陈昊然 | 人工智能工程师 | 大模型与机器学习

中国,北京市,北京,北京市海淀区中关村大街 27 号,100080
+86 138 0013 8000 • chen.haoran.ai@example.com • https://www.haoran-ai.com
@haoran-ai • @haoran-chen-ai • 知乎 @haoran-ai

简介

- 人工智能工程师,专注大语言模型、RAG 与机器学习工程化,具备从数据、训练、评估到部署的全链路经验。
- 熟练使用 PyTorch、TensorFlow 与 Transformers,熟悉向量数据库、LangChain 和 MLOps,能够将算法原型转化为稳定服务。
- 有推荐系统、智能问答、文本理解等落地经验,注重模型效果、推理成本与线上稳定性。
- 具备跨团队协作与技术分享能力,持续关注生成式 AI、多模态与智能体方向。

教育背景

清华大学 **https://www.tsinghua.edu.cn**

硕士,计算机科学与技术,成绩:3.9/4.0 2016 年 9 月–2019 年 6 月

- 研究方向为自然语言处理与机器学习,参与实验室文本理解与知识抽取项目。
- 完成大规模文本分类与实体识别实验,发表 1 篇中文会议论文。
- 担任机器学习与数据结构课程助教,负责习题课与项目辅导。

课程:机器学习、深度学习、自然语言处理、计算机视觉、概率图模型、分布式系统、算法设计与分析

工作经历

云智科技有限公司 **https://www.yunzhi-ai.com**

高级人工智能工程师 2022 年 7 月至今

- 负责企业级大模型应用平台的核心算法与工程架构,支持知识库问答、文档摘要和智能客服等场景。
- 设计 RAG 检索增强生成链路,优化文档切分、召回、重排与提示词策略,问答准确率提升 23%。
- 主导 7B/13B 开源模型的领域微调与 LoRA 实验,建立离线评测与人工反馈闭环,降低幻觉率。
- 推动模型服务容器化与灰度发布,结合缓存、批处理与量化将推理成本降低约 35%。
- 与产品、数据和后端团队协作,制定技术方案与迭代计划,指导 3 名初级算法工程师。

关键字:大语言模型、RAG、LoRA、PyTorch、MLOps、模型量化

星图数据科技有限公司 **https://www.xingtu-data.com**

人工智能工程师 2019 年 7 月–2022 年 6 月

- 参与个性化推荐系统研发,负责特征工程、召回与排序模型迭代,提升点击率与转化率。
- 使用 TensorFlow 构建深度学习排序模型,设计离线训练与在线推理流程,支撑千万级用户请求。
- 搭建 A/B 实验与模型监控看板,分析关键指标并快速定位效果波动。
- 与数据平台团队合作优化 Spark 特征流水线,缩短模型训练数据准备时间。

关键字:推荐系统、TensorFlow、特征工程、Spark、A/B 实验

语言

中文:母语或双语水平 **关键字**:普通话

英语:完全专业水平 **关键字**:CET-6、技术文档阅读、英文技术分享

专业技能

机器学习与深度学习:专家 **关键字**:PyTorch、TensorFlow、Transformers、模型微调、强化学习

大模型应用:高级 **关键字**:LLM、RAG、LangChain、向量数据库、Prompt Engineering

工程与部署：高级　　**关键字**：Python、Docker、Kubernetes、FastAPI、MLOps

数据处理：高级　　**关键字**：SQL、Spark、Pandas、数据清洗

荣誉

优秀员工奖

云智科技有限公司　　2023 年 12 月

- 因在大模型应用平台项目中的突出贡献获奖。
- 主导 RAG 链路优化与推理成本治理，推动平台在多个业务线落地。

国家奖学金

清华大学　　2017 年 10 月

- 研究生阶段学业成绩与科研表现优异。
- 参与自然语言处理方向研究，完成多项实验与论文撰写工作。

证书

AWS Certified Machine Learning - Specialty　**https://aws.amazon.com/certification/certified-machine-learning-specialty/**

Amazon Web Services　　2021 年 8 月

TensorFlow Developer Certificate　**https://www.tensorflow.org/certificate**

Google　　2020 年 3 月

出版刊物

中国计算机学会自然语言处理专委会会议　**https://doi.org/10.12345/nlp.2021.001**

基于预训练语言模型的中文文本分类方法研究　　2021 年 10 月

- 研究预训练语言模型在中文文本分类任务中的迁移能力与微调策略。
- 通过对比实验分析不同预训练模型、数据增强与提示学习方法的效果。
- 在多个中文文本分类数据集上验证方法的有效性与稳定性。

项目

企业知识库智能问答系统　**https://github.com/haoran-ai/enterprise-rag**

基于 RAG 的企业知识库问答平台，支持文档解析、检索与答案生成。　　2023 年 1 月–2023 年 12 月

- 设计文档解析、向量化、召回、重排与生成一体化链路。
- 引入多路召回与答案引用，提升复杂问题的可追溯性与准确率。
- 建立离线评测集与线上反馈机制，持续优化提示词与模型策略。

关键字：RAG、大模型、向量检索、FastAPI

智能客服对话分析平台　**https://github.com/haoran-ai/dialog-insight**

面向客服对话的意图识别、情绪分析与质检辅助平台。　　2021 年 3 月–2022 年 12 月

- 构建对话文本清洗、意图分类与情绪识别流水线。
- 使用预训练模型提升小样本场景下的分类效果。
- 为质检团队提供可视化分析看板与高风险对话提醒。

关键字：NLP、意图识别、情绪分析、可视化

兴趣爱好

技术社区：开源贡献、技术写作、学术分享

运动：长跑、羽毛球、徒步

志愿服务

中国计算机学会青年计算机科技论坛　**https://www.ccf.org.cn**

人工智能方向志愿者　　2020 年 1 月–2022 年 12 月

- 参与人工智能主题论坛与技术沙龙的筹备与组织工作。
- 协助邀请讲者、整理活动资料并维护社区交流群。
- 推动高校学生与产业界工程师之间的技术交流。