余胜辉

18656417980 | truetoysh@gmail.com
https://blog.csdn.net/TrueYSH

专业技能

大语言模型：熟练 DeepSeek、Qwen、ChatGLM、LLAMA 训练与部署，擅长 Pre-train、SFT、RLHF 优化模型性能

高效部署：熟练 Sglang、vLLM 等框架加速推理，熟悉 LLaMA-Factory、Unsloth 微调提效

智能检索：熟练 RAG 提升问答精度，熟悉 GraphRAG 增强复杂场景应用

AI 产品化：基于 Dify 开发高性能 AI 应用与自动化 Workflow

多模态创新：熟悉 GLM-4V、Qwen2.5-VL，推动多模态解决方案

NLP 实战：擅长知识图谱与 NER，优化文本处理效率

深度学习：熟练 Transformer、BERT、GPT，驱动智能生成任务

机器学习：掌握 XGBoost、RandomForest，解决多样化业务问题

开发部署：熟练 Linux系统、Docker部署，保障项目快速上线

数据管理：熟练 MySQL、Redis、Neo4j，熟悉 Milvus 向量检索优化

敏捷开发：熟练 FastAPI、Flask，擅长 Prompt 工程提升交互体验

工具调用：熟悉 MCP工具调用，优化模型Function Call能力 自动化完成任务

工作经历

北京维思陆科技有限公司 2025年04月 - 至今

大模型算法工程师 研发部

主导大模型产品线从零到一的研发

参与产品竞品和数字人调研及开发

爬取RAG所需数据并进行数据清洗

主导大模型推理基座及智能体框架选型

实现生产环境在纯内网环境下迁移

北京元客方舟科技有限公司 2023年09月 - 2025年04月

NLP算法工程师 研发部

负责NLP算法的研发与优化,包括文本分类、NER、语义匹配和生成任务。

参与企业知识图谱的构建与维护,优化实体识别和关系抽取流程,提升知识查询效率。

主导RAG(检索增强生成)系统的设计与实现,显著提升智能问答系统的准确性和用户体验。

参与多模态数字人项目,负责语音、文本和动画模块的整合与优化。

北京争上游科技有限公司 2021年07月 - 2023年09月

NLP算法工程师 技术部

设计并实现文本数据预处理Pipeline,包括清洗、标注和特征工程。

开发并优化NER模型和知识图谱模块,提升系统对复杂查询的智能化响应。

通过分布式训练和超参数调优优化模型训练流程,缩短训练周期30%。

教育经历

珠海科技学院 2017年09月 - 2021年06月

计算机科学与技术 本科 珠海

项目经历

XX智慧就医智能体 2025年04月 - 至今

背景:

主导公司产品的更新，将大模型技术融入传统智慧就医。带领原有技术团队速度熟悉及应用大模型技术。

技术细节：

基于xinference部署BGE-m3、BGE-reranker、CosyVoice-300M、fish-speech-1.5等模型

清洗医疗数据构建RAG数据库、制定召回重排序规则。

构建工作流 实现就医咨询及分导诊 构建智慧路由优化意图识别和实体查询

工作流接入TTS及ASR实现语音输入输出

配合医院网安完成测试环境纯内网迁移并成功投入生产

成果：

项目从零研发到落地 与甲方紧密沟通 主动推进工作进度 加快研发周期

工作流创新 突破原有问题分类器 实现意图识别与实体提取双突破速度提升%70

良医问药智能问答数字人系统 2024年06月 - 2025年03月

背景:

为大健康咨询开发智能数字人,支持语音交互和知识查询,服务于健康管理平台。

使用senseVoice实现语音转文本,优化ASR模型在医疗术语上的识别率。

基于Dify框架构建workflow实现智能体,接入数字人实现智能应答。

基于Neo4j构建医药知识图谱,包含疾病、症状、药材等实体及关系。

集成FishSpeech(文本转语音)和Sonic(数字人动画),实现多模态输出。

技术细节:

ASR优化: 针对专业词汇扩充训练语料,提升识别准确率至95%。

工作流: 使用Xinference推理框架部署,大模型实现意图识别,

知识图谱: 设计三元组结构(如“疾病-症状-药方”),支持复杂查询。

多模态同步: 通过时间戳对齐语音和动画,延迟控制在300ms以内。

成果:

系统响应延迟300ms,10次测试唤醒成功率100%。

支持200+并发访问,满足企业级需求。

RAG智能问答系统 2023年09月 - 2024年06月

背景:

为科研公司开发文献知识检索工具,提升科研人员获取信息的效率。

使用GPT-4进行Query增强,生成多样化查询表达。

基于FAISS和BGE-large模型实现向量化检索,支持语义匹配。

优化RAG流程,结合检索结果和生成模型输出高质量答案。

技术细节:

Query增强: 通过GPT-4改写用户输入,增加检索覆盖率。

向量检索: 使用BGE-large生成1024维嵌入,FAISS索引加速Top-K查询。

RAG优化: 引入上下文截断策略,减少生成模型幻觉问题。

成果:

检索召回率提升30%以上,Top3准确率达88%。

RAGAS评估准确性达0.91,用户体验显著提升。

大模型训练及自动化评价系统 2023年02月 - 2023年09月

背景:

研发大模型并设计自动化评价系统,提升语料质量和模型性能。

参与模型的预训练和SFT微调,优化医疗领域生成能力。

开发自动化评价系统,使用SimCSE+MLP评估生成文本的语义连贯性。

技术细节:

预训练: 清洗10TB医疗语料,使用分布式训练(8块A100 GPU)。

微调: 设计SFT任务,标注5000条高质量问答对。

评价系统: SimCSE提取句向量,MLP回归预测连贯性得分。

成果:

节省80%人力成本,连贯性评分从0.7提升至0.9。

模型在医疗问答任务上表现优于基线10%。

争上游大健康智能体 2022年06月 - 2023年02月

背景:

开发健康咨询智能体,支持在线诊断和健康建议。

使用BERT+BiLSTM实现NER,识别症状、疾病等实体。

构建Neo4j知识图谱,存储健康领域结构化知识。

基于SimCSE+Redis优化检索模块,提升响应速度。

技术细节:

NER: BERT提取特征,BiLSTM捕捉序列依赖,准确率达92%。

知识图谱: 设计“用户-症状-疾病-建议”关系模型。

检索优化: SimCSE生成语义向量,Redis缓存热门查询。

成果:

NER准确率92%,检索响应时间缩短至1.5秒。

系统支持日均5000次查询,稳定性达99.9%。

肺炎知识图谱 2022年02月 - 2022年06月

背景:

构建肺炎相关知识图谱,支持多维度特征关联和查询。

使用BERT+BiLSTM+CRF实现实体抽取,识别疾病、症状、药物等。

基于SimCSE+ES进行知识融合,解决实体歧义问题。

使用Neo4j存储和管理图谱数据。

技术细节:

实体抽取: CRF层提升边界识别精度,F1分数90%。

知识融合: SimCSE计算实体相似度,阈值设为0.85。

图谱存储: 设计多层关系(如“疾病-症状-治疗”),支持复杂查询。

成果:

实体识别准确率92%,知识融合精度95%。

图谱支持实时查询,响应时间小于1秒。

疾病筛查与诊断系统 2021年07月 - 2022年02月

背景:

开发智能诊断系统,识别早期肺癌风险,支持医院决策。

数据清洗: 使用正则表达式(RE)处理非结构化医疗记录。

模型构建: 基于XGBoost训练分类模型,预测肺癌风险。

技术细节:

数据预处理: 提取关键特征(如年龄、吸烟史、CT结果)。

模型训练: 使用网格搜索优化超参数,特征重要性分析提升解释性。

成果:

模型AUC提升至0.85,假阳性率降低10%。

实现高危患者实时预警,辅助医生诊断。

团队协作:

在跨部门合作中表现出色,善于与产品、工程团队沟通需求和技术方案。

学习能力:

持续关注NLP领域最新进展(如大模型、RAG),快速掌握新技术并应用于实践。

个人总结

工作认真负责，参与过大模型从零到一的训练。熟练各类算法，对底层算法有研究。也在积极了解时兴SOTA，梦想是在大模型行业有所成就。