李冠群

19152094604 | liguanqun07@163.com | 北京
22岁 | 男 | 汉族 | 共青团员

教育经历

北京师范大学 985 211 双一流 2024年09月 - 2026年06月

应用统计 硕士 统计学院

保研至北京师范大学

主修课程：机器学习； 数据挖掘； 非结构化数据分析； 深度学习

首都经济贸易大学 2020年09月 - 2024年06月

数据科学与大数据技术 本科 统计学院

专业排名：**1/36** **学院排名**：**1/139** **GPA**：**4.25 / 5.00**

英语水平：CET-4（525） CET-6（472）

主修课程：数学分析（95）； 数据科学的概率基础（95）； 数据科学的统计基础（97）； 数据结构（98）； python数据分析（96）； 计算机网络技术与应用（95）； 超高维数据分析（99）； 数据采集与存储（94）

专业技能：熟悉transformer、PyTorch、git、Python、Linux、MySQL、Excel、了解JAVA；计算机三级（优秀）

实习经历

云知声智能科技股份有限公司 2024年06月 - 2024年08月

大模型优化实习生 AI Labs 北京

项目背景：目前采用的是基于BERT的规则化应答，存在机械，呆板等问题。因此要研发新一代的对话系统。

工作内容：对车载语音的nlg任务进行优化，主要针对天气和音乐领域应答的优化，并能够对其他领域的问题作出基本响应。利用GPT4生成训练数据，并在此基础上人为修正，使用k8s在集群上对模型进行全量SFT，DPO对齐，使用了Megatron以及Deepspeed框架对14b模型微调训练。

项目成果：微调后的14b模型达到了仅用prompt的70b模型效果，使其能够达到产品经理制定的金标准得分，并成功部署上线。

科研竞赛

基于大语言模型的零售商品多维度分类研究 ——品类、属性与消费群体(华北赛区一等奖) 2024年04月 - 2024年06月

项目目标：本项目旨在利用先进的人工智能技术，特别是大语言模型，构建高效、准确的商品自动分类模型，提升零售企业的商品管理质量和消费服务水平，实现商品信息与消费群体的精准匹配 。

需求调研：针对企业零售商品标注数据量大、标签丰富、人工标注成本高的问题，将商业需求转化为技术问题，重点关注品类、标签和消费群体三个维度 。

数据处理：通过多次调整提示（prompt）撰写，优化大语言模型的预测能力。

模型构建：基于蚂蚁商业联盟提供的百万条人工标注数据，提出了从预训练模型到大语言模型的递进式零售商品多维度分类算法，采用DoRA等显存优化技术在sft阶段进行微调训练，从而提升算法准确度和泛化能力 。

实验结论：本研究通过对预训练模型以及大语言模型针对具体的下游任务微调训练使得模型准确率得到了显著的提升。实现了零售商品的高效、精准分类，提升了零售企业的商品管理质量和消费服务水平，对零售行业的发展具有重要意义 。

商品多级品类自动标注及多维标签识别（北京市优秀毕业论文） 2024年01月 - 2024年04月

商品多级分类是将商品自动分类到所对应的多层次分类体系。本文通过对比试验验证了传统深度学习方法的效果。结合两种文本特征提取和四种传统深度学习模型（CNN、RNN、LSTM、GRU），结果显示结合 Word2Vec 的 LSTM 和 GRU 效果最佳。其次，本文研究了不同预训练模型在商品多级分类中的表现差异。对比了 BERT、ALBERT、RoBERTa、ERNIE 四种模型，发现 ERNIE和BERT具有最佳效果和泛化能力。

商品多标签分类是将商品分类并赋予多个能够描述商品的特征标签。本文探讨了如何利用生成式模型进行商品的多标签分类。研究了四种生成式模型，其中包括 GPT-1.0、Alpaca、Gemma 和 ChatGPT-3.5-turbo 模型，通过一轮提示对话任务使模型输出固定格式的分类结果，并对 Gemma 和 ChatGPT-3.5-turbo 模型进行了微调。最终，利用 BLEU 指标评估了模型效果，得出了微调后的 ChatGPT-3.5-turbo 是最佳模型的结论。

荣誉奖项

计算机软件著作权登记证书	2024
美国大学生数学建模竞赛M奖	2022
全国大学生大数据分析技术技能大赛省级一等奖	2022
二等科研创新奖学金（学院2%）	2021
连续三次校级三好学生（学院5%）	2020~2023
连续三次学习一等奖学金（专业5%）	2020~2023
北京市优秀毕业生	2024
应用统计研究生案例大赛华北赛区一等奖	2024