

Large Language Model Deployment

企业AI大模型全链解决方案

借力AI实现企业（再次）腾飞

前言

- 合作共赢，打造企业核心竞争力

1. 当前社会&竞争的共识

每家企业，都必须尽快引入AI大模型

2. 先借助外力，快速让AI落地

小栖原生AI团队 + 贵司业务负责人 = 极速降本增效 + 无限可能

3. 再逐步培养、配备更多AI人才

小栖AI企业培训 + 沉浸式业务实练 = 组织形成更多AI人才

4. 业务导向，高质量低成本干大事

小模型 + 工程化能力 + 高拓展框架 = 花小钱干大事



整体架构&实施方案

- 从硬件到业务层，AI全链解决方案

应用层

App / Web / API、业务系统集成

AI引擎层

数据智能标注、模型训练（推理框架）、向量化引擎、模型评估体系

硬件层

GPU集群 / 存储 / 网络

数据层

结构化/非结构化数据库、知识图谱存储

安全性

私有化部署
端到端加密
访问控制
监控日志

硬件配置-最小硬件投入获得最佳效果

- 基于业务评估后规划最优硬件解决方案

模型	部署方式	配置需求	总显存大小	算力(TFlops)	cuda核心数
DeepSeek-R1671B 完整版	GPU部署	H20*8	96G*8(768G)	40	4096
DeepSeek-R1671B 完整版	GPU部署	H100*8	80G*8(640G)	51	14892
DeepSeekQ4量化版	GPU部署	H20*4	96G*4(384G)	40	4096
DeepSeekQ4量化版	GPU+CPU部署	H20*1	96G	40	4096
DeepSeek蒸馏版本Qwen32B	GPU部署	H20*1	96G	40	4096
DeepSeek蒸馏版本Qwen7B	GPU部署	H20*1	96G	40	4096
table Diffusion(图像工具)	GPU部署	H20*1	96G	40	4096
DeepSeek蒸馏版本Qwen32B	GPU部署	H100	96G	51	14892
DeepSeek蒸馏版本Qwen32B	GPU部署	A40	48G	37.4	10752

核心模块

- 模型部署与优化

实时指标（响应延迟、
资源占用）、自动化扩
缩容

性能监控

容器化（Docker/
K8s）、模型微调
（LoRA/P-Tuning）

部署方案

预训练模型，如
DeepSeek、LLAMA 2、
ChatGLM或自研模型

模型选型

说明：

- 模型选型可基于实际需求和成本自由选择；
- 多模型间可以自由切换；
- 已有成果可共享（无需二次训练）；

核心模块 - 大数据加速标注

通过实体关系模型+DSL描述语言来拆分不同段落

标注前

患者因反酸、腹胀入院，进食后明显，伴间断少量呕吐，后症状持续，就诊北京安贞医院，完善检查，胃镜(2021-09-27，外院):胃内大量黄绿色潴留液，胃窦变形狭窄，后壁可见溃疡，幽门无法通过;活检病理我院会诊:(胃窦后壁)低分化腺癌，伴印戒细胞癌，Lauren分型:弥漫型;免疫组化结果显示:EGFR(-),HER2(0),MLH1(+),MSH2(+),MSH6(+),PMS2(+),pan-TRK(-),PD-L1(22C3)(CPS=0),CK(+), 分子检测原拆分前位杂交结果显示:EBER(-)。腹部增强CT(2021-10-09，外院)示:胃壁增厚、胃液潴留。胆囊不均匀增厚，胆囊底部软组织病变。PET-CT(2021-09-29，外院)示:胃窦-十二指肠走行区团块状软组织密度影，SUVmax52x33x62mm，其周多个结节样代谢增高灶;腹膜后代谢增高肿大淋巴结，大小15x8mm。阅片可见腹膜后肿大淋巴结位于主动脉旁。肿瘤标志物(2021-09-29，外院):示:CA199142U/ml。我院多学科查房考虑患者分期cT3N1，建议评估腹膜状态

