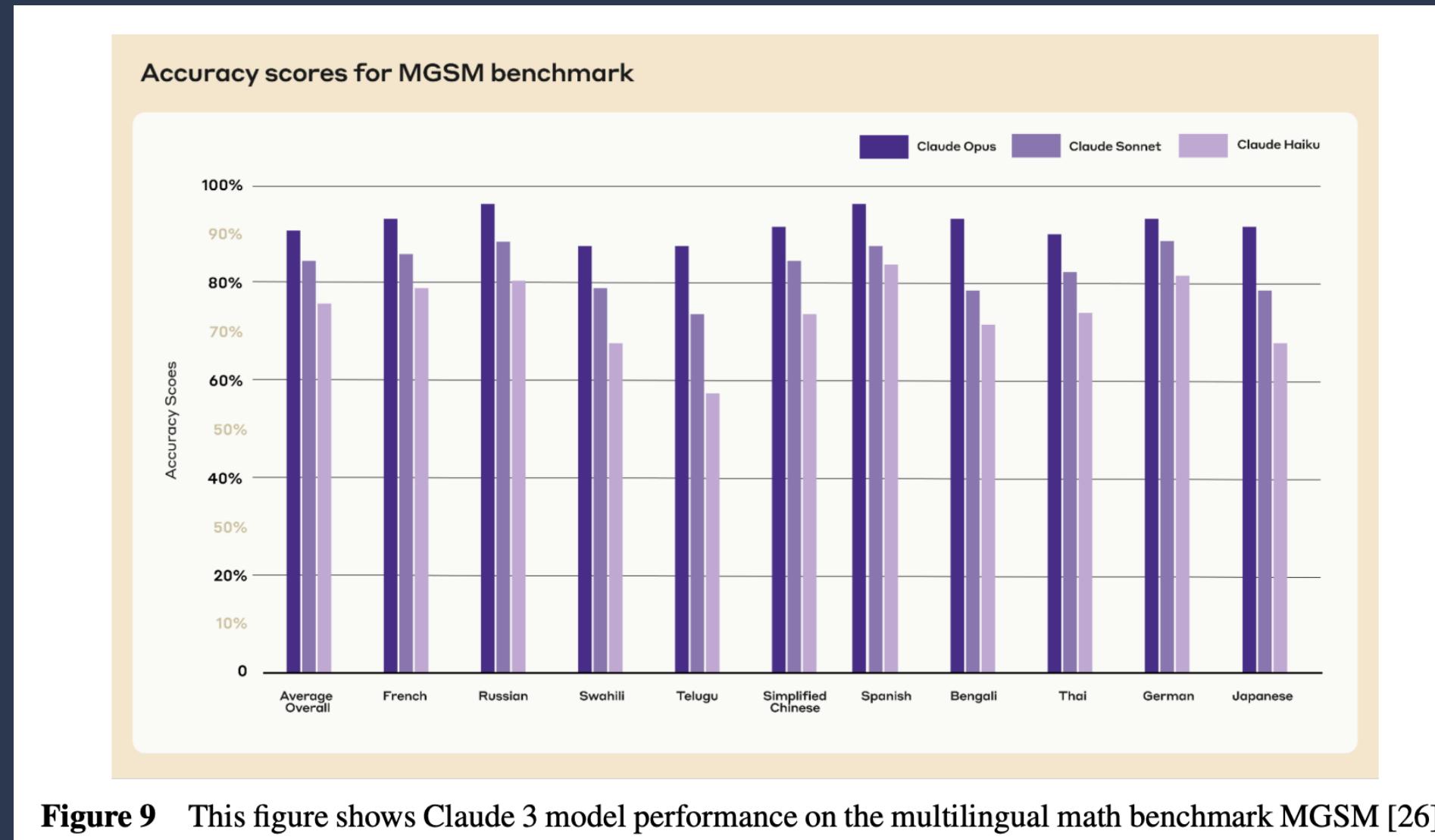


Figure 9 This figure shows Claude 3 model performance on the multilingual math benchmark MGSM [26].

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	Claude 2.1	Claude 2	Claude Instant 1.2	
Multilingual MMLU (Reasoning)	5-shot	79.1%	69.0%	65.2%	63.4%	63.1%	61.2%

Table 5 This table shows results on the multilingual MMLU benchmark. Claude 3 Opus outperforms its predecessor, Claude 2.1, by 15.7%.

