

AI 驱动 软件研发 全面进入数字化时代

中国·北京 08.18-19

AI+
software
Development
Digital
summit



大模型落地面向企业级 客户服务场景的实践

Sunny Zhang 浪潮信息

科技生态圈峰会 + 深度研习 —— 1000+ 技术团队的选择



2023K+
全球软件研发行业创新峰会
上海站

会议时间 | 06.09-10



2023K+
全球软件研发行业创新峰会
北京站

会议时间 | 07.21-22



2024K+
全球软件研发行业创新峰会
深圳站

会议时间 | 05.17-18



K+峰会详情



会议时间 | 08.18-19
AiDD AI+软件研发数字峰会
北京站



会议时间 | 11.17-18
AiDD AI+软件研发数字峰会
深圳站



AiDD峰会详情

▶ 演讲嘉宾



Sunny Zhang

浪潮信息 大模型高级产品经理

博士研究生，浪潮信息人工智能与高性能应用软件部高级产品经理，源大模型产品负责人。长期从事高性能计算和人工智能计算的创新与产品化工作；参与国家级自然科学基金3项，省部级科技进步奖6项；发表国内外SCI、EI期刊及会议论文10余篇。

目录

CONTENTS

1. 大模型已落地ToC客户服务
2. ToB客户服务的特点和痛点
3. 大模型落地ToB客户服务的解决思路
4. 源1.0落地ToB客户服务的技术实践

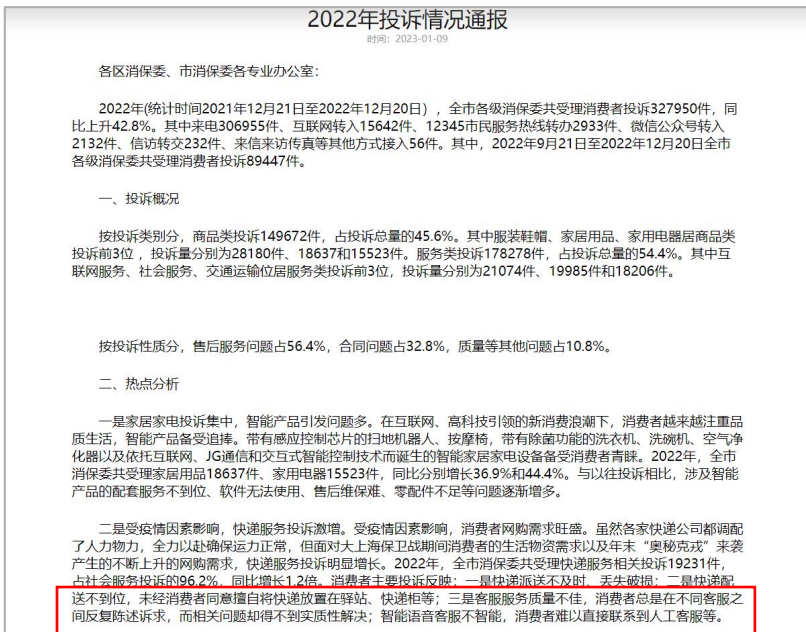
PART 01

大模型落地ToC客户服务



智能客服是企业最希望AIGC落地的场景

- **人工坐席数量庞大，工作强度大，企业成本高。**根据CC-CMM行业标准的统计,2022年呼叫中心总座席数约为135万。人工坐席每天接听电话几十个到几百个，一通电话成本可能高达几十元。
- **企业急需降本增效，智能客服市场规模大。**企业沙利文《2023年中国智能客服市场报告》指出，2022中国智能客服市场规模达到66.8亿元，预计到2027年市场规模有望增长至181.3亿元。
- **目前智能客服机器人水平有限，客户体验差，转人工量不减反增。**根据CC-CMM对近200家企业的调查显示，实施智能语音之后，有超过三分之一的企业热线，转人工量不减反增，对人工座席反而造成更大的话务压力。



上海市消保委《2022年投诉情况通报》

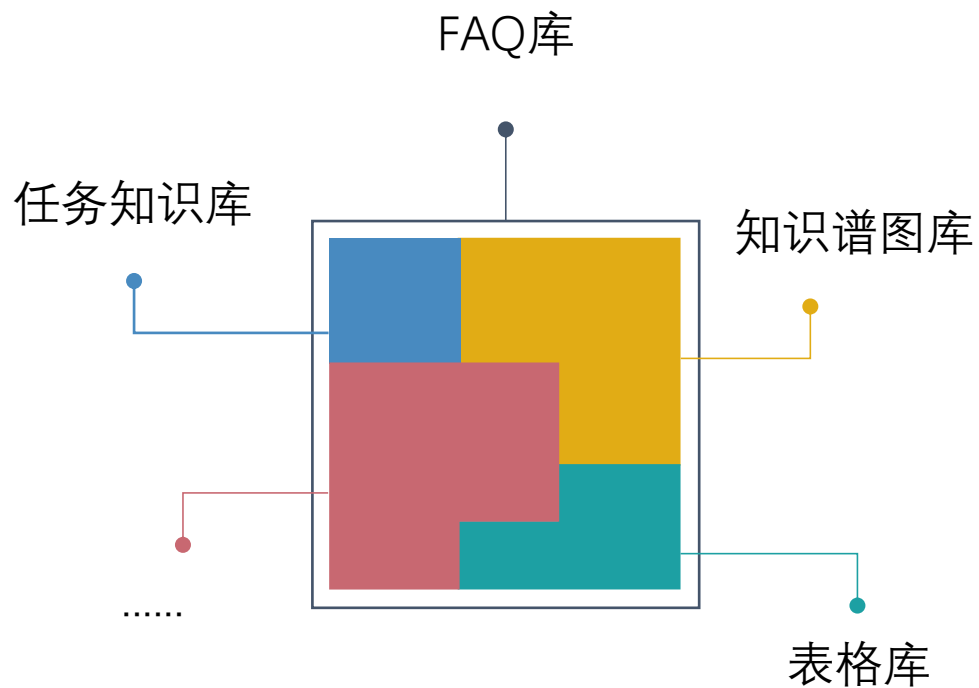


人民网-人民日报（2023年05月23日）

▶ 传统智能客服被诟病的地方

成本高

“有多少人工，就有多少智能”



起点是人工建知识库

天花板低

“那么多人工，还是没多少智能”



我在直播间买了东西已经收到货了，但是没有收到中奖的奖品。



您好，直播间中奖奖品满足派发条件后，正常是在7个工作日左右发出，还请您耐心等待。



已经超过7天了啊……



您好，商品标签等完整的情况下，我们支持7天无理由退货。



我不是退货啊……是没有收到直播间的奖品。



您好，直播间中奖奖品满足派发条件后，正常是在7个工作日左右发出，还请您耐心等待。



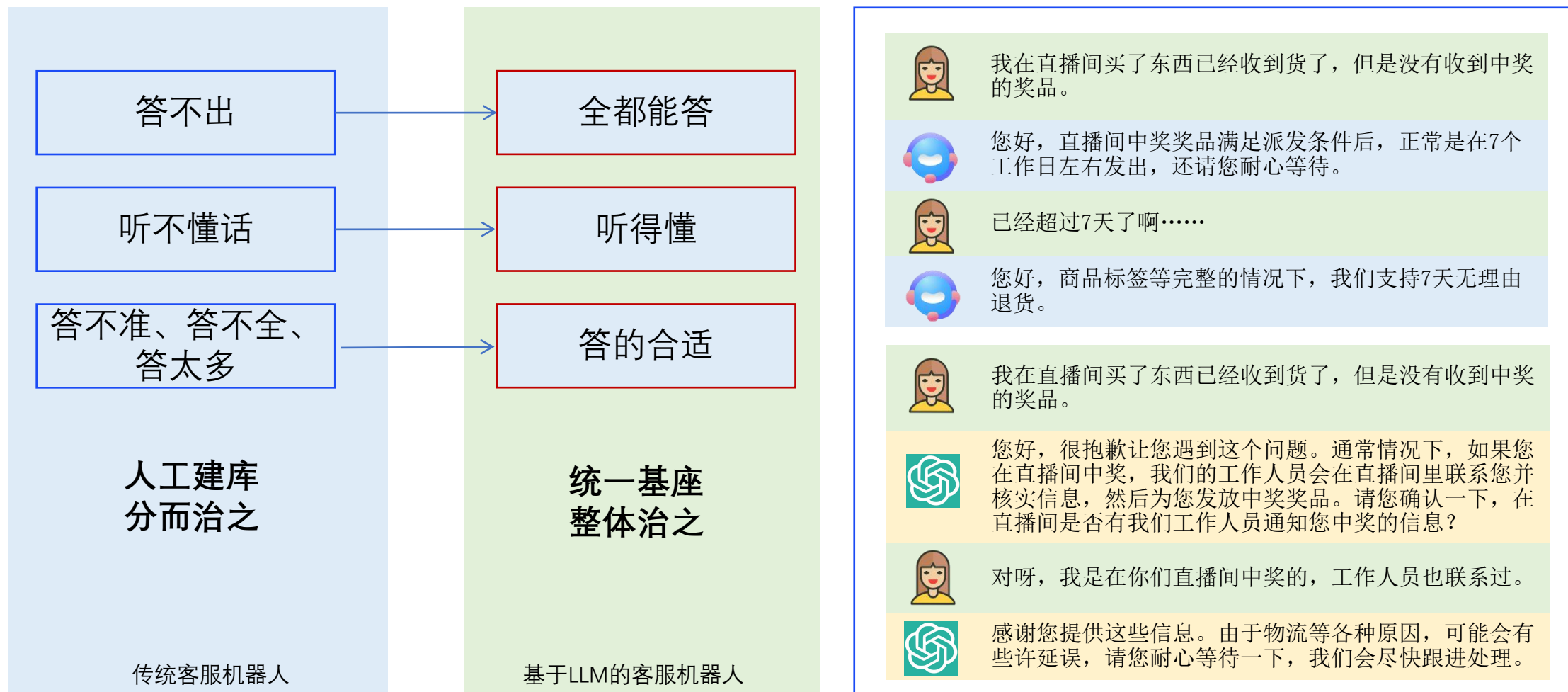
……怎么回事……



对不起，我不太明白您的意思～您可以尝试换一种方式提问～

本质上是搜索匹配

▶ 以ChatGPT为代表的LLM成为智能客服 “智能” 的新希望



▶ 大模型已有ToC客服服务的落地案例

业界已有不少企业尝试基于ChatGPT来搭建智能客服机器人

- **微调行业大模型**：采用行业数据、客户服务数据进行再训练和微调，得到垂直领域大模型
- **角色设定（Prompt工程）**：用文字来进行角色的设定，将ChatGPT设定为客服角色
- **嵌入（Embedding）**：将文本转换为一个向量，去企业向量数据库进行检索，检索出相关内容
- **插件**：识别出实体和关键词，采用外部插件，获取到相关信息