2021-11-11行腹腔镜探查、HIPEC+PORT术。术中见PCI=4，左下腹壁活检病理考虑为低分化腺癌，术中快速细胞学可见个别可疑癌细胞。诊断cT4aN1M1，P1Cy1，遂术中行多西他赛160ml热灌注化疗。转化治疗:2021-1-26至2021-12-03行第1周期转化TS+T方案化疗，具体方案:紫杉醇80mg，iv，D1、8;紫杉醇30mg，ip，D1、8;替吉奥120mg，p.o.分两次(早60mg，晚60mg)，D1-D14;Q21d。化疗副反应:口腔/喉溃疡，3级，表现为d10开始溃疡明显，影响进食，每日口服肠内营养不足500ml，化验结果示低蛋白血症、低钠低钾血症;另伴恶心、腹泻、面部及后背皮疹伴瘙痒。予以静脉营养及白蛋白对症支持1周后患者口腔溃疡营养状况较前好转，已进流食。

于2021-12-26、2022-01-04行第2周期转化化疗，化疗方案为“TS+T”，考虑到上次化疗出现3级溃疡不良反应，与替吉奥相关可能性大，予以替吉奥减量，具体化疗药物和剂量:紫杉醇80mg，iv，D1、8;紫杉醇30mg，ip，D1、8;替吉奥100mg，p.o.分两次(早40mg，晚60mg)，D1-D14;Q21d。化疗副反应:D12出现喷射性呕吐1次，为胃内容物，伴腹泻、食欲减低，每日口服半流食、肠内营养600ml左右。考虑梗阻加重，于2022-1-17给予空肠营养管置入。2022-01-19第1次评估:复查腹盆CT可见病变较前缓解，ycT4aN2M1，经腹腔化疗灌洗未引流出液体，Cy状态不详。复查肿瘤:CA19948.7→10.87;CA12579.4→58.8。评估获益SD。

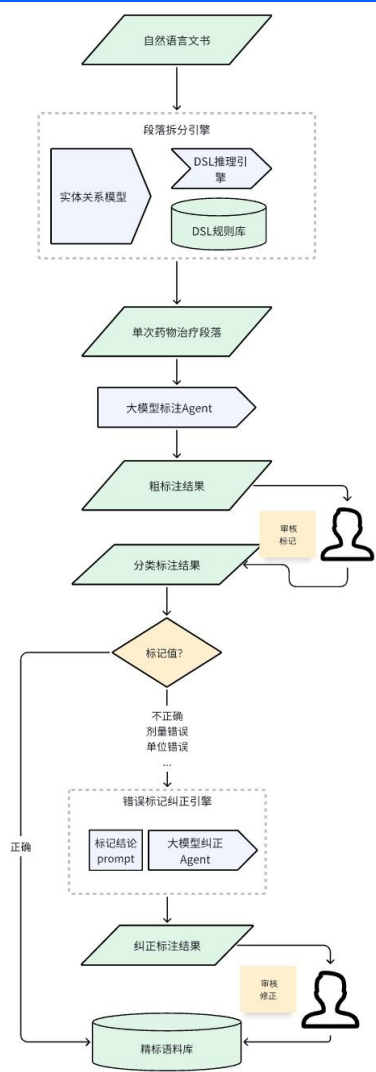


标注后

2021-11-11行腹腔镜探查、HIPEC+PORT术。术中见PCI=4，左下腹壁活检病理考虑为低分化腺癌，术中快速细胞学可见个别可疑癌细胞。诊断cT4aN1M1，P1Cy1，遂术中行多西他赛160ml热灌注化疗。

