姜晓东

男 | 38岁 13911314163 10604152@qq.com

11年工作经验 | 大模型算法 | 期望薪资：30-40K | 期望城市：北京

个人优势

1. 精通LLM应用开发与LoRA微调，能够高效实现模型的优化与定制。
2. 熟练运用vLLM、TensorRT，SGLang显著提升模型推理效率。
3. 熟练掌握LangChain、LlamaIndex、RAGflow框架以及RAG技术，构建高效的知识驱动应用。
4. 熟练通过Prompt Engineering进行Prompt迭代优化，提升模型输出质量。
5. 了解RLHF强化学习PPO、GRPO、DPO技术，助力模型性能提升。
6. 掌握分布式训练框架DeepSpeed、Megatron- LM、Colossal-AI来加速模型训练过程。
7. 熟练使用FastAPI与Docker快速部署应用，高效开展MVP最小可行性验证。
8. 熟练运用Neo4j等图数据库构建知识图谱，实现复杂数据关联分析。
9. 对Agent、MCP、A2A等技术有一定了解与应用。

工作经历

成都青牛泰科信息科技有限公司 算法工程师 2024.07-2025.04

负责大模型应用系统的算法设计与优化，主导构建两个核心系统：智能客服系统（RAG + 图谱）与专利语义检索系统，实现大模型与向量检索、知识结构融合的深度协同，显著提升复杂任务的召回率与响应质量。

1.RAG + 知识图谱客服系统：构建包含 Milvus + Neo4j +BM25算法的混合召回检索架构，引入 gte-Qwen2 与 bge-base-zh 中文嵌入模型；融合 LangChain 记忆模块、多轮语义识别与 Cross-Encoder 精排；基于 Qwen2-32B 微调与 vLLM 部署实现低延迟对话生成

2.专利语义检索系统优化：构建高质量三元组数据集，使用 Triplet Loss + LoRA 微调嵌入模型，提升 Top-K 检索效果；使用 mistral AI 与 LLaMA-Factory 优化权利要求摘要生成；基于 QwQ-32B 等大模型构建 LLM 打分排序系统，配合精细化 Prompt 工程提升 Top-20 召回率

3.大模型微调与推理部署：使用 Unsloth 微调 DeepSeek / Qwen 系列模型，结合 vLLM 实现大模型的高效部署与快速迭代测试

4.Prompt Engineering + 多维度理解：针对排序任务，从简单打分 prompt 演化为基于“技术创新点 + 功能 + 领域”分权重结构化模板，增强模型对复杂语义差异的感知与排序能力

马恒达北京信息技术股份有限公司（Google项目） 算法工程师 2021.09-2024.05

1.熟练使用 Google 内部 colab, g3，cider，bigstore，gcloud，vizier，Xmanager 工具以及 sql 语句等。

2.熟练使用 Numpy,Pandas,Matplotlib 对时间序列数据进行清洗，去噪，连续性填充，分析等预处理工作。

3.熟练使用时间序列模型的机器学习模型 sklearn及 pytorch，tensorflow 等框架的深度学习模型进行浆果产量预测实验，评估，fine-tuning。

4.熟练使用 cider对g3环境下的代码进行更新。

5.熟练使用 github 进行代码迁移升级。

6.熟练使用 Xmanager/Vizer 进行模型微调。

京北方信息技术股份有限公司 NLP算法工程师 2018.03-2021.09

1.使用 Numpy、Pandas 等进行数据清洗处理工作；

2.使用 Tensorflow,Pytorch,keras,sklearn 等深度学习框架进行模型搭建；

3.使用 Python 语言对非结构化数据，进行数据的预处理，数据预标注等工作；

4.根据业务需求使用自然语言处理技术完成命名实体识别（NER）、文本分类、情感分析、文本摘要、机器翻译等；

5.智能保险问答中使用 BiLSTM+CRF/BERT+CRF模型进行 NER ,使用 Bert模型进行命名实体审核；

6.在情感分类/机器翻译/舆情分析等任务使用 RNN/LSTM/GRU/ Bert 等模型；

7.使用 TextRank 、seq2seq、PGN等模型进行文本摘要任务，使用 gensim 工具训练词向量；

8.使用 VisluaBERT、ViLBERT 模型基于文本和图像/视频进行多模态的多分类任务；

9.了解BERT-wwm在中文领域词向量处理方法,以及XLnet处理长文本；

10.了解动态量化,微调,知识蒸馏,剪枝 等优化方法。

知学云（北京）科技股份有限公司 python开发 2015.06-2018.03

工作描述：

1． 负责数据开放平台的开发，支持各个业务方使用。

2． 负责开发数据收集、数据处理、数据解析、入库等服务接口。

3． 负责开发人才库接口，提供便捷的人才分类、人才管理等。

4． 负责公司定制化项目的开发。

北京欧美思教育科技发展有限公司 Python 2014.05-2015.05

工作描述：

1. 对产品的原始需求做总结分析,提供技术解决方案. 参与开发和维护后端服务框架.