部分或全部使用了这几种方法，而且对应每种方法目前也有一些开源的框架工具。

微调	Prompt工程	Embedding	插件
如，银行的智能客服可以基于基础模型引入行业数据、企业数据、客户服务数据进行模型再训练和微调	如：你是某电商XX公司的客服，你们公司的产品为XXX，你的职责是为顾客提供24小时在线答疑服务。你善于倾听客户需求、细致耐心地解答问题，总是以最佳的用户体验为出发点去工作。	如： 问：给我推荐一下最近比较流行的口红？ 可以先根据“问题”检索对应的产品库，然后根据这些背景信息，让模型给出合适的回答。	如搜索引擎、天气插件等 问：根据最近全国的天气情况，给我规划一份旅游出行计划，出游时间是7天。
学习领域新知识	激发能力	扩展领域信息	扩展能力边界

▶ 大模型已有ToC客服服务的落地案例

LLM引入到现有客服系统，可以在原有客服系统基础上，优化客服系统功能以及提高客服机器人回复质量。

国内外客服系统对比分析		
相关公司	客服系统功能	智能客服机器人功能
LiveChat	多个网站消息，一个账户解决	通过客户交流判断客户资质，挖掘潜在客户
	便捷接入网站，拖动创建小组件	使用多样化的产品展示方式吸引客户注意力
	AI为真人客服提供建议	帮助客户处理简单的问题，并将重要业务转接至人工
HubSpot	对话收件箱收集各平台消息	撰写邮件用于营销或挖掘潜在客户（OpenAI）
	设置知识库，便于一键回复	与HubSpot CRM集成后，可以识别客户并进行个性化对话
	自动生成任务并排列优先级	提供基础的客户支持，包括客户资格判断、会议预定等
容联云	多渠道接入	拖拽式组件，实现可视化定义机器人
	回复内容支持全媒体形式	精准回答问题、特定任务执行、闲聊陪伴
	跟踪客户访问路径，精准营销	BI智能分析图表多维度分析智能机器人工作效果与客户反馈
多客 (未上市)	平台API对接，确保店铺安全	售前主动催单促进转化，售后要求评价促进自我完善
	支持移动端回复	机器人设置有行业包，可以根据所处行业进行选择
	接入ChatGPT助力快速回复	快速响应与简单客户服务

资料来源：LiveChat，ChatBot，HubSpot，OXYGEN，容联·七陌，多客，国海证券研究所

▶ 大模型已有ToC客服服务的效果咋样

问答不错，对话不行；通用知识不错，专有知识不行

说明	示例
单轮问答不错	对于用户问题理解力强，并且能够给出较丰富的答案
多轮对话不行	受限于模型输入的长度，以及对于对话策略的掌握
通用知识不错	LLM+Web检索，具备丰富的开放域知识
专有知识不行	依赖于向量知识库的丰富度，检索的准确度，以及模型对于专有知识的理解力
总结摘要不错	根据内容进行总结不错
知识推理不行	根据文本内容进行推理的能力有限

PART 02

ToB客户服务的特点和痛点

▶ ToB与ToC的客服服务有什么区别



C端用户



客服



B端用户



工程师

用户 都3天了，我怎么还没收到货

用户 有台服务器，看不到硬盘

用户 我这个银行卡，为啥转了5000元就不让转了？

用户 你好，这台服务器NF5468M6，安装redhat7.9 找不到硬盘，请问这个型号raid卡驱动在哪里有下载

专业度要求高

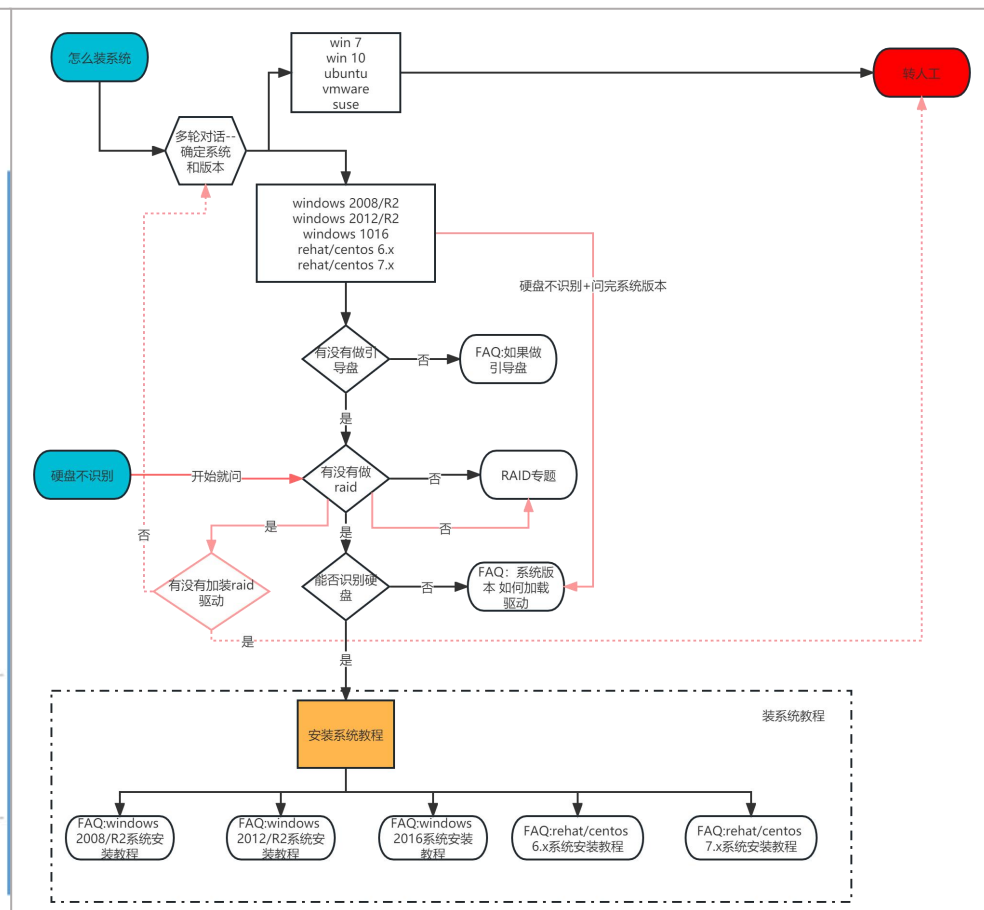
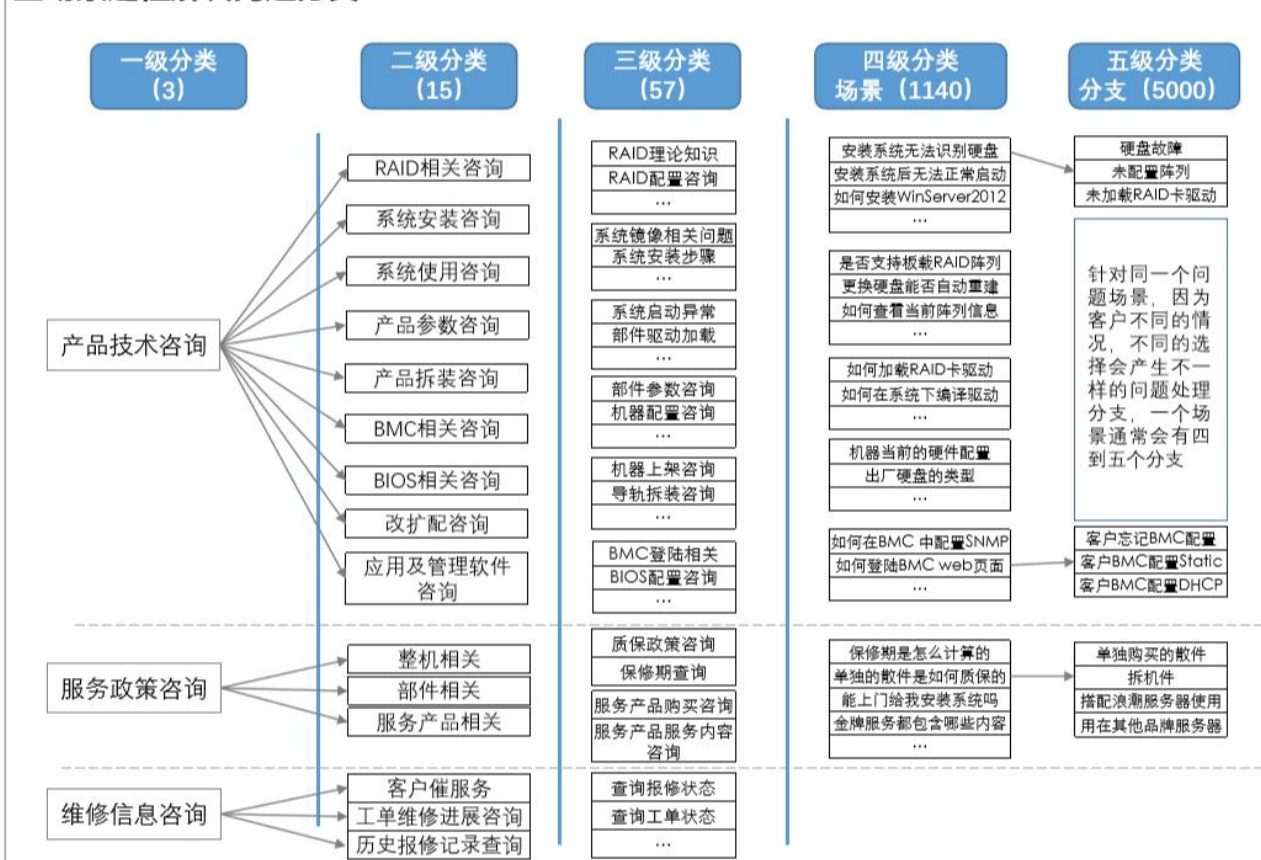
场景更复杂