转化治疗:2021-1-26、2021-12-03行第1周期转化TS+T方案化疗，具体方案:紫杉醇80mg，iv，D1、8;紫杉醇30mg，ip，D1、8;替吉奥120mg，p.o.分两次(早60mg，晚60mg)，D1-D14;Q21d。化疗副反应:口腔/喉溃疡，3级，表现为d10开始溃疡明显，影响进食，每口服肠内营养不足500ml，化验结果示低蛋白血症、低钠低钾血症;另伴恶心、腹泻、面部及后背皮疹伴瘙痒。予以静脉营养及白蛋白对症支持1周后患者口腔溃疡营养状况较前好转，已进流食。

于2021-12-26、2022-01-04行第2周期转化化疗，化疗方案为“TS+T”，考虑到上次化疗出现3级溃疡不良反应，与替吉奥相关可能性大，予以替吉奥减量，具体化疗药物和剂量:紫杉醇80mg，iv，D1、8;紫杉醇30mg，ip，D1、8;替吉奥100mg，p.o.分两次(早40mg，晚60mg)，D1-D14;Q21d。化疗副反应:D12出现喷射性呕吐1次，为胃内容物，伴腹泻、食欲减低，每日口服半流食、肠内营养600ml左右。



核心模块 - 大数据加速标注

1. 大模型预标注后，交给专人审核 (无需纠正具体字段，只需标记问题的类别)

任务标注

返回任务列表

ID: 221	已标注
ID: 222	已标注
ID: 223	已标注
ID: 224	已标注
ID: 225	已标注
ID: 226	已标注
ID: 227	已标注
ID: 228	已标注
ID: 229	
ID: 230	

< 11 / 156 >

2021-12-26电子肠镜示：乙状结肠冗长。病检示：（胃窦+胃角）粘膜组织慢性炎症伴活动及轻度肠化及溃疡形成。局灶可见游离腺上皮细胞巢增生伴高级别上皮内瘤变，未见粘膜肌层。病变组织少，建议结合临床。【免疫组化结果：CK（+），HP抗体（-），Ki67（index50%）】。又与2021-12-15XXX肿瘤医院检查CT示：1、胃体、胃窦胃壁不规则增厚，考虑为癌，肿瘤于肝左叶、胰脏关系密切。

标注详情 - 0

检测时间:

检测方法:

检测平台:

免疫组化检测详情 - 0

免疫组化名称: CK

检测抗体名称:

检测结果: +

免疫组化检测详情 - 1

免疫组化名称: HP

检测抗体名称:

检测结果: -

免疫组化检测详情 - 2

免疫组化名称: Ki67

检测抗体名称:

检测结果: index 50%

当前字段: 免疫组化名称

☐ 免疫组化名称错误

☐ 把不同切片的免疫组化信息整合到检测结果中, 例如:A+,B+,C+,D-,E-

☐ 包含非免疫组化名称的切片信息

☐ 缺少免疫组化详情

完成标注

删除标注

病历:

%s

结构化信息:

```json

%s

```

我需要你帮我核实上述从病历中提取的结构化信息是否有误，只需要处理以下两个维度的问题：

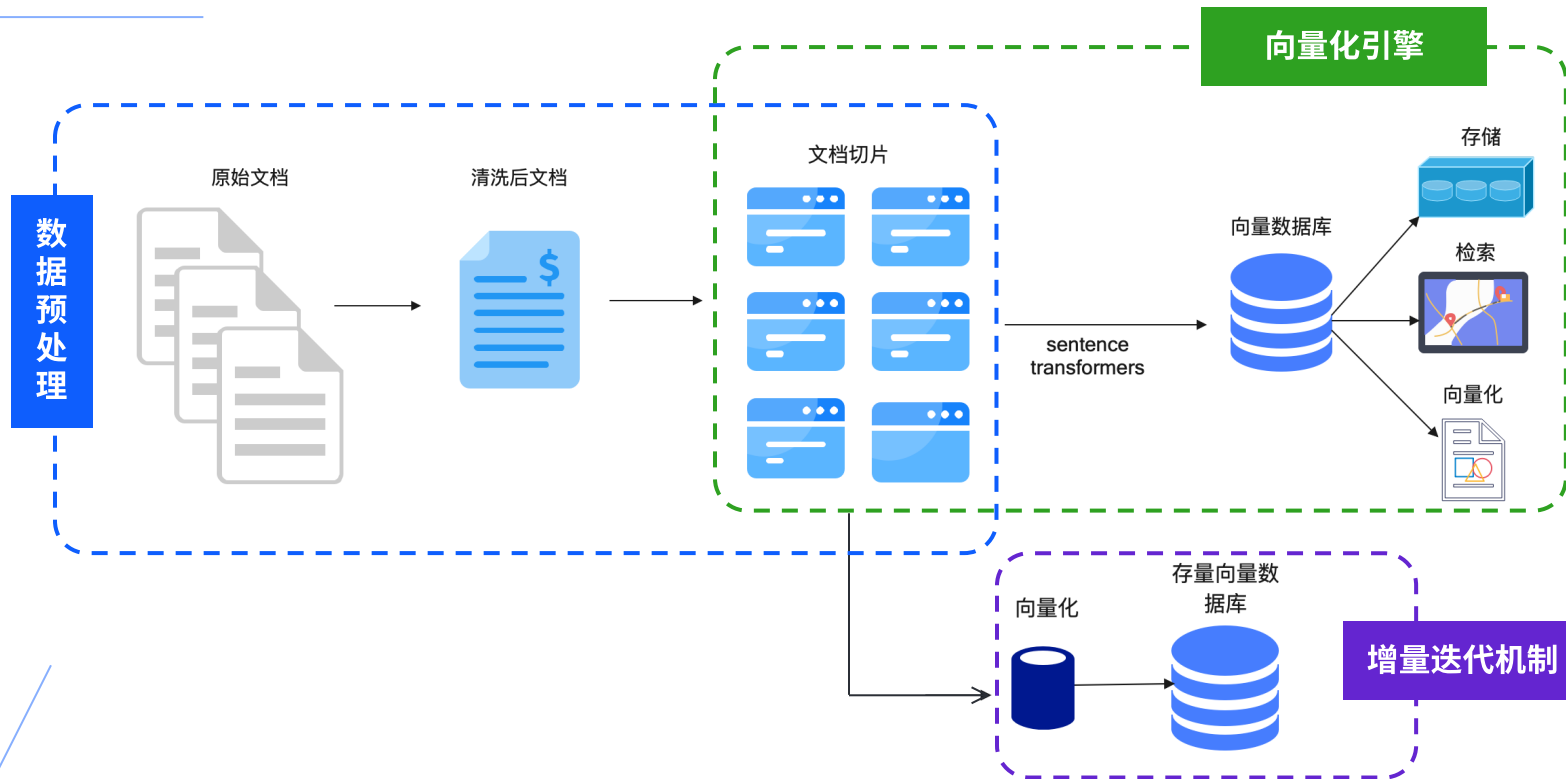
1. 提取信息中，是否有原文中计划执行或建议执行的内容，这类内容不应提取。
2. 核对`方案开始时间`是否提取正确，如果病历中提到多个进行治疗的时间点，则应该将多个时间合并，以`,`分隔。
3. 核对`药物剂量单位`字段是否提取正确。它应该表示药物剂量的单位，如ml，mg等。如果没有写单位，则取值空字符串。
4. 核对是否混淆了`用药频率`、`给药时段`和`药物剂量单位`。常见的用药频率类似：qid，每日四次；bid，每日两次；tid，每日三次；qod：隔日一次。而给药时间一般会如：`d1,6`：第一天和第6天；`1-5`：第一至第五天。