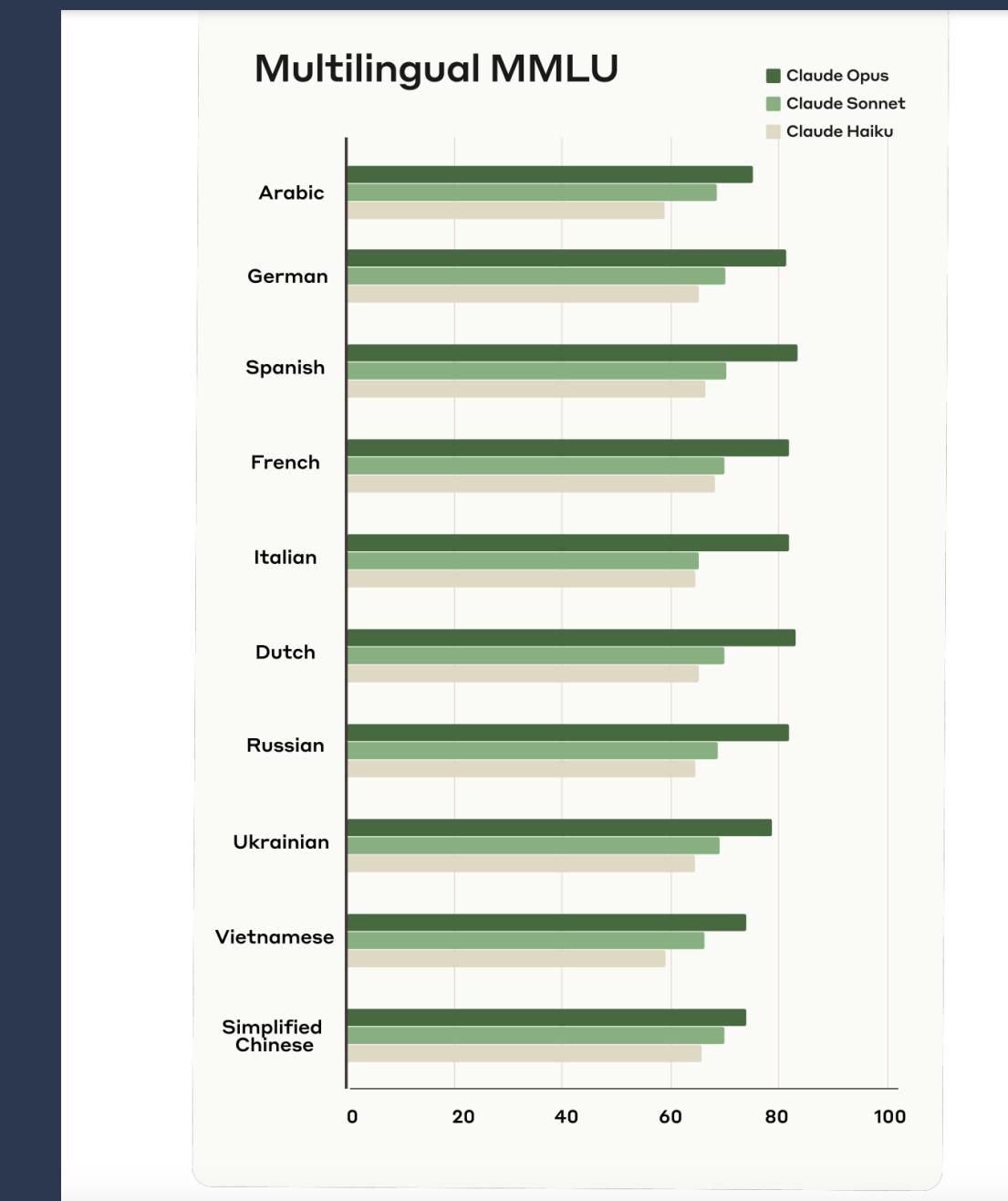