2. 负责与前端开发人员合作, 完成系统前后端通信的 API 设计和开发实现.

3. 持续重构与维护组件保持可用性和稳定性.

4. 与测试人员沟通, 确认模块的功能与开发质量.

5. 为客户提供数据支持.

项目经历

专利权利要求检索（查重）与LLM排序增强 小组算法leader 2024.11-2025.04

构建面向专利权利要求书的语义检索系统，融合向量检索、嵌入模型微调、大模型排序与微调优化，显著提升 Top-K 召回性能。

核心工作与优化策略：

1. 高质量嵌入向量构建与存储（Milvus）

使用 gte-base-zh 与 gte-Qwen2-1.5B/7B-instruct 对专利摘要、说明书与权利要求进行语义编码、向量数据存入 Milvus，实现高效 Top-K 相似度检索

2. LLM摘要增强生成（Recall 提升）

基于 LLM API（如 mistral AI）生成权利要求摘要，提升检索文本的语义代表性、使用 LLaMA-Factory 进行有监督与无监督微调，增强摘要生成质量，优化整体召回表现

3. 嵌入模型微调（Triplet Loss + LoRA）

构建三元组样本集，采用 Triplet Loss 训练嵌入模型，增强相似度分数的判别能力、应用 LoRA 精调方案，提升嵌入表达效果

在 Top-100 / Top-50 / Top-20 任务上对比实验，分析不同召回策略性能提升

4. 基于 LLM 的排序优化

在 Top-100 检索结果中引入 LLM 打分排序，进一步优化 Top-50 / Top-20 精度

替换基础的rerank模型为LLM模型，提升排序阶段召回率

5. Prompt Engineering 策略进化

、Prompt 从早期的简单“打分对比”演化为包含技术创新点、功能应用、技术领域权重调控的复杂模板

、提升 LLM 对文本语义细节的理解与排序能力

6. 大模型微调与部署优化

使用 Unsloth QLoRA精调 DeepSeek-R1-Distill-Qwen-32B 与 QwQ-32B-bnb-4bit，提升特定任务表现、结合 vLLM 部署，显著加速推理响应并支持高并发测试场景

项目成果：

• 摘要增强 显著提升原始语义检索的召回率（Recall）

• Triplet Loss + LoRA 微调模型 大幅优化 Top-100 / Top-50 / Top-20 的相似度检索表现

• LLM排序机制 使 Top-50 Recall 接近原始 Top-100 水平，进一步通过 Prompt Engineering 提升 Top-20 结果质量

• 微调后的 QwQ-32B-bnb-4bit 模型 在排序阶段显著提升最终召回效果

整体系统在各 Top-K 水平的 Recall 上均超越传统语义检索基线，验证了嵌入微调、摘要优化与生成排序协同优化的有效性。

（RAG+知识图谱）智能客服系统　小组算法leader　2024.07-2024.10

项目概述：

构建面向长文本（PDF专利、操作手册、行业文档等）的多通路检索增强生成式问答系统（RAG），集成语义、关键词、图谱三类检索路径，结合Prompt模板化与记忆管理，提升问答准确性与系统可扩展性。

核心工作与技术实现：

1.基于 PyMuPDF 提取 PDF 文本内容，采用段落 + Tokens + Overlap 策略进行分块处理，提升语义完整性。

2.多通路检索索引构建,使用中文嵌入模型（如 gte-Qwen2 / bge）生成向量并写入 Milvus 向量数据库;提取关键词构建 BM25（rank_bm25）关键词索引库，支持关键词精确命中;利用 NER + 关系抽取构建结构化三元组，构建 Neo4j 图谱数据库，增强结构化问答能力。

3.混合召回与去重融合,针对用户问题并行触发语义检索、关键词检索、知识图谱关系检索;采用 MinHash + LSH 技术对多路召回结果去重并聚合，提高片段多样性与覆盖率。