一步一步思考，如果原结构化信息正确无误，则以`CORRECT`结尾结束回答；如果原结构化信息有误，则指出有误的地方，并且给出修正后的结构化信息。

2. 根据问题标记错误多的类别， 设计自动纠正agent的prompt

核心模块-知识库向量化

- 预数据处理/向量化引擎/增量迭代机制



核心模块-训练任务管理平台wandb

workspace

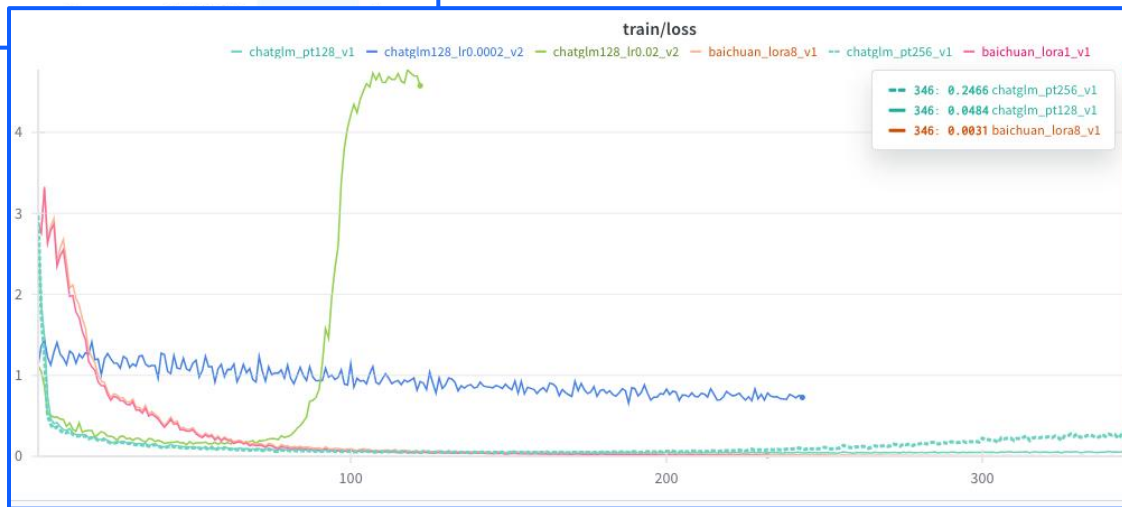
Search runs

Filter Group Sort Tag Move Create Sweep

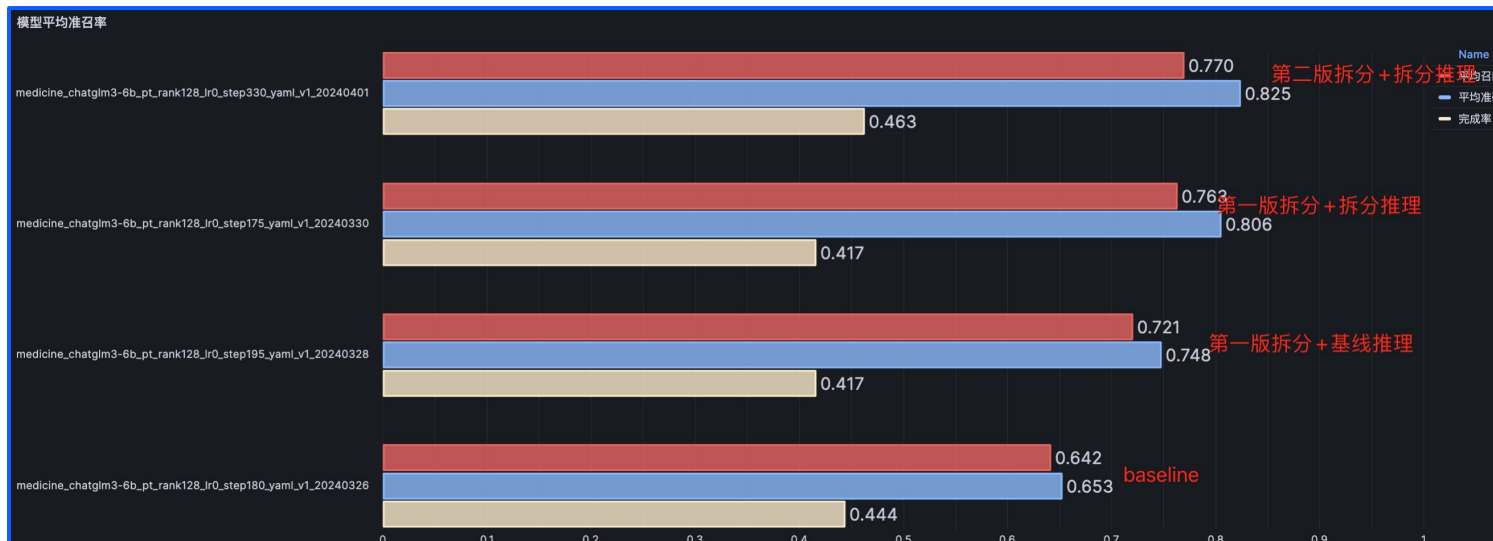
☐	eye	Name (6 visualized)	State	Notes	Use	Tag	Created	Runtime	Sweep	_from_
☒	eye	chatglm_pt128_v1	Finished	Add notes	498811		1mo ago	1d 11h 18m	-	-
☐	eye	chatglm128_lr0.000...	Crashed	Add notes	498811		2mo ago	2d 23m 15s	-	-
☒	eye	chatglm128_lr0.02_v2	Killed	Add notes	498811		2mo ago	1d 30m 47s	-	-
☒	eye	baichuan_lora8_v1	Finished	Add notes	498811					

Jobs

Automat.



核心模块-模型评估系统



Inspect value

```
1 [
2   {
3     "标本部位": "贵门",
4     "标本名称": "",
5     "标本类型": "组织学",
6     "标本获取方式": "内镜活检",
7     "标本采集时间": "10天前",
8     "病理报告时间": "",
9     "组织机构": "宝应县人民医院",