对话轮次更多

▶ ToB与ToC的客服服务有什么区别

■ 专业度要求高、场景更复杂

全场景远程解决问题分类





用户问题专业度高

用户问题难度较高
交互体验表现不佳

- 售后侧用户咨询和报修非常专业、解答难度较高；
- 智能客服对于用户提问存在理解偏差，导致答非所问；
- 智能客服呈现给用户的答案过于冗余，还需用户定位。

2022-08-29 14:20:50用户



Inspur3008imr、LSI9361-8i/9364-8i Raid卡配置热备

机器人2022-08-29 14:20:50

您好：

- 1.当发现如下界面时，按Ctrl+R进入raid配置界面
- 2.如果要配置某一阵列的热备盘，应该已经配置过了raid阵列，选择对应的raid阵列后，按“F2”键，选择“Manage Ded. HS”。
- 3.此时，跳出对话框，选择磁盘成为这个阵列的热备盘，之后点击“OK”即可。
- 4.此时检查“PD Mgmt”界面，则会发现有磁盘状态为“Hotspare”
- 5.如果要配置全局热备盘，则在“PD Mgmt”界面，选中某一磁盘，按

2022-08-29 14:19:11用户



如何做热备？

机器人2022-08-29 14:19:11

您好,请提供机器序列号转人工，不同的raid卡配置热备的方法也不一样。

ctrl A

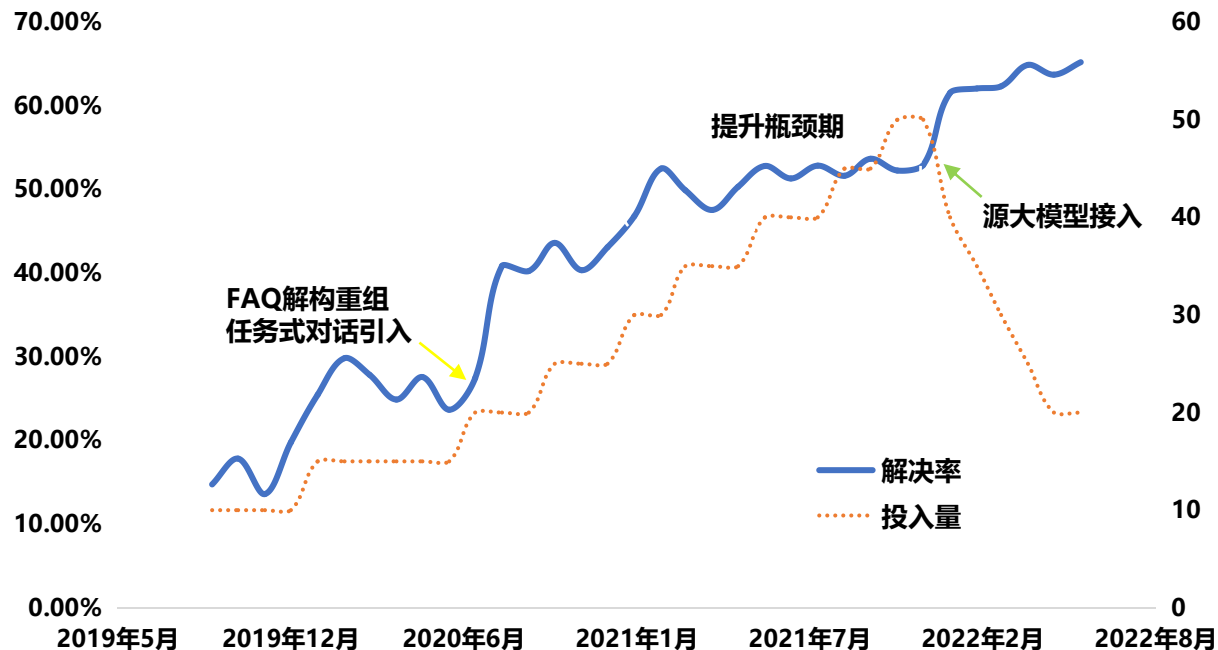
http://www.4008600011.com/archives/9346#110_Hotspare

ctrl R

<http://www.4008600011.com/archives/429#i>

解决率易进入瓶颈

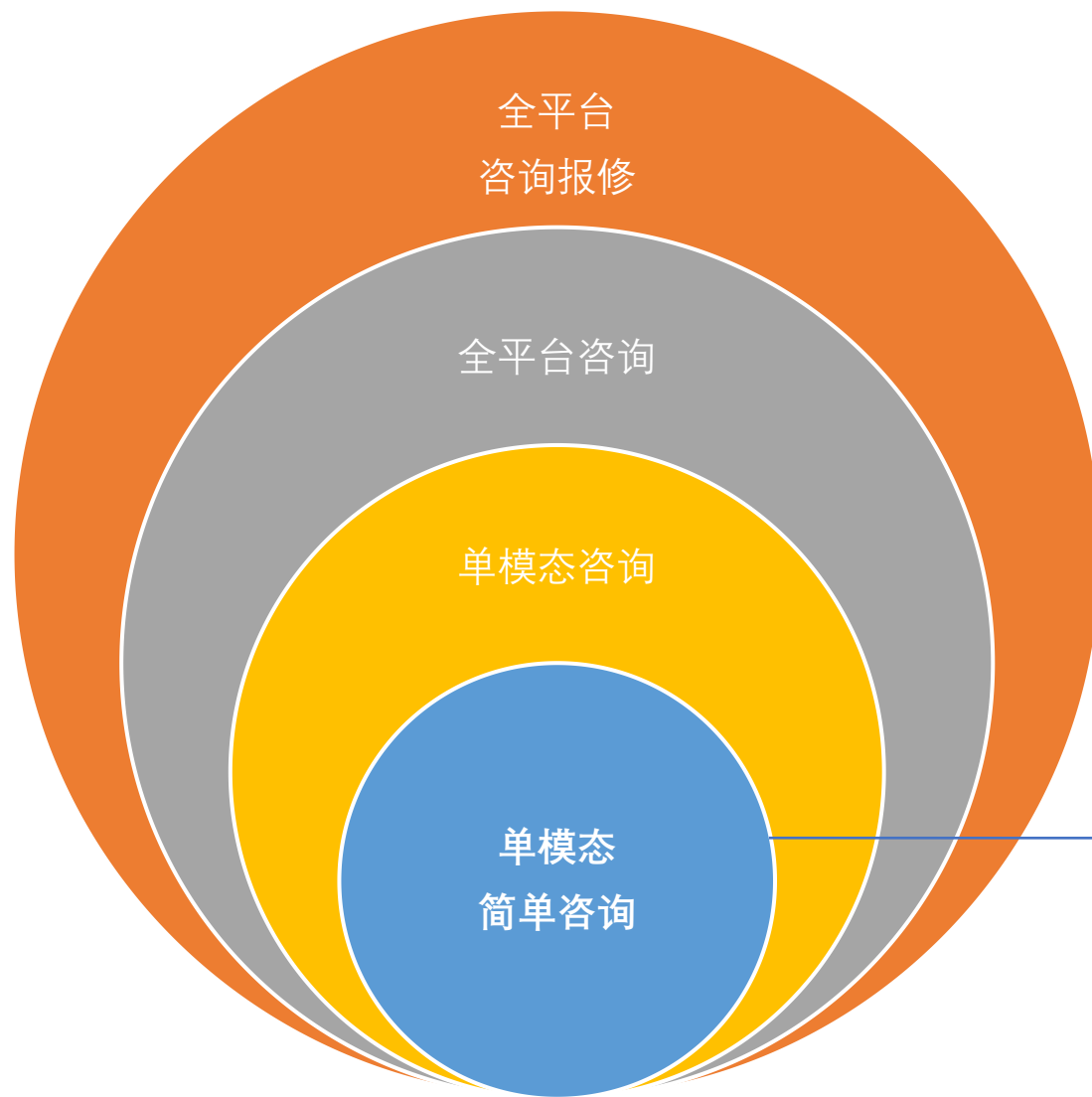
2019.9-2022.6 智能客服机器人解决率



人工训练高价低效
效果提升遭遇瓶颈

- 行业领域越专业越细分，人工训练成本就越高；
- 人工训练需要维护FAQ、任务式等多个知识库多个小模型，效率低下，标准也难以拉齐；
- 浪潮售后侧用户问题解决率的提升进入瓶颈期，亟待模式革新。