4.精排与段落筛选,使用 CrossEncoder(bge-reranker-large)模型 对召回片段进行精细打分排序，提升最终 Top-K 精度表现;基于召回内容构建最终 Prompt 输入片段，提高 LLM 生成质量。

5. Prompt模板设计与生成优化,针对不同任务场景设计 Prompt 模板（结构对比、技术点抽象等），提升模型可控性。

6.大模型推理,使用 vLLM 部署 LLM 后端（如 DeepSeek、Qwen2）支持高性能生成。

7.LangChain 长短期记忆管理,接入 ConversationBufferMemory 与 VectorStoreMemory，实现用户多轮问答历史追踪与知识上下文增强。

8.前后端部署与原型集成,后端使用 FastAPI 实现服务部署，前端基于 Gradio 构建交互原型，支持问答展示、知识引用与召回路径可视化。

基于时间序列模型的浆果产量预测　Data scientist　2021.09-2024.05

客户有多年浆果产量数据，基于浆果多年历史数据以及每周的产量，去预测未来52周的浆果产量，客户从而预估合同的订单范围，最大限度的减少损失。

数据集：十年多的时间序列结构化数据。

1.使用 bigstore，bigquery 进行数据的存储，使用 sql对 bigquery 的数据进行分析提取。

2.使用 pandas，numpy 等数据科学工具在 cloab 上对数据清洗更新，生成模型所需的时间序列数据。

3.使用 sklearn的 lm,huber,rf,svr,tree,sgd 等多个机器学习模型，进行特征工程分析，模型预测。

4.使用 pyhtorch及 tensorflow 框架的时间序列深度学习模型 LSTM,TimeNet,Transformer 等多个模型进行产量预测。

5.对多个模型的输出结果进行评估以及动态及静态加权，并提交结果给客户。

6.使用 matplotlib，seaborn，plotly 等工具对模型预测的结果与人类预测结果进行评估比较 WMAPE。

7.使用 vizier,xmanager 对模型的多个参数进行 fine-tuning ，选出最好的模型以及参数，放到实际生产，减少预测的误差。

8.对不同种类，区域，品种的浆果数据做更详细的分析，对数据进行 normalization 等处理。

9.研究更多的时间序列模型如：Autoformer,Dlinear,LightTS ;以及 google 内部工具 vertex AI 等来预测浆果产量。

10.研究天气等数据与浆果产量的相关性以及进行 WMAPE 评估。

11.研究大语言模型对时间序列的预测。

12.把目前的代码向 github 进行迁移升级，使用到 GitHub 仓库，以及 gcolud 下的 cloud run，vertex AI等。

研究浆果Brix变化的驱动因素 Data scientist 2022.03-2022.12

内容:

更高的 Brix 影响浆果的溢价,本项目了解影响 Brix 的因素,比如天气,环境,土地管理等,从而向种植者提供针对性的建议来提高浆果的 Brix.

数据集:5年左右的 brix 及浆果产量数据.

1.使用 bigstore，bigquery 进行数据的存储，使用 sql对 bigquery 的数据进行分析提取。

2.使用 pandas，matplotlib 等数据科学工具在 cloab 上对数据进行可视化及数据质量分析。

3.对浆果数据进行生长季节的分解以及理解,对数据进行比较。

4.对天气和种植区域的影响进行时间序列分析,特征工程,使用 sklearn的 lm,xgb,tree,rf 等进行建模。

5.使用统计学 ARIMA 等模型以及 Vertex AI 模型进行建模。

6.获取天气,如温度,湿度,光照,紫外线等与 Brix 的相关性进行因果分析。

业绩:

get所有因素的相关性

金融领域多模态分类系统 NLP算法工程师 2021.04-2021.08

该项目是基于京北方旗下金融事业部积累大量的文本和图像数据,包括更大规模的多模态数据集,项目利用多模态技术融合多数据源,利用文本和图像提取技术完成了金融数据分类业务问题,为公司解决了多模态技术痛点. 有数据部门提供金融领域相关图片和对应的文本数据

项目职责:

数据集:2 万张图片,以及对应的 train.josnl,test.jsonl.

