

聂浩宇

男 | 23岁 | 中共党员 | 求职意向: 算法工程师 | 期望城市: 京深沪杭
联系方式: 18354270795 | nhy18354270795@163.com



教育背景

卡内基梅隆大学 (Carnegie Mellon University)

人工智能工程理学硕士, GPA: 3.87/4.00

匹兹堡, 美国

2025 年 8 月 - 2026 年 12 月

西南交通大学

计算机科学与技术学士, GPA: 3.86/4.00, 均分91

成都, 中国

2021 年 9 月 - 2025 年 6 月

技术技能

语言: 中文 (母语)、英语 (IELTS: 7.5 / CET-6: 588 / GRE: 325)

编程语言: Java、Python、Golang、C、C++、JavaScript、Kotlin

深度学习框架: PyTorch、TensorFlow、Hugging Face Transformers、DeepSpeed、vLLM、scikit-learn

大模型技术: Multimodal(Vision/Audio)、SFT/PEFT、PPO/DPO/GRPO、RAG、Harness Engineering

大数据与云: Spark(PySpark)、PgVector/PostgreSQL、Elasticsearch、Docker、Kubernetes、AWS/GCP

实习经历

顺丰科技

大模型算法工程师

2026 年 6 月 - 2026 年 8 月

- 参与研发基于 LLM Agent 的算法决策释疑工具链, 构建“AST分析-脚本自生成-并发测试-Git同步”的 4 阶 Skill 架构; 成功落地于 PySpark+Gurobi 复杂模型, 全自动提取 29 条约束规则与隐藏逻辑; 将异常归因效率提升 10-20 倍, 测试通过率达 100%
- 负责基于 AgentScope 搭建的AI释疑 Harness 框架长短期记忆结构、上下文压缩、事实提取与检索的性能测试和逻辑优化

海程邦达(A股上市)

大数据工程师

2024 年 6 月 - 2024 年 8 月

- 基于OpenAI API与RAG技术研发智能客服系统, 通过多Agent协作的Text2SQL框架将任务解构为Schema检索、SQL 生成及自修复流程, 实现复杂业务数据的结构化问答, 深度优化了空运产品的咨询流程。
- 主导 7.4 万企业数据建模, 利用 IV/WOE 及分箱单调性校验完成特征工程; 基于逻辑回归构建预测模型, 通过 KS、AUC 指标评估并识别“双峰”分布特征, 为高潜力客户识别提供量化依据。

学术成果

Haoyu Nie*, Shengyuan Lin*, Zhiling Li*. Quantitative Diagnosis of Action Dynamics, Instruction Following, and Exploration Constraints in NitroGen (*Equal contribution.)

Submitted to TMLR (Transactions on Machine Learning Research), under review.

项目经历

基于提示词引导的增强型长视频时序定位与理解系统 项目组长

2026 年 3 月 - 2026 年 5 月

- 针对长视频理解中的“采样信息丢失”与“时序幻觉”瓶颈, 基于 LLaVA-NeXT 构建一套集成了输入侧自适应重采样与输出侧结构化锚定的全栈框架, 实现对 3 分钟及以上长视频的精准时序定位
- 双阶段重采样管线: 对长视频进行粗采样语义扫描, 通过SigLIP计算各视频帧与Query相关性, 在Top-K窗口分别进行高帧率、高分辨率精采样, 旨在将Token预算精准分配至关键事件帧, OracleIoU相较均匀采样提升28.66%
- 将连续时间轴离散化为 1000 个可学习Bin Token, 并引入 <start> / <end> 语义锚点 token, 将时序定位问题转化为离散 token预测任务。设计复合损失函数 (时序token加权交叉熵 + 距离敏感定位损失), 结合 LoRA 微调, 使模型时序解析率从 61.4% 提升至接近 100%, mIoU@70提升约 4 倍 (0.165 → 0.675)
- 引入 DIVID 双塔架构, 实现高分辨率空间特征与高帧率时间特征的异步融合。为提升模型感知容量并维持高效模型推理设计混合专家模型 (MoE) 投影器以及结合视觉 Token的动态路由方案

可解释性动作质量评估深度学习系统设计与优化 项目组长

2025 年 9 月 - 2025 年 12 月

- 基于 I3D 视频特征卷积与Transformer交叉注意力机制, 构建了一套程序感知 (Procedure-aware) 的深度学习框架, 实现对复杂跳水动作 (起跳、滞空、入水) 的精细化时空对齐与建模
- 集成 Detectron2 (目标检测) 与 HRNet (姿态估计) 搭建自动化评估管线, 利用Savitzky-Golay滤波器对16个关键点轨迹进行时域降噪, 精准提取翻腾紧凑度、入水垂直度、水花大小等 5 项核心运动学指标
- 设计了包含得分回归、时序分割及姿态评分的多任务回归头, 引入基于物理规则的运动学指标作为辅助监督信号, 通过结构化先验引导模型提取具有物理意义的解耦特征, 显著增强模型预测的逻辑性与透明度
- 在 FineDiving 数据集上验证, 模型 Spearman 秩相关系数提升至 0.9302, 相对L2距离降低 6.46%, 并在高精度时序分割指标 tIoU@0.75 上提升 5.68%
- 实验中模型在 30 个 Epoch 时即可达到基准模型 200 个 Epoch 的性能水平, 大幅缩短了实验迭代周期

基于大语言模型的机器翻译平台开发 个人独立完成

2025 年 1 月 - 2025 年 6 月

- 通过LoRA微调与RLHF技术, 构建了支持通用及特定领域 (法律、科学、新闻) 的中英双语翻译系统
- 筛选并预处理中英语料集CCMT2019, 使用Qwen API合成特定领域优劣翻译奖励模型训练集
- 采用LoRA微调通用模型 (Qwen2.5-7B-Instruct), 结合SFT与PPO算法使用4张RTX 4090显卡优化领域模型, 解决专业术语与语境适配问题。最终实现在法律 (BLEU提升 136%、COMET提升60%)、科学 (BLEU提升18%)、新闻 (COMET提升 24%) 三个特定领域的指标显著提高