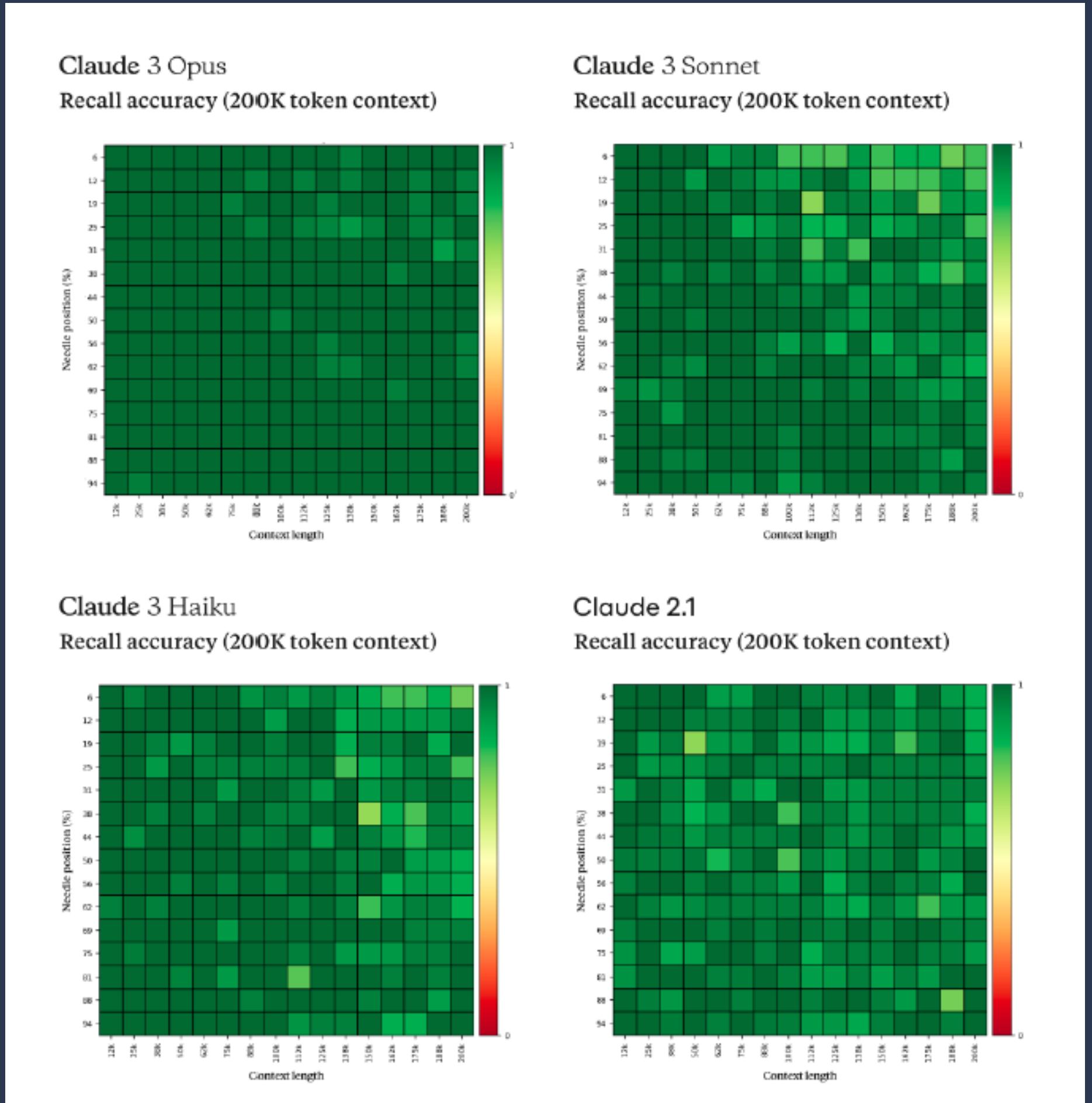