1、以视频关键帧/图像和附带的文本信息为输入，输出: 0 非金融领域,1,金融相关领域,进行二分类

2、对原始数据进行数据分析,文本:统计文本长度分布,标签的分布,图像:宽高分布

3、使用 faster-RCNN 网络进行图像特征抽取 生成 image.npy 文件方便后续多模态模型使用

4、构建 1.Resnet 和 RNN/GRU 结合的多模态模型,2.VisualBERT 单流模型,3.VilBERT 双流模型

5、多模型分布式训练和预测, 三种不同模型 baseline 分别在 55%,60%,65% 左右

6、使用知识蒸馏进行优化，将 ViLBERT作为 teacher，Resnet+RNN 网络作为 student

7、模型训练服务器:CPU: 64C，128G 内存，1T硬盘 > * GPU: 4*Tesla T4,

模型部署服务器:CPU: 64C，128G 内存，1T硬盘，10M 带宽

金融APP借贷平台文本摘要 NLP算法工程师 2020.11-2021.04

金融借贷 app 后台有大量客户和客服的对话,以及客户的建议,对此数据进行收集并提取主要意图,从而进行相关业务调整,更加了解客户与市场的需求

文本摘要项目分为 抽取式和生成式,

本项目抽取式采用 TextRank模型 抽取摘要

生成式分别采用 seq2seq 和 PGN 模型生成摘要,并使用 gensim 训练小规模的行业词向量

项目职责:

1.数据由客户提供 10万条 train.csv,2万条 test.csv,数据预处理:提取特定的文本以及对数据进行清洗,将有标签的数据 train:dev 分为 8:2

2.采用 TextRank 模型进行摘要抽取 ,模型抽取 baeline V1.0

3.采用 seq2seq,模型生成 baseline V2.0,PGN 模型,生成 baseline V3.0
PGN+coverage 模型进行优化 生成 baseline V4.0
使用 Beam-search Decode 进行优化 生成 baseline V5.0
使用 ROUGE 算法对模型进行评估,并最终选取 baseline V5.0
ROUGE-N1, precision,recall, f1-score:22,36,22
ROUGE-N2,precision,recall, f1-score:7,12,7
ROUGE-L,precision,recall, f1-score:35,31,29
环境:Pythorch1.6.0 ,jieba 工具 gensim 工具

保险领域智能问答(NER) NLP算法工程师 2020.04-2020.11

在保险领域智能问答应用研究中，用户提问时大量使用缩写、简写的保险名称，降低了问题语义理解的准确率，对保险数据进行抽样，通过 BiLSTM+CRF 模型， 从保险数据中抽取疾病、险类险别，保险产品，药品实体，自动抽取这些信息能更加高效 ，后续对数据和模型进行优化,在识别率和召回率得到了大大的提升。
项目职责:
命名实体审核: 处理结构化数据
1、数据由客户提十万条 ,正样本 train.csv 2 万条 作为正样本
2、构建 RNN 模型进行命名实体审核,使用 bert-chinese 预训练模型来获得中文汉字的向量表示.
3、模型训练预测,在验证集上 baseline,acc=70%,后续对数据进行增强,负样本增强逐步优化 86%.
命名实体识别: 处理非结构化数据
1、模型的标签定义,对四种实体进行设计类别标签 B.I.O 共需要设计 9 类标签，
2、训练集验证集共一万条，按 8:2分配，文本长度 ： 20
3、模型选用 Bilstm Crf模型,后续优化使用 BERT CRF ,以及 BERT Bilstm Crf
4、进行训练模型，保存模型 ,最终 BERT Bilstm Crf 在表现上最好 ,实体识别 acc,recall,F1-score 均达到 85%以上
硬件设备：16C，32G 内存的 CPU，GTX1080Ti 11G
neo4j 建立图数据库:
1、使用 Cypher 语言将命名实体写入图数据库,写入的数据供在线部分进行查询，
2、在线上部分根据用户输入内容来匹配对应的实体
句子主题相关部分:
判断用户的最近两次回复是否围绕同一主题, 来决定问答机器人是否也根据自己上一次的回复来讨论相关内容. 本质上来说是一个二分类任务
1、使用 bert-chinese 预训练模型，为了符合垂直领域在后续的全连接层进行 fine-tune
2、文本的平均长度为 15，训练集正负样本共 6 万条数据，验证集 1 万条，
3、模型训练，预测 baseline 70% Bert 后续优化可以达到 85%，textcnn 为 82%.
4、BERT，QPS 为 80ms,后续使用 textcnn 网络,textcnn，QPS 为 10ms，最终使用 textcnn 上线.