10    "切缘": [],
11    "病理": [
12      {
```

Inspect value

```
"标本采集时间": "2013-08-12",
"病理": [
  {
    "HPV": [],
    "复发": [],
    "病理诊断": [
      {
        "病理诊断名称": "低分化癌",
        "病理诊断存在状态": "肯定"
      },
      {
        "病理诊断名称": "鳞状细胞癌",
        "病理诊断存在状态": "可能性大"
      }
    ]
  }
]
```

base_model_name

chatglm3-6b
chatglm3-6b
chatglm3-6b
chatglm3-6b
chatglm3-6b
chatglm3-6b

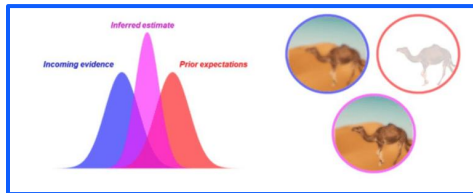
模型完成情况统计

	record数	平均召回率	平均准确率	开始时间 ↓	结束时间
chatglm3-6b	36	0.0314	0.0889	2024-02-02 16:12...	2024-02-02 16:20...
chatglm3-6b	36	0.689	0.676	2024-01-27 09:30...	2024-01-27 09:35...
chatglm3-6b	36	0.705	0.686	2024-01-27 09:18...	2024-01-27 09:26...
chatglm3-6b	2	0.500	0.500	2024-01-23 12:38...	2024-01-23 12:38...
chatglm3-6b	36	0.717	0.633	2024-01-18 12:57...	2024-01-18 13:15...
chatglm3-6b	36	0.517	0.564	2024-01-16 23:19...	2024-01-16 23:35...