Figure 10 This figure shows results from the Multilingual MMLU evaluation on Claude 3 models.

Bedrock 在RAG上的优势

✓ Claude-V3 Benchmark – 200K 超长上下文



- Claude 3 系列模型，全部都支持 200K token 的上下文窗口。
- 此外针对某些特殊场景，会开放支持 1M token 的上下文窗口；
- 大海捞针（Needle In A Haystack, NIAH）召回率表现优异；
- 甚至还能识别出测试本身的局限，比如发现某“目标”句子明显是后来人为添加进原始文本的

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	Claude 2.1
All context lengths	99.4%	95.4%	95.9%	94.5%
200k context length	98.3%	91.4%	91.9%	92.7%

Bedrock 在RAG上的优势

✓ Claude V3 不同尺寸模型的定位

Opus

- 加速研究与开发任务,如文献综述、假设生成等
- 利用智能体和工具实现端到端的复杂任务自动化
- 专业级虚拟分析师,加速决策过程
- 先进的交互式编码,解决复杂问题

Sonnet

- 分析大量企业知识,获取数据洞察
- 为金融和市场模型进行预测
- 个性化客户分析和营销
- 提高日常知识工作的生产力

Haiku

- 支持多语种的实时交互式聊天机器人
- 大规模实时内容审核
- 从文本和图像中提取数据
- 快速、低成本的大规模翻译

Bedrock 在RAG上的优势

✓ Claude V3 性价比



Claude 3

The best models in the world for **speed**, **intelligence**, and **cost**

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4T	GPT-3.5T	Gemini Pro 1.0
Intelligence (Multi-benchmark approximation)	💡💡💡	💡💡	💡💡	💡💡💡	💡	💡
Speed (Approximation; after optimizations)	🏎️🏎️	🏎️🏎️🏎️	🏎️🏎️	🏎️🏎️🏎️	🏎️🏎️🏎️	🏎️🏎️
Multimodal	✓	✓	✓	✓	✗	✓
Cost (\$ per MTok)	Input: Output: \$15 \$75	\$3 \$15	\$0.25 \$1.25	\$10 \$30	\$0.5 - \$1.5 \$1.5 - \$2	\$0.47 ¹ \$1.42 ¹

With Claude-V3 Sonnet/Opus

应对复杂的上下文输入和指令

举例：RAG问答

With Claude-V3 Haiku

应对简单的/容易定义的任务

举例：

1. Few-shot 意图分类

2. IUR 任务 - Query重写

Bedrock 在RAG上的优势

✓ Cohere Embedding (100+ Multilingual能力)