金融信息舆情分析 NLP算法工程师 2019.05-2020.03

金融行业咨询信息十分丰富，难以靠人工阅读分析大量的相关的资讯，利用 NLP 处理技术对每一条舆情做情感分析，从而辅助做出商机发现和风险预警,投资决策，对于相关金融信息可以进行实时在线监测，评估。
项目职责:
1、由客户提供数据五万条, 按照 8:1:1 分为 train,dev,test，文本长度 100-5000 不等，标签:

有益，中性，无益

2、数据预处理，提取并清洗相关数据，去除特除符号等

3、使用 BERT,textcnn,XL-Net 模型分别进行模型搭建

4、模型训练和预测，三种模型 acc分别为 70%，65%，73%左右

5、使用 flask 进行模型部署上线，属于 toB项目，对 QPS 没有太高要求，100ms。

金融领域机器翻译项目 机器翻译 2018.10-2019.04

公司针对国际金融领域的一些最新资讯，文章进行机器翻译

项目职责:

1、由数据部门提供数据一万篇金融领域英汉双语文章，作为训练语料 字数 1 千-3 万不等,

2、数据预处理 ,提取数据并清洗数据，使用 gensim 训练行业词向量

2、机器翻译模型选用 seq2seq GRU,transformer 等模型 ，

3、模型训练，翻译文本的生成

4、使用 rouge对两种模型进行评估，以及谷歌翻译同时评估，

5、transformer 在客观评价指标 rouge-1，rouge-2，rouge-l 在精确率，召回率，F1-score 都略高

在主观评价指标：1.流畅度 2.相关性 3. 助盲性 由人工打分，也相对较好

6、在公司内网使用 flask 进行部署上线

基于投诉风险预测情感分类项目 NLP算法工程师 2018.03-2018.09

金融证券银行理财投行等每天会有很多客户的投诉，因为数据量太大，难免有没有及时处理的情况， 造成客户的越级投诉，影响公司信誉，所以为了准确预测投诉客户越级风险，投诉处理管控机制优化，客服线条投诉处理资源利用率提升，降低投诉升级概率，采用 RNN 神经网络构建用户越级投诉 风险情感分类模型,从而对相关业务进行整改并控制投诉量

项目职责:

1、由客户提供数据集十万条, 制定三大类越级投诉风险判定标准（一般不满，比较不满，非常不满），针对现有样本进行风险标注

2、数据预处理，数据清洗，处理脏数据,按照(8:1:1)划分为 train,dev, test,

3、模型与框架 Pytorch、RNN/GRU/LSTM、textcnn;

4、模型训练及预测 在测试集上 baseline,acc =75%,后续优化几个模型版本 acc = 84%

数据中台管理系统 python开发 2016.09-2017.12

该项目使用Django框架实现数据中台数据管理系统。应用于公司数据出入管理，数据统计，数据抓取，数据查询等。对公司客户的数据进行安全管理，分析，安全，规范，起到一个系统化中间作用。

1. 用户注册登录，邮箱验证，权限设计等.

2. 对业务方日常入库数据进行统计，可视化展示.

3. 提供简历数据查询，简历数据重推，简历解析可视化接口.

4. 提供职位数据统计，职位解析，职位抓取接口。

5. 开发简历、职位数据中间件，检

员工任务进度管理系统 python开发 2016.03-2016.09

项目描述:

对公司员工工作进度进行统一管理，可根据不同部门、不同职位实现不同权限，只有个人可更改

个人的工作进度等管 理层人员可查看所管辖范围内的所有进度信息；
基于 django admin 源码实现通过数据库中每个注册后的数据表生成这个数据表的增删改查等一系列的 url 路径以及所要展示的字段列、是否为可添加数据、是否为可编辑数据的自定义组件以及根据不同用户角色生成不同 url 的权限组件跟可自定义显示条数的分页组件。

学一学社吧 python开发 2015.08-2016.03

项目描述：
学一学社吧是一个基于搜索的互动式知识问答分享平台，可以搜索自己未知的问题，也可以解答、回复和点赞他人问题的一款软件。
技术栈：flask、mysql、redis
1. 负责网站用户中心登录与注册；
2. 问吧首页的分类与展示；
3. 用户基本信息修改,发表个人提问,个人关注；

魔术鱼在线美术 python开发 2014.06-2015.06

项目描述：
魔术鱼主要为在线美术类学习内容及线上运营服务为 B 端机构服务，包含在线小班课教学平台、美术课，以及机构管理等功能，
1．负责整项目机构部分的开发；
2．对机构部分进行数据库表设及机构后台的开发；
3. 学校管理系统，学员系统。

教育经历

北亚利桑那大学 硕士 Computer Science 2024-2025

山西农业大学 本科 经济管理(第二学位) 2008-2012

山西农业大学 本科 食品质量与安全 2008-2012

学生会主席