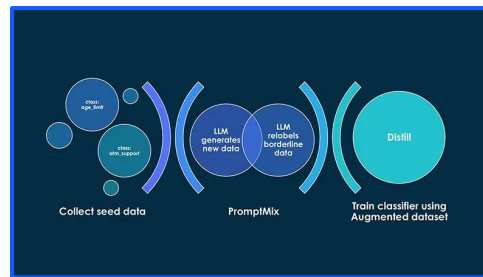
核心模块-低成本推理&模型量化压缩



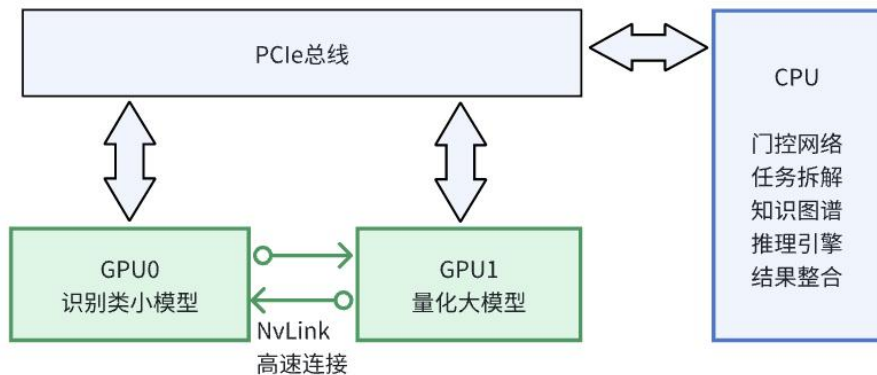
gguf量化



多LoRA联合推理



模型蒸馏



推理硬件需求

- 32CPU
- 64G内存
- 4GPU
- 单个GPU显存不小于24G
- TFLOPS不小于30
- Nvlink连接速率不小于100G/s

核心模块

- 交互系统设计



前端架构

- 低代码平台配置
- 多终端适配（PC/移动）



后端服务

- RESTful API
- 消息队列
(Kafka/RabbitMQ)



用户体验

- 对话式UI
- 意图识别
- 多轮会话管理

数据安全与合规

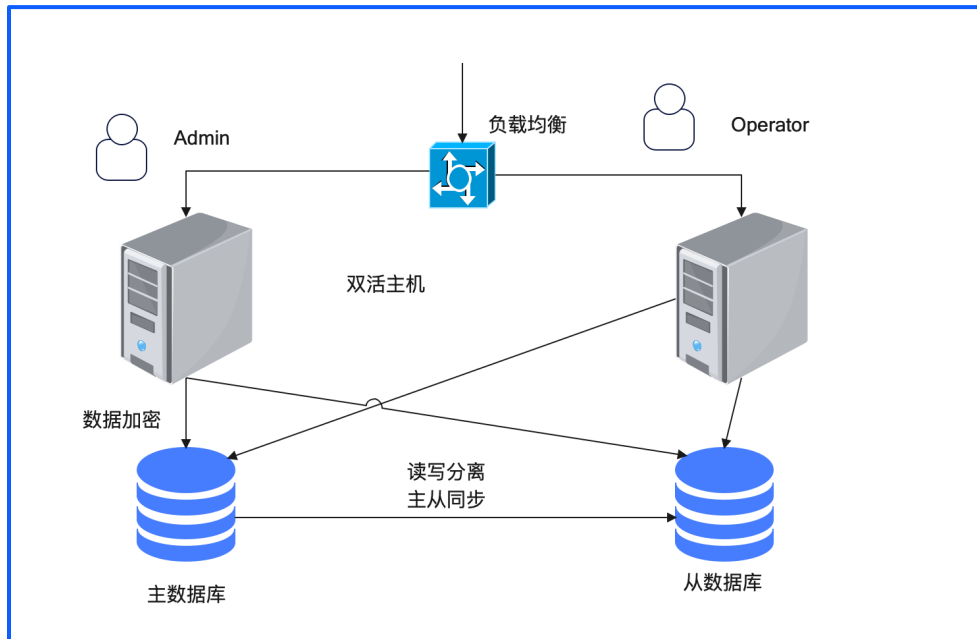
- 技术措施 & 灾备方案

技术措施:

- 传输加密 (TLS/SSL)、静态加密 (AES-256)。
- 角色权限 (RBAC)、审计日志 (ELK Stack)。

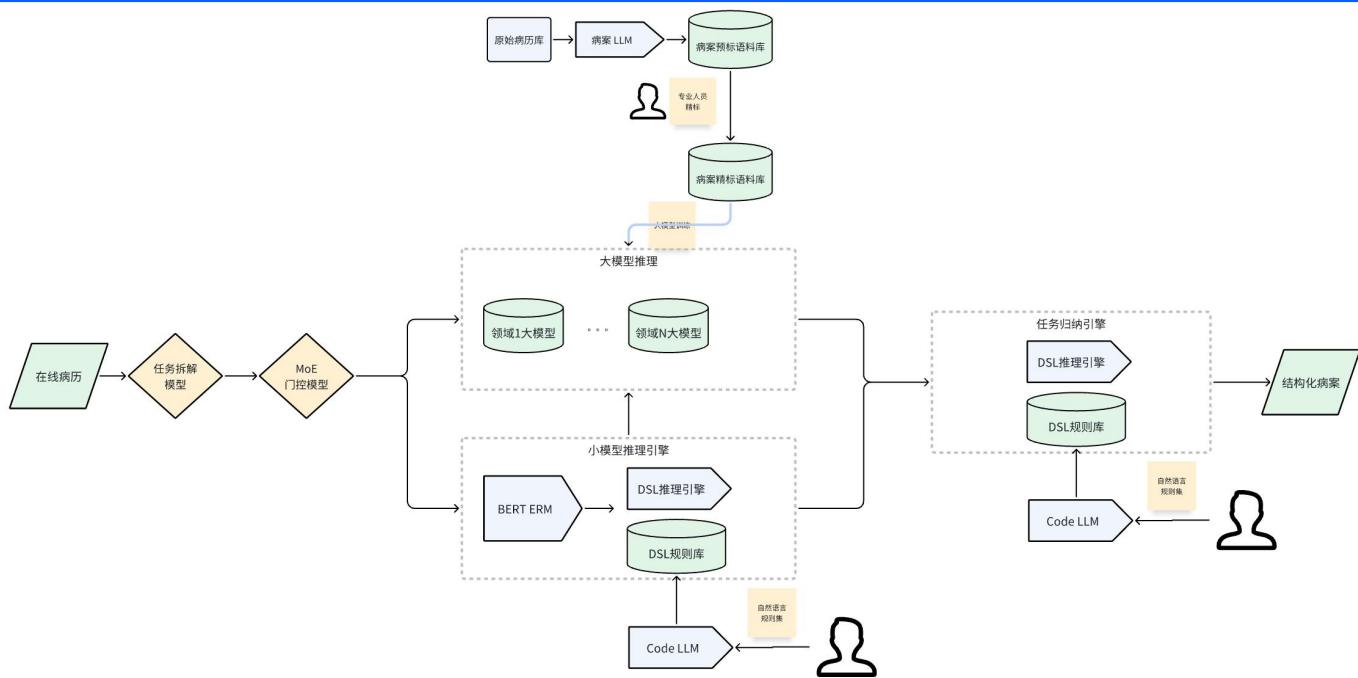
安全架构:

- 双主机多活方案+负载均衡
- 主从数据同步与备份
- 核心数据加密存储
- 独立的角色权限管理



成功案例

- 医疗专病库系统架构和流程



项目管理与支持

- 专业的服务流程和服务保障



实施流程



服务人员



服务支持

- 7×10（特殊情况实时对应）
- 定期系统健康检查
- 定制化升级



合作共赢