

刘佳琛 ♂

浙江 | ai算法工程师 | 3年经验

☎ 电话: 18039296071

✉ 邮箱: ljc3424@163.com

工作经验

2023-06 ~ 2025-06	腾讯科技股份有限公司	ai算法工程师
2022-07 ~ 2023-06	腾讯科技股份有限公司	数仓工程师
2022-01 ~ 2022-06	杭州扶志科技有限公司	算法工程师

技能特长

大模型相关

- 1. 架构设计: 基于 RAG 与 Agent 技术构建企业级智能系统、实现业务场景的 AI 解决方案
- 2. 大模型原理: 熟练掌握DeepSeek (MOE、MLA、MTP等)、Hunyuan、GPT、LLama等主流LLM内部核心原理
- 3. 大模型微调: FULL Finetune 全参数微调, Lora策略的低成本微调, DPO使用偏好进行策略优化
- 4. 大模型工程能力: 熟练掌握Prompt工程、PDF处理pipeline、基于RAG的召回优化与资源压缩
- 5. 评估体系: 通过实际业务建立多维评估机制、确保模型性能与业务需求对齐
- 6. 框架开发: 具有 LangChain、LangGraph Multi Agent 的开发经验

算法相关

- 1. 掌握大模型底层原理: 旋转位置编码、多头注意力、MOE混合专家模型、多 Token 预测等
- 2. 掌握 NLP 方向经典架构: Word2Vec、RNN、LSTM、Transformer、Bert等
- 3. 熟悉机器学习相关算法: AC自动机、KNN、线性回归、逻辑回归、K-means 聚类
- 4. 熟悉 CV 方向卷积神经网络的基本原理, YOLO架构

数据相关

- 1. 掌握 Spark、Flink 大数据计算引擎、具备离线/实时项目落地经验
- 2. 熟练常用的大数据组件 Kafka、Flink、Spark、Hive、StarRocks等
- 3. 掌握 MySQL 常用的业务数据库

项目经验

2024-04 ~ 2025-06	ChatBI-Multi-Agent数据中台	ai算法工程师
-------------------	------------------------	---------

项目背景

基于公司现有组织架构, 无法满足业务部门众多需求的开发, 以及在精品、联运、增长、广告等核心场景下都无法通过离线/实时数仓动态适配全部的需求, 一味堆积报表数量和数据分析师的数量并不是最优解, 故需要研发一个面向公司业务智能数据中台从数仓开发、业务排错、营销方案评估等多方面来加速业务问题定位以及日常取数效率, 洞察数据潜在联系, 更好地为业务决策做支撑, 同时压缩需求从提出到上线的链路, 反向促进数据模型的迭代

方案

意图识别与要素提取

- 1.基于 Fine Tuning+Prompt 的方案, 通过 llm 能力, 对 query 进行意图识别和要素字典提取
- 2.对关键槽位为空等异常情况的识别处理, 通过上下文对话内容进行用户补全, 指代消解生成对应json模板, 降低模型生成复杂度
- 3.基于指标需求和业务意图要素构造共近8000样本, 通过大模型增强至10000+条, 并通过人工校验。训练集: 测试集=5: 1
- 4.自建数据集进行Lora微调, 测试集准确率达99.1%

数据，库，表，列精准召回

- 1.基于数仓建模思想的问题分流，使得简单问题快速解答，高难度问题解决链路健全
- 2.建立业务标题树实现多源知识的精准召回，建立多级规则引擎，将嵌入的动态问题改写，保证召回关键词准确
- 4.实现了核心节点关键词match代替混合检索，降低算力消耗，提高召回度
- 5.基于DPO Tuning对Rerank模型进行微调，过滤业务噪音

SQL生成

- 1.基于 Few Shot+Prompt 的方案，基于子query的SQL生成以及多query SQL汇总，为SQL生成提供依据避免自然语言直接生成SQL的高幻觉度
- 2.引入SQL Glot语法校验、Explain Analyze性能校验+用户行列权限，确保SQL可执行性和安全性
- 3.执行sql，并调用Echars Agent生成可视化图表面向用户展示

效果

- 1.实现自然语言到SQL的自动转换，操作时间从30分钟(需拖拽字段/写SQL)缩短至10秒
- 2.实现SQL生成准确率达94.2%+
- 3.支撑日均3000+次即席查询，大大缩短业务人员取数延迟
- 4.基于multi agent协同工作+Supervisor监督机制提高项目稳定性，确保问题解决链路健壮性与弹性变更能力

2023-06 ~ 2024-04

RAG智能客服(应用小宝)

ai算法工程师

项目背景

应用小宝旨在为玩家提供实时、个性化、高准确率的游戏攻略支持，覆盖角色培养、关卡通关、装备搭配、战术策略等核心需求。通过AI降低游戏学习成本，让玩家专注乐趣，而非繁琐攻略搜索，同时为游戏厂商提供玩家行为分析的新维度