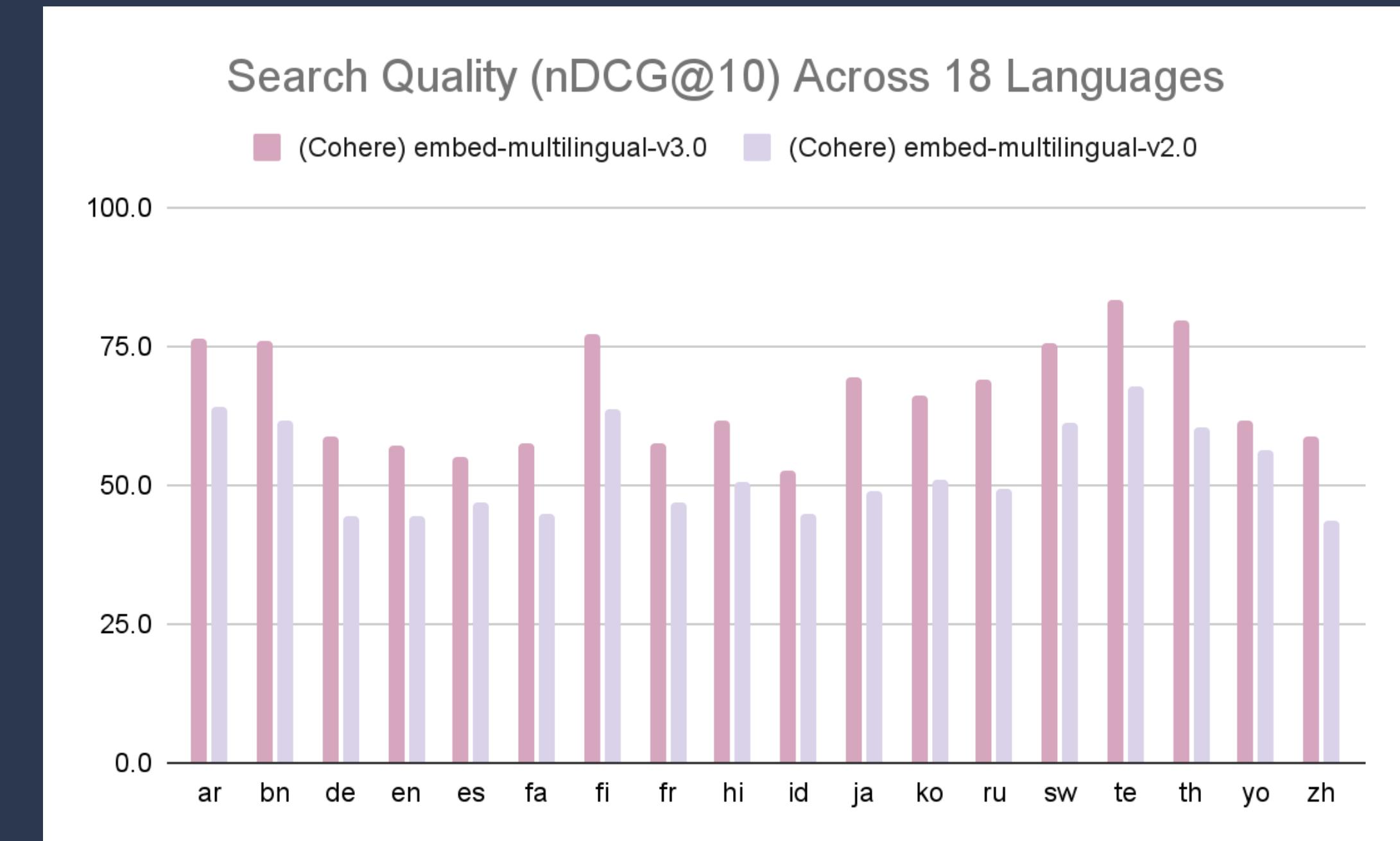
效果优秀

- 英文上超过bge-en-large-v.15
- METB Leaderboard 排名第二

Model	Dimensions	BEIR (nDCG@10, 14 datasets, higher=better)
embed-english-v3.0	1024	55.9
embed-english-light-v3.0	384	52.0
embed-multilingual-v3.0	1024	54.6
embed-multilingual-light-v3.0	384	50.9
Other Models		
BM25	-	43.0
OpenAI ada-002	1536	49.8
GTR-Base	768	44.1
GTR-XXL	768	48.6
DupMAE	768	49.7