▶ 对话刻板不自然，仅能解决简单咨询问题



对话刻板不拟人

PART 03

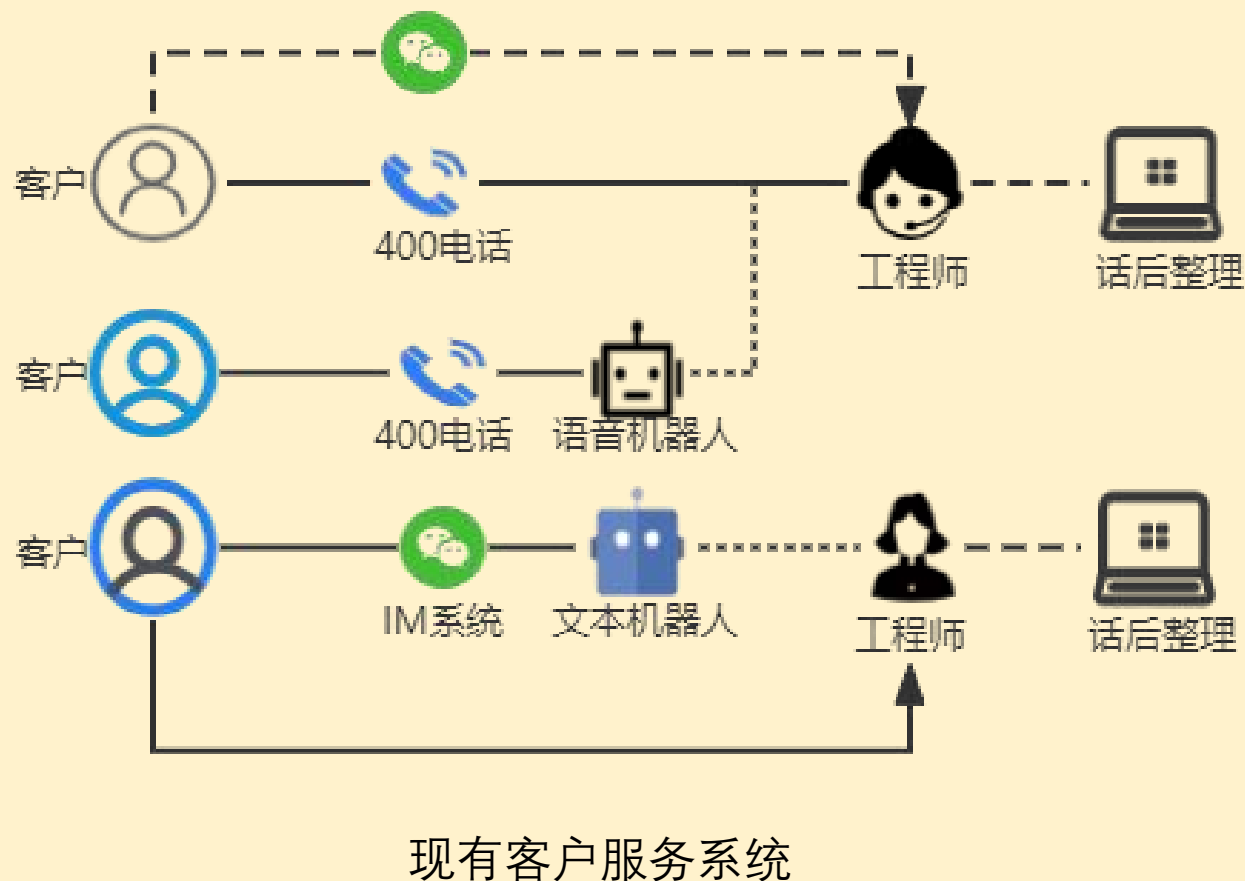
大模型落地ToB客户服务的解决思路

▶ 落地场景——IT企业的ToB客户服务

落地场景：浪潮信息客户服务部门的智能客户服务系统，浪潮信息是全球领先的IT基础设施产品、方案和服务供应商，致力为用户提供各类智慧计算解决方案，**客户服务部门承担着公司支持浪潮信息服务器、存储、边缘计算等八大产品线的产品售后服务**，包括技术、产品、故障、服务四大类1000多个细分场景问题。

客户服务系统现状：目前用户接入渠道有400电话、IM对话系统（微信公众号），有人工坐席、文本机器人（IM对话机器人）、语音机器人、数字人。IM对话机器人经过两年多的人工训练，其问题解决率的提升速度逐步放缓趋平，人工成本投入却持续攀升，边际效用递减显著；与此同时，IM对话（实时通信）在整个咨询报修中的占比提升也遭遇了瓶颈，热线工程师每天仍然要面对海量的电话咨询。