方案

- 1. 收集IEG部门，合作游戏厂商的用户手册和专业知识
- 2. 使用 marker 对 pdf 进行内容层级切割，转为makerdown格式，进行混合分块，数据清洗，去重，doc2query 处理
- 3. 对query进行改写，多轮对话补充，同义词增强，使用 ES 作为向量库对所有query的 embedding 向量进行存储
- 4. 通过AC自动机算法识别出问题中的实体，从预先定义好的图数据库中查询游戏和游戏，游戏和道具等等之间的关系，形成关列表
- 5. 将关列表作为约束条件，进行多路召回，返回内容进行 rerank，最后放入llm+prompt进行生成
- 6. 通过 Lora，DPO，FULL Finetune 对模型进行微调训练，通过建立多级评估体系对RAG系统进行打分

效果

- 1. 评估准确率在 98%左右
- 2. 查询 P95 < 500 ms

2022-07 ~ 2023-06

应用宝会员体系数仓建模

数仓工程师

项目背景

会员福利数据旨在详细描述会员福利相关的数据指标和体系，支持会员管理和福利发放等业务需求，包含了会员等级、积分、付费、商城等多个方面的数据指标和定义，例行化处理大盘成本优化和会员客诉问题

方案

- 1.指标体系：详细定义会员等级、积分、付费、商城等多个方面的指标，包括定义、粒度、口径等
- 2.数仓架构：积分发放，会员体系，宝券等数仓相关架构设计
- 3.数据质量：通过编写python脚本通过max、sum，空值率等指标对修改表进行统一对比，确保数据质量

例行化职责

- 1.成本优化：制定告警规则(查询失败率>10%，查询成本>5元，查询次数>100)，配置看板，每日进行告警，优化
- 2.会员客诉问题：每周四统一处理客诉，排查用户是否是应用宝染色用户，用户绑定的是qq账号还是微信账号等问题

效果

- 1.通过详细的指标体系，可以全面监控和分析会员福利相关的数据，支持会员管理和福利发放等业务决策
- 2.明确的数仓架构和埋点规范可以提高数据的准确性和完整性，支持高效的数据分析和报表生成
- 3.通过积分、付费等指标的监控，可以优化会员福利策略，提高会员的活跃度和满意度

项目背景

由于原始数仓在进行指标开发时，并未合理区分定义各数仓层，dwd、ads各层区分不明，各主题域表阶段性调整修改导致出现很多同含义的ods，由此开发的数仓较为混乱，因此展开针对各业务对象的数仓重构

方案

1. 业务方对各主题域各业务对象指标对应表进行梳理，由数分、产品经理等业务部门明确核心业务对象，定义其关键指标
2. 数仓部门对ods表进行整合，新建最新ods表，保障数据质量，对历史ods表进行下线处理
3. 将原有分散的指标逻辑整合为《指标口径手册》，并基于此进行方案重定义分层处理

效果

1. ods表减少30%，dwd表冗余字段减少25%
2. 减少重复开发以及数据口径问题争议，再开发逻辑指标复用大大减少开发时间

项目背景

纠错大师内所需图像失真恢复、图像分割、OCR、试卷还原以及题库匹配等功能，70%功能需要结合深度学习模型进行研发。实现图像实时分割，降低OCR流程，提高识别准确性

负责事项

- 1.使用labelme对图像数据进行打标
- 2.几何失真模拟(正弦/余弦波形变换、旋转、折叠)，光学特性模拟(高斯滤波、对比度调整、反转)，基于2K真实数据集的增强扩展
- 3.基于YOLOv5的实时分割框架，使用官方预训练模型初始化，结合业务自适应微调
- 4.采用paddle的PP-OCRv2模型微调，进行文字识别，调用用科大讯飞公式识别API构建LaTeX-OCR训练集，存储到OSS bucket中，定期人工手动核验，训练LateXOCR模型进行公式识别
- 5.基于YOLOv5的印刷/手写体分离，印刷体用通用OCR识别，利用OpenCV高斯去噪，自适应二值化等方式使得图片更平滑，初步实现试卷去手写的还原操作

效果

- 1.PDF失真恢复mAP值达94%
- 2.实现秒级处理速度(<1s/页)
- 3.提高了整条答疑链的性能，提高了用户体验度

教育背景

自我评价

工作积极认真，细心负责，善于在工作中提出问题、发现问题、解决问题，有较强的分析能力；勤奋好学，踏实肯干，动手能力强，认真负责，有很强的社会责任感；坚毅不拔，吃苦耐劳，喜欢迎接新挑战。