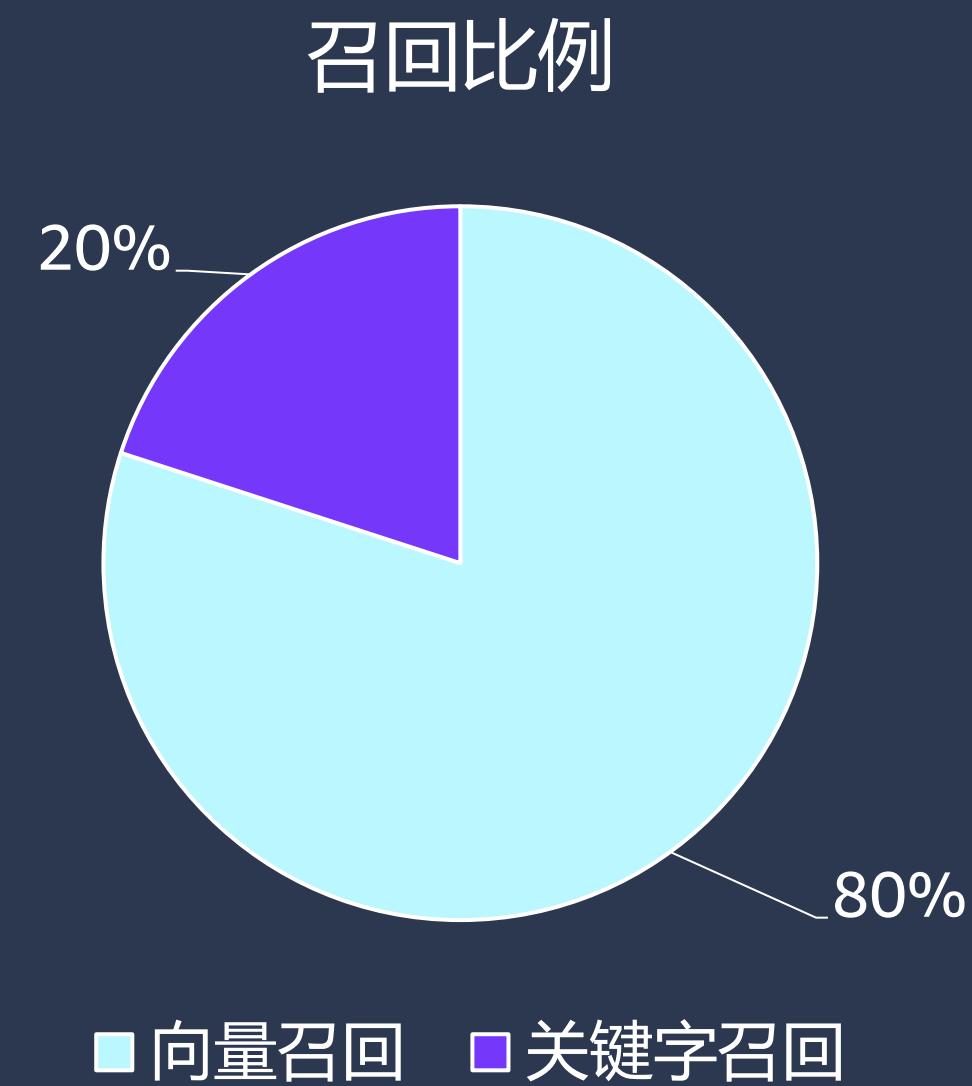
多语言能力均衡

- 具备跨语言对齐效果，可以通过中文召回类似语义的英文



OpenSearch 在RAG上的优势

- ✓ 多路召回 (向量 + 关键字)
- ✓ Hybird召回 (Reference Doc)
- ✓ Sparse Vector Retrieval



口 优缺点互补:

- 向量模型不理解垂直领域专词
- 向量模型有时出现语义相似但主题不相似情况
- 可解释性弱，不易通过补丁解决bad case
- 倒排召回，缺乏语义信息，仅靠关键词匹配

OpenSearch 在RAG上的优势

- ✓ 多路召回 (向量 + 关键字)
- ✓ Hybird召回 (Reference Doc)
- ✓ Sparse Vector Retrieval

好处：简单的内置多路召回，无需在客户端实现融合逻辑

- Step1: 创建Search pipeline

```
PUT /_search/pipeline/my-pipeline
{
  "description": "Post-processor for hybrid search",
  "phase_results_processors": [
    {
      "normalization-processor": {
        "normalization": {
          "technique": "l2" # min-max
        },
        "combination": {
          "technique": "arithmetic_mean" # geometric mean or
harmonic mean
        }
      }
    }
  ]
}
```

- Step2: 搜索时指定search pipeline

```
POST my_index/_search?search_pipeline=<pipeline>
{
  "query": {
    "hybrid": [
      {}, // First Query
      {}, // Second Query
      ... // Other Queries
    ]
  }
}
```

OpenSearch 在RAG上的优势

- ✓ 1. 多路召回 (向量 + 关键字)
- ✓ 2. Hybrid召回
- ✓ 3. Sparse Vector Retrieval(term expansion)

- 效果示意

Without Query Expansion

Query: Where did Beyonce perform in 2011?

In 2011, documents obtained by WikiLeaks revealed that Beyoncé was one of many entertainers who performed for the family of Libyan ruler Muammar Gaddafi. Rolling Stone reported that the music industry was urging them to return the money they earned for the concerts; a spokesperson for Beyoncé later confirmed to The Huffington Post that she donated the money to the Clinton Bush Haiti Fund. Later that year she became the first solo female artist to headline the main Pyramid stage at the 2011 Glastonbury Festival in over twenty years, and was named the highest-paid performer in the world per minute.

With Query Expansion

Query: Where did Beyonce perform in 2011?

[stage, appeared, festival] [bey, she, music, artist] [performed, performance]

In 2011, documents obtained by WikiLeaks revealed that Beyoncé was one of many entertainers who performed for the family of Libyan ruler Muammar Gaddafi. Rolling Stone reported that the music industry was urging them to return the money they earned for the concerts; a spokesperson for Beyoncé later confirmed to The Huffington Post that she donated the money to the Clinton Bush Haiti Fund. Later that year she became the first solo female artist to headline the main Pyramid stage at the 2011 Glastonbury Festival in over twenty years, and was named the highest-paid performer in the world per minute.

- Sparse Vector 在近似RAG场景的数据集上带来的提升

Dataset	BeIR/fiqa				squad_v2	
	Method\Metric	Recall@1	Recall@4	Recall@10	Recall@1	Recall@4
bm25	0.112	0.215	0.297	0.59	0.771	0.851
dense	0.156	0.316	0.398	0.671	0.872	0.925
sparse	0.196	0.334	0.438	0.684	0.865	0.926
hybird_dense_sparse	0.203	0.362	0.456	0.704	0.885	0.942
hybird_dense_bm25	0.156	0.316	0.394	0.671	0.871	0.925

THANKS

软件正在重新定义世界
Software Is Redefining The World