补充——赋能——替代



补充: 离线使用, 补充现有客户服务系统知识库、工单的内容, 减少人工工作量

赋能: 在线使用, 增加现有客服系统中的功能, 提高整个客服系统的智能化水平

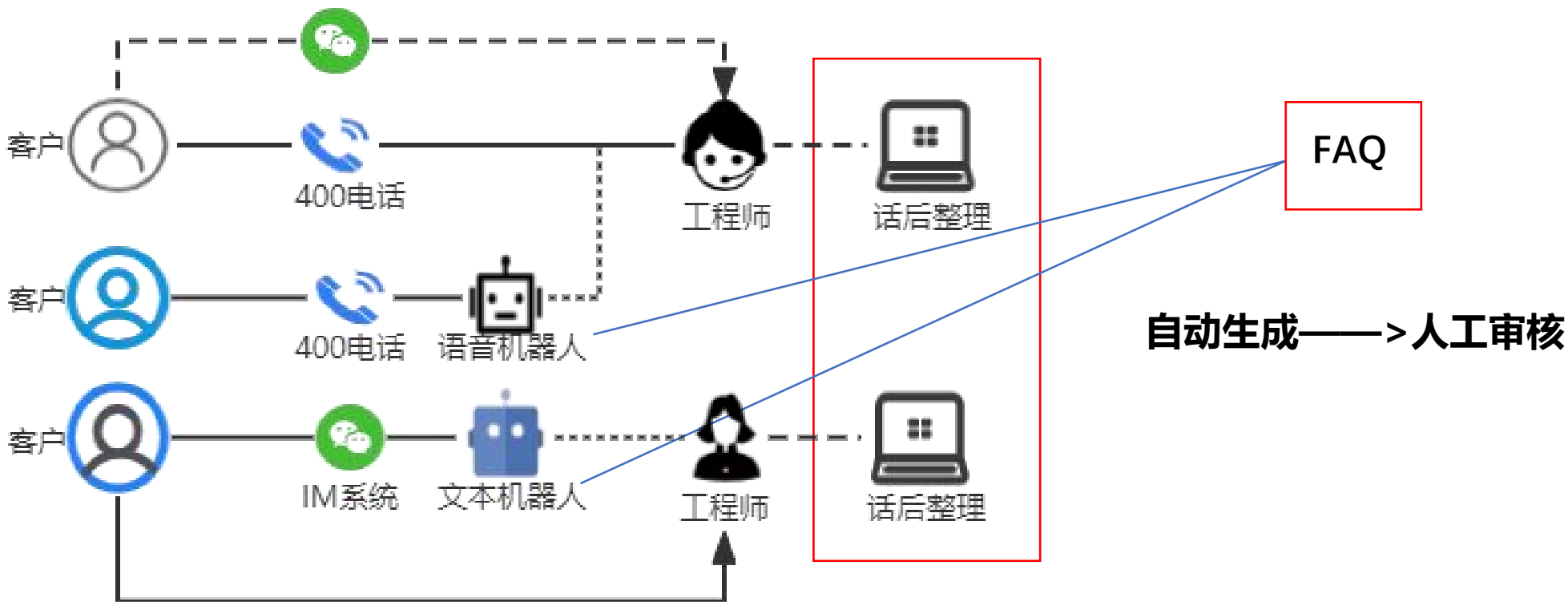
替代: 在线使用, 替代语音机器人、文本机器人、替代人工坐席

► 补充——时效性、准确性要求低的场景

任务1：FAQ生成：基于文档资料库，自动生成（问题，答案）对，扩充FAQ

任务2：相似问生成：基于基础问题，自动生成语义相似表达不同的相似问题

任务3：话后整理自动填写：基于客服服务对话流，自动生成话后整理的工单字段，包括问题描述、咨询过程描述、问题分类标签和解决方案分类标签。

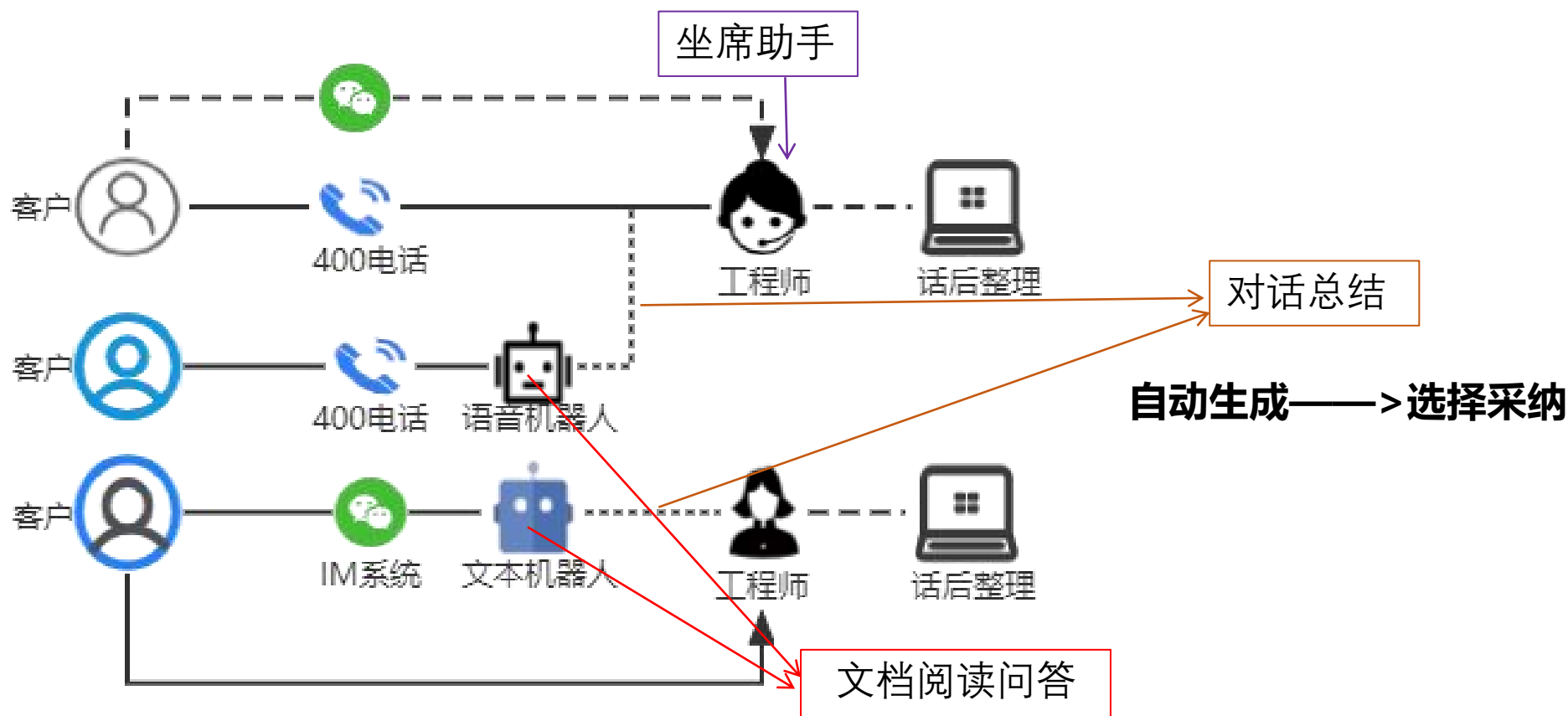


▶ 赋能——准确性要求低的场景

任务1：对话总结：转人工时，基于对话流，自动生成摘要总结，供人工坐席参考

任务2：坐席助手：基于对话流，自动生成知识推荐内容，供人工坐席参考

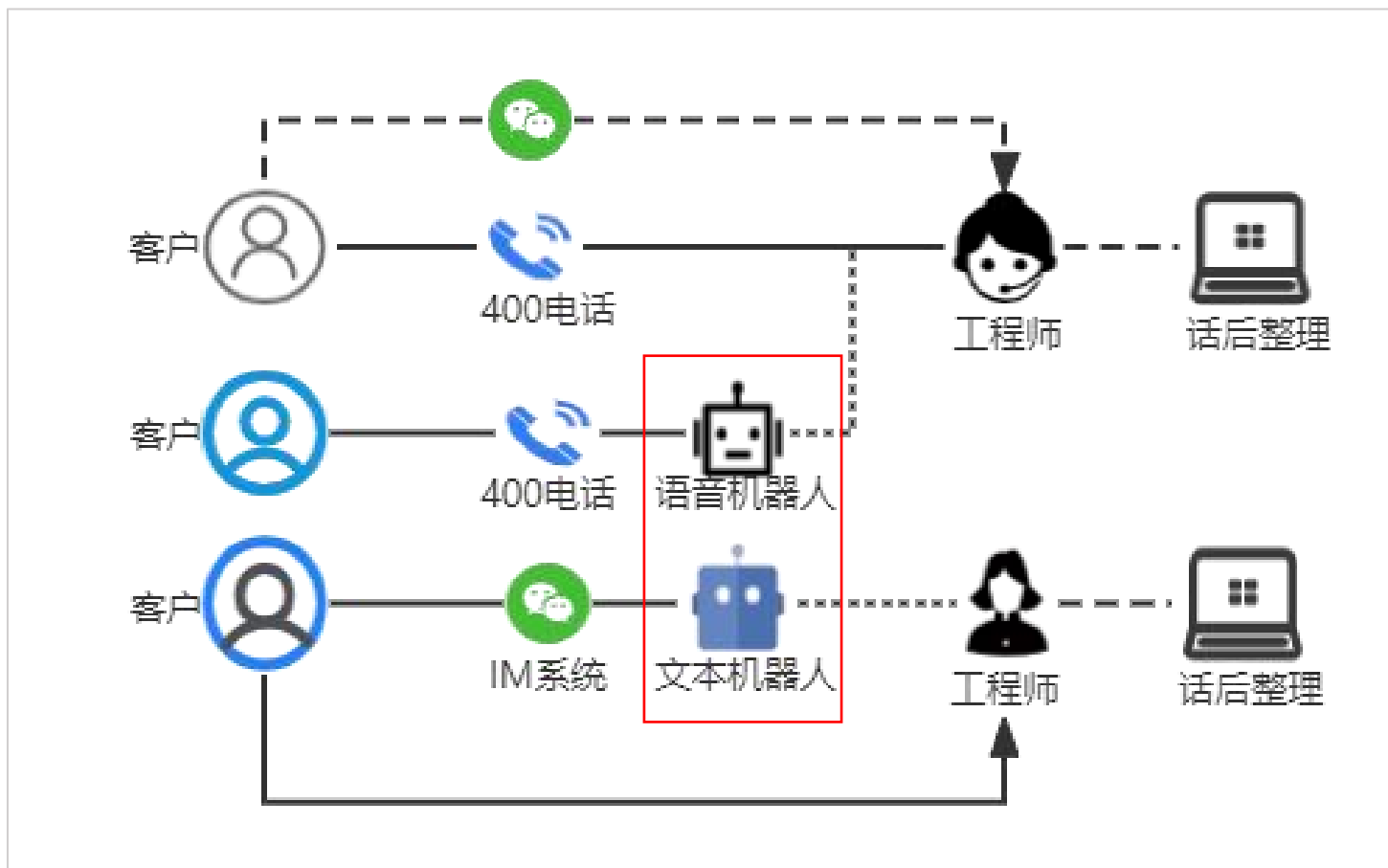
任务3：文档阅读问答：针对用户问题，基于文档阅读理解的问答。作为机器人的兜底回复方案。



► 替代——时效性、准确性要求高的场景

任务1：语音机器人、文本机器人：基于大模型实现单模态的客服对话机器人

任务2：多模态机器人：基于大模型实现多模态的客服对话机器人



单模态机器人：用户输入的内容不包含图片，在进行语义检索、机器人生成时也不包含图片。但可以将原有文档中的图片作为辅助内容输出。传统的语音机器人、文本机器人都是单模态机器人。

多模态机器人：用户输入的内容包含图片、或含有图片的文档，机器人可以理解图片，基于图片进行语义检索，生成的内容主要是文本，附带的图片来自于知识库，而不是由模型进行生成。

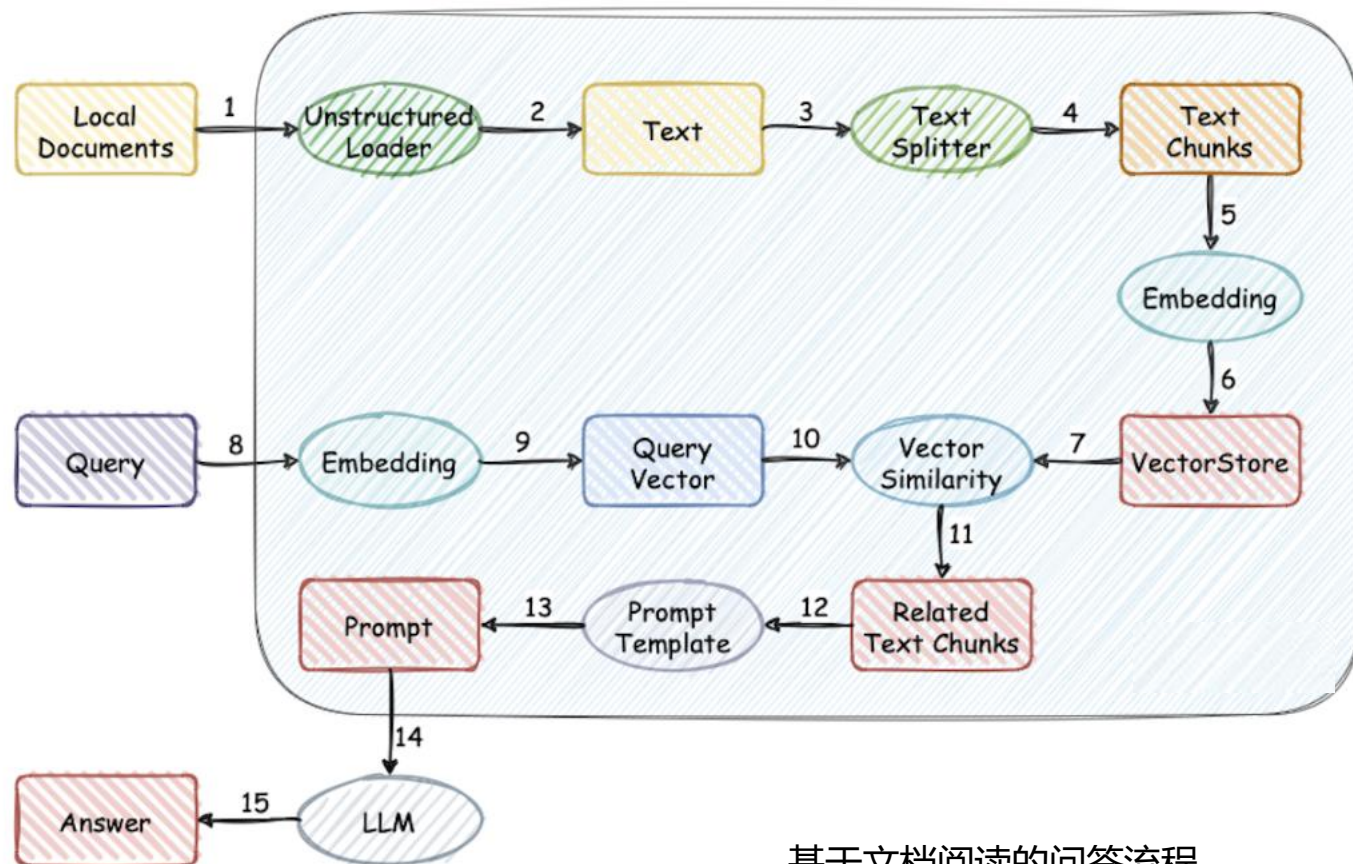
是否能理解图片和视频

PART 04

源1.0落地ToB客户服务的技术实践

基于文档阅读的问答技术实践

■ 核心是语义编码和检索，依赖于人工编写prompt模版，精度会累积影响

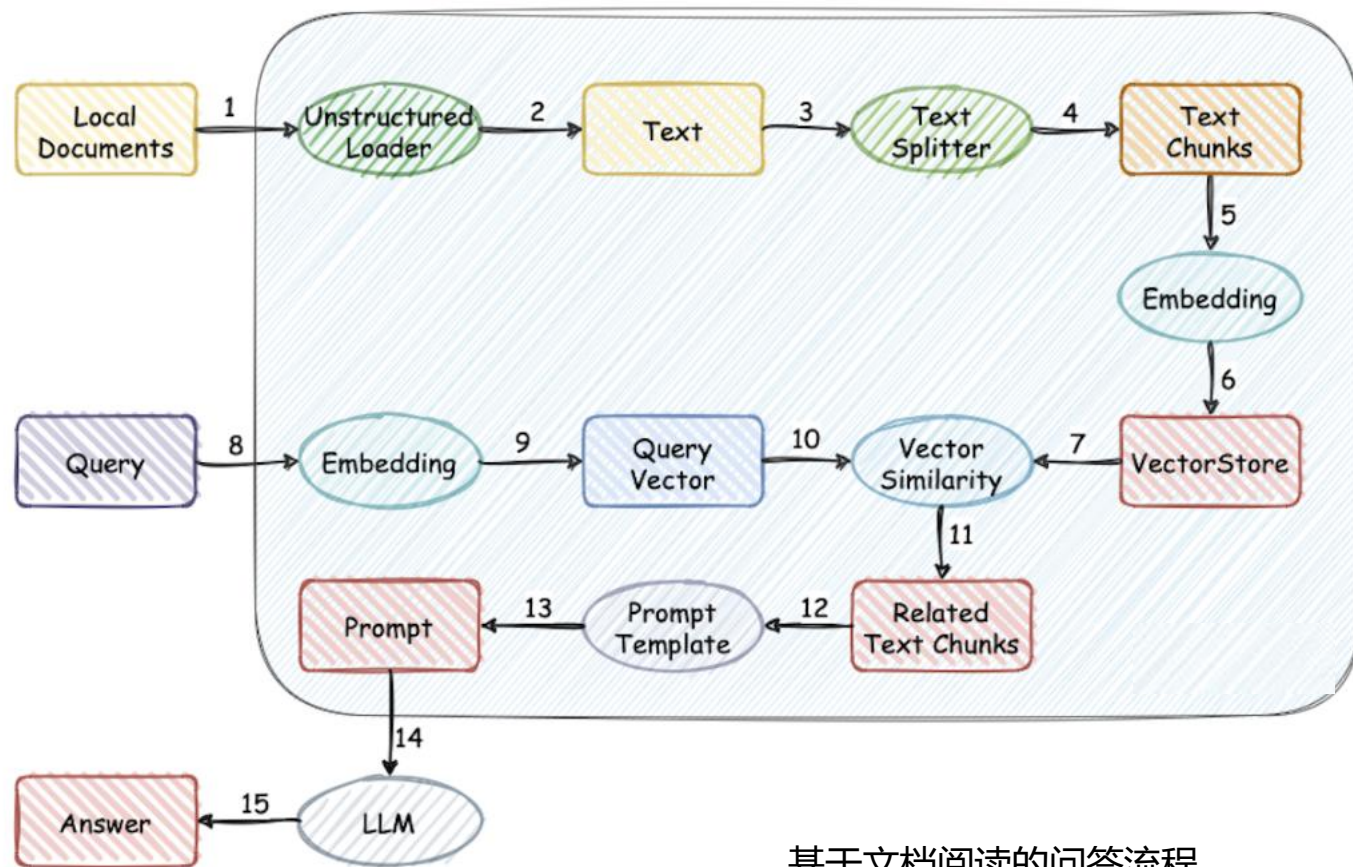


基于文档阅读的问答流程

1. **文本分块**，按规则进行分割（段落、句子结尾、字数控制）
2. **文本转向量**，采用text2vec基础模型text2vec-large-chinese进行微调过的模型
3. **向量数据库**：Faiss
4. **query**：并不是所有用户消息都是query，只有关键词匹配到的才被认为是query
5. **Prompt模版**，根据任务场景，人工进行的撰写
6. **评测**：随机抽取，人工标注是否正确

基于文档阅读的问答技术实践

■ Pipeline的流程已走通，但每个步骤都还存在技术上的挑战，整体精度有待提高



基于文档阅读的问答流程

存在的问题：

- 1. 文档内容识别**，文档的格式多样，含有众多的图表，仅采用OCR、图片理解模型不能把所有信息识别到，目前仅保留了文本，抛弃了其他内容。
- 2. 文本分块**，目标是完整的知识点，但仅仅采用规则来进行分块，存在知识点被分割、不完整的情况。
- 3. 大模型**，目前存在prompt敏感，生成领域结果不准确的现象，同时推理时间长，无法支持在线的多路并发。

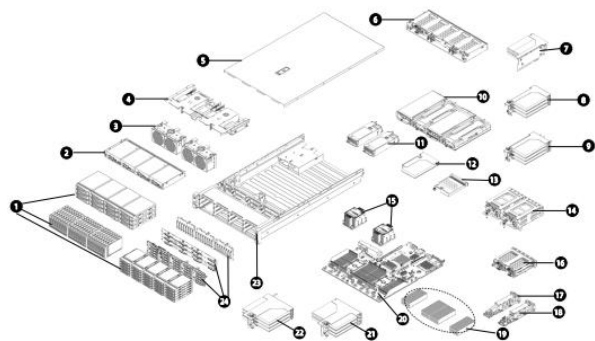
最终结果，离线平均65%，在线平均45%，可作为组件补充到现有的机器人中。

基于文档阅读的问答技术实践

分步骤说明-文档内容识别

3.6 物理结构

图 3-7 物理结构图



编号	模块名称	编号	模块名称
1	前置硬盘模组	13	电容板
2	中置硬盘模组	14	后置3.5英寸硬盘模组
3	风扇模组	15	散热器模组
4	导风罩	16	后置2.5英寸硬盘模组
5	机箱上盖	17	后置M.2模组
6	后置2.5英寸硬盘模组	18	后置E1.S模组
7	右侧转接卡模组（8卡配置）	19	内存
8	中间转接卡模组（8卡配置）	20	主板
9	左侧转接卡模组（8卡配置）	21	右侧转接卡模组（11卡配置）
10	GPU模组（4GPU配置）	22	左侧转接卡模组（11卡配置）
11	电源模块	23	机箱
12	转接卡模组（4GPU配置）	24	前置硬盘背板

内容识别

仅识别文本

3.6 物理结构

图 3-7 物理结构图

编号 模块名称 编号 模块名称

1 前置硬盘模组 13 电容板
2 中置硬盘模组 14 后置3.5英寸硬盘模组
3 风扇模组 15 散热器模组
4 导风罩 16 后置2.5英寸硬盘模组
5 机箱上盖 17 后置M.2模组
6 后置2.5英寸硬盘模组 18 后置E1.S模组
7 右侧转接卡模组（8卡配置） 19 内存
8 中间转接卡模组（8卡配置） 20 主板
9 左侧转接卡模组（8卡配置） 21 右侧转接卡模组（11卡配置）
10 GPU模组（4GPU配置） 22 左侧转接卡模组（11卡配置）
11 电源模块 23 机箱
12 转接卡模组（4GPU配置） 24 前置硬盘背板

数据处理

添加标题、段落标签
文本处理（删空格、换行等）

标题1：浪潮英信服务器
NF5280M6 用户手册 V1.4

标题2：物理结构

段落：

1 前置硬盘模组
13 电容板
2 中置硬盘模组
14 后置3.5英寸硬盘模组
3 风扇模组
15 散热器模组
4 导风罩
16 后置2.5英寸硬盘模组
5 机箱上盖 17 后置M.2模组
6 后置2.5英寸硬盘模组
18 后置E1.S模组
7 右侧转接卡模组（8卡配置）
19 内存
8 中间转接卡模组（8卡配置）
20 主板
9 左侧转接卡模组（8卡配置）
21 右侧转接卡模组（11卡配置）
10 GPU模组（4GPU配置）
22 左侧转接卡模组（11卡配置）
11 电源模块 23 机箱
12 转接卡模组（4GPU配置）
24 前置硬盘背板

基于文档阅读的问答技术实践

■ 分步骤说明--文本分块

因源1.0支持的输入tokens为2048，加之ToB客户服务的知识点教长，**文本分块的max tokens设置为512tokens**，可以满足top3的语义相关段落作为背景信息。除了max tokens进行设置，还启用规则匹配（段落结尾）。分词采用结巴分词工具，词表采用源1.0大模型的词表+领域扩展词。

具体标题1：浪潮英信服务器 NF5280M6 用户手册 V1.4
标题2：常见硬件问题
段落：
开机不加电
故障描述：
按电源开关按键后，电源开关按键上的电源指示灯为橙色且不变绿，硬盘活动状态指示灯不亮，同时显示器无显示输出，服务器机箱风扇不转动。
处理建议：
查看服务器后窗上电源模块指示灯的状态。
电源模块指示灯不亮或为琥珀色
a. 如果电源模块指示灯不亮或为琥珀色，则表示供电异常。请检查电源插座是否能正常工作以及电源线是否安插到位，相应电源模块上的供电线供电是否正常工作；
b. 确保电源插座可正常工作，电源线安插到位且供电线可正常工作后：电源模块指示灯仍就不亮或为琥珀色，则可能为电源模块故障，请使用同型号机器同配置的电源模块更换该电源模块，排查是否为电源模块故障；电源模块指示灯变为绿色，但电源开关按键仍就为橙色，请拨打浪潮客户服务热线(400-860-0011)或邮件联系我们(lckf@inspur.com)。
电源模块指示灯均为绿色
a. 如果电源模块指示灯均为绿色，请拔掉供电线，重新拔插电源模块，安插好供电线后再次开机验证；
b. 如果拔插电源模块测试仍然无效，则使用同型号机器同配置的电源模块更换该机器的电源模块，测试是否为电源模块故障；
c. 如果以上操作无法解决故障问题，请拨打浪潮客户服务热线(400-860-0011)或邮件联系我们(lckf@inspur.com)。
加电无显示
故障描述：
按电源开关按键后，电源开关按键上的电源指示灯由橙色变为绿色，机箱风扇正常转动，但是显示器一直无输出。
处理建议：
检查显示器是否加电正常；
若加电正常，但显示器仍无显示输出，则检查显示器与服务器 VGA 接口是否正常工作；
如果 VGA 接口连接正常，但显示器仍无显示输出，则尝试更换显示器测试；
如果更换显示器无效，则可能是 VGA 接口异常。请尝试登陆 BMC Web 界面，打开BMC 远程 KVM（具体参考文档请参见《固件升级和配置》）确认是否显示正常；如果KVM 能正常显示，可能是主板 VGA 接口存在异常，请联系浪潮客服；如果 KVM 不能正常显示，请记录具体的告警信息；
如果以上操作无法解决，请拨打浪潮客户服务热线(400-860-0011)或邮件联系我们(lckf@inspur.com)，并告知的告警信息和故障现象。

文本分块

≤ 512tokens
and 段落结束
and 句子结束

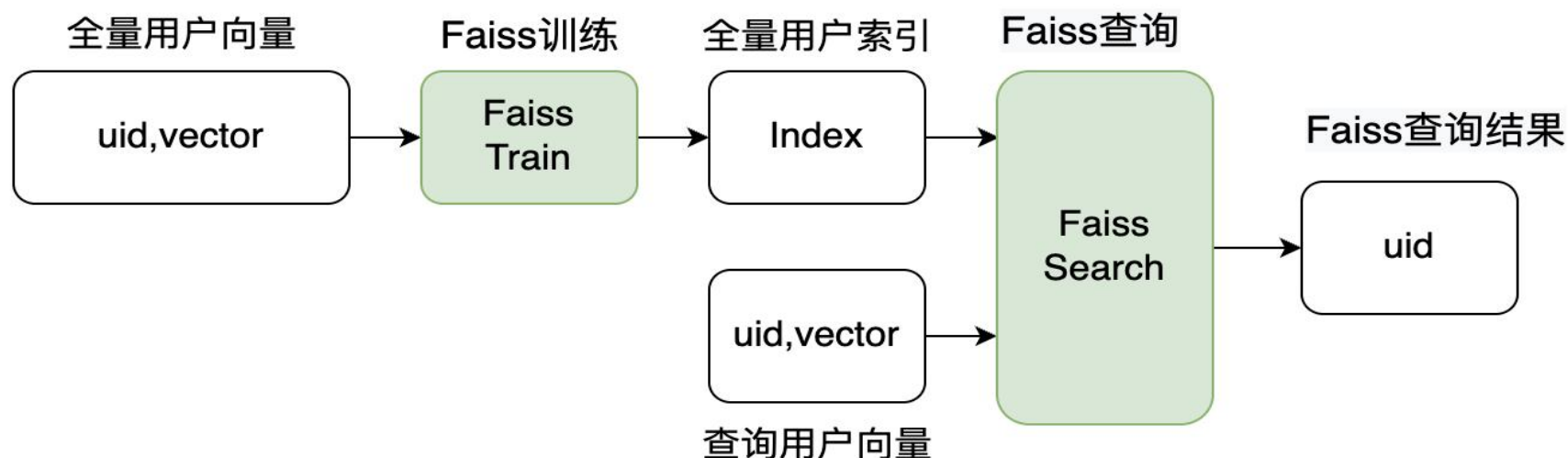
具体标题1：浪潮英信服务器 NF5280M6 用户手册 V1.4
标题2：常见硬件问题
段落：
开机不加电
故障描述：
按电源开关按键后，电源开关按键上的电源指示灯为橙色且不变绿，硬盘活动状态指示灯不亮，同时显示器无显示输出，服务器机箱风扇不转动。
处理建议：
查看服务器后窗上电源模块指示灯的状态。
电源模块指示灯不亮或为琥珀色
a. 如果电源模块指示灯不亮或为琥珀色，则表示供电异常。请检查电源插座是否能正常工作以及电源线是否安插到位，相应电源模块上的供电线供电是否正常工作；
b. 确保电源插座可正常工作，电源线安插到位且供电线可正常工作后：电源模块指示灯仍就不亮或为琥珀色，则可能为电源模块故障，请使用同型号机器同配置的电源模块更换该电源模块，排查是否为电源模块故障；电源模块指示灯变为绿色，但电源开关按键仍就为橙色，请拨打浪潮客户服务热线(400-860-0011)或邮件联系我们(lckf@inspur.com)。

电源模块指示灯均为绿色
a. 如果电源模块指示灯均为绿色，请拔掉供电线，重新拔插电源模块，安插好供电线后再次开机验证；
b. 如果拔插电源模块测试仍然无效，则使用同型号机器同配置的电源模块更换该机器的电源模块，测试是否为电源模块故障；
c. 如果以上操作无法解决故障问题，请拨打浪潮客户服务热线(400-860-0011)或邮件联系我们(lckf@inspur.com)。
加电无显示
故障描述：
按电源开关按键后，电源开关按键上的电源指示灯由橙色变为绿色，机箱风扇正常转动，但是显示器一直无输出。
处理建议：
检查显示器是否加电正常；
若加电正常，但显示器仍无显示输出，则检查显示器与服务器 VGA 接口是否正常工作；

► 基于文档阅读的问答技术实践

■ 分步骤说明--基于向量的检索

我们试验了几种语义编码模型（Embedding Model），如OpenAI Embedding API、ernie-3.0-base-zh，性能最好的是text2vec-large-chinese。向量检索数据库采用Faiss，原因是Faiss效率较高（平衡了性能和精度），1000w+条知识&单节点，检索时间为1.2s。



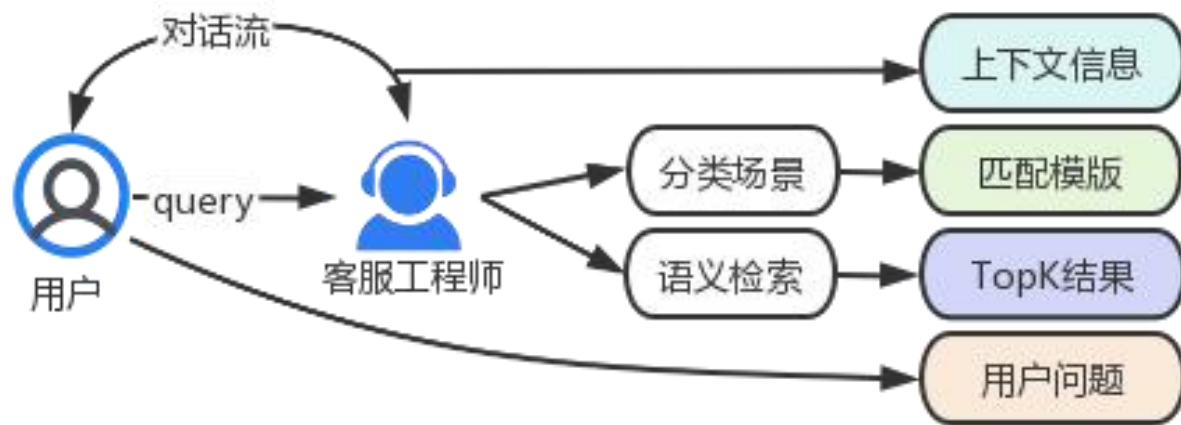
Faiss 流程图

► 基于文档阅读的问答技术实践

■ 分步骤说明--面向场景的Prompt撰写和匹配

目前浪潮信息客户服务部已梳理了相对独立的1000+的场景问题，而每个场景问题下又有不同的分支处理流程。如，硬盘不识别问题（一个场景问题），它可能的原因是没有配Raid，没有安装驱动，硬件设备故障（三条可能的分支）。

针对用户的单个问题（query），我们实际给到模型的Prompt包含四部分内容：**Prompt场景描述+问题相关背景信息+上下文信息+用户问题**。其中，场景描述、问题相关背景信息这两部分是可有，而上下文信息和用户问题是必有。

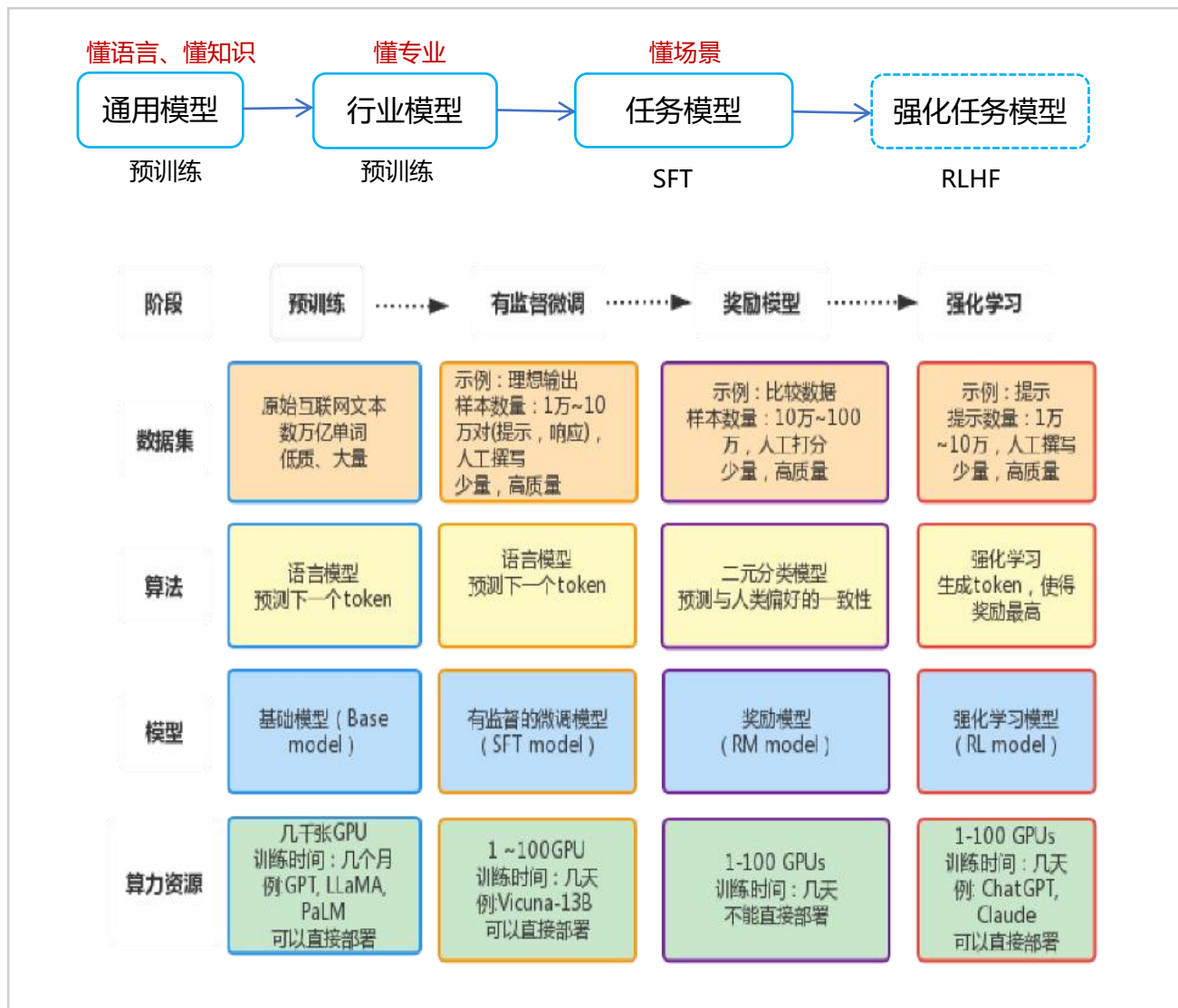


场景：硬盘不识别

场景模版：

当用户遇到的是硬盘不识别的问题，首先要询问用户有没有做Raid，如果确认是没有做Raid，则引导用户去按步骤去做Raid。用户遇到的是硬盘不识别问题，同时已经做过Raid，则需要询问用户是否根据产品型号选择了正确的驱动，并进行安装。每种产品.....

单模态客服机器人-微调训练



1. 微调训练流程：目前基于基础通用大模型微调训练垂直领域模型，业界目前通常方法也是**自监督预训练-有监督微调-强化学习**，最终得到**信息行业——客户服务LLM**。

2. 数据，目前基座模型预训练数据都在**万亿tokens**。

类别	数据量
预训练	百亿级别tokens（几十G）
有监督微调	几千~几万条/单领域
数据来源	学科书籍论文、行业网站、行业百科、企业文档、对话数据

3. 算力，算力与参数规模、训练数据量成正比，可以采用 **8 × 参数量 × tokens**来估算计算量（FLOPs）。

4. 全参数微调VS部分参数微调，算力够用的情况下，建议全量参数微调。

▶ 单模态客服机器人-客户服务LLM应用

现状：LLM、语义检索能力，采用**LLM+语义检索+prompt构造**的方法实现客服机器人

- 问题：**
- 1. 需要人工构造每个场景合适的Prompt，成本高
 - 2. 偏向于单轮对话、上下文多轮对话，**但无法进行启发式对话。**

用户	喂，你好，就是我刚刚打过电话了，就是说我现在装那个那个服务器啊，他现在就是说。不是硬盘到那就到那一步了。	说状态，不清晰
工程师	看不到硬盘是吧。	确认问题
用户	对，是不是缺少那个RAID驱动啊。	询问原因
工程师	有可能，你做RAID吗？	启发式问询
用户	做了，我看上面写的应该是RAID1是。	回答，但可能错
工程师	怎么做的RAID？或者看见那个时候有什么快捷键的提示。	启发式问询
用户	呃，有个CTRL加A还是加啥，就是能配置他应该是这个是那个集成的嘛，集成的那个主板	回答，但矛盾
工程师	CTRL加A一般不是集成的。	说出矛盾

工程师与用户的真实语音对话

用户	喂，你好，就是我刚刚打过电话了，就是说我现在装那个那个服务器啊，他现在就是说。不是硬盘到那就到那一步了。
AI	非常抱歉听到您遇到了问题，可以请您再具体描述一下遇到的情况吗？
用户	看不到硬盘
AI	您的描述，您现在遇到的问题是，在安装过程中遇到了硬盘安装的问题，无法进入下一步。根据您的描述，有几种可能的原因。
用户	最有可能是啥原因啊？
AI	根据您的描述，最有可能的原因是硬盘没有正确连接或者BIOS设置不正确。因此，我建议您先检查硬盘的连接和BIOS设置是否正确，
用户	硬盘链接了，我看上面写的应该是RAID1，是不是那个RAID驱动啊。
AI	如果您使用的是RAID1模式，那么需要安装相应的RAID卡和RAID驱动程序。您可以查看服务器的型号和RAID卡型号，然后下载相应的驱动程序，接着安装。

AI与用户的真实文本对话

▶ 总结、思考和展望

我们的总结、思考和展望

- ToB的客户服务因其对于企业有极大价值，但又因其场景复杂&用户多样，**是LLM最值得去探索落地、并且非常有挑战的领域。**
- 目前落地的探索都还停留在，原有NLP解决的问题，现在用LLM再做一遍，并没有突破原有NLP的范围，**未来需要更多的突破思维。**
- LLM固有存在的问题，如幻觉、推理时间长，并没有从根本上解决，增加知识库引入搜索，撰写prompt强化输入引导，都使得系统变的复杂、人工成本增加。**未来需要探索在算法层面更大的突破。**
- **道阻且长,行则将至。**未来，LLM落地到ToB的客户服务，核心要解决的**一是能够深层次的理解多模态的信息和知识，学习到专家工程师沉淀到文档、对话中的经验**，能够采用合适地对话策略，像真人一样提供客户服务；**二是推理成本和用户体验的问题。**

LLM的总结、思考和展望

目前，大语言模型应用在ToB客户服务场景中的现状是已被广泛使用，成为企业客户服务的有效工具，可用于自动问答系统和智能客服平台，提供快速、可扩展的客户支持服务，并降低企业的成本，提高服务质量和客户体验。但同时还存在一些限制和挑战，例如，复杂或涉及个人敏感信息的问题需要谨慎处理，需要大量的人工和技术投入来构建和改进。未来，大语言模型技术将有如下的进一步思考和展望：

- **发掘更多的ToB场景**，如面向销售和营销等方向，以扩大大语言模型技术的适用范围和商业化。
- **探索更复杂的对话场景**，如多轮对话、人机交互等，以提高服务质量和精度。
- 结合AI技术的其他分支和新技术的发展，如语音识别、语义理解等，**以提高对话质量和便利性。**
- **进一步探索大语言模型技术与其他系统的集成**，如CRM、ERP等，以更好地支持企业客户服务，提高效率和准确性。
- **关注大语言模型技术的可信度和安全性**，赋予用户对于隐私保护的可控性等，以促进这一技术的发展和应用。

感谢聆听

