黄朝晖

邮箱: andrew.z.huang@outlook.com ｜ 手机: +86 188-1171-1781
GitHub: github.com/J-L-Andrew ｜ LinkedIn: linkedin.com/in/andrewhuangpku

教育背景

- **北京大学,** 力学与工程科学系, 工程力学 | 博士研究生 2020.09 - 2025.06(预计)
  - **2023.09 - 2024.10 耶鲁大学联合培养**、国家留学基金委奖学金、北京大学校长奖学金、北京大学李惠荣奖学金、北京大学优秀科研奖（3 次）
  - 相关课程：机器学习、深度学习技术与应用、计算智能、并行程序设计
- **佐治亚理工学院,** 计算机学院, 机器学习 | 硕士研究生 2024.03 - 2025.06(预计)
  - 主要课程：人工智能、计算机视觉、强化学习、GPU 硬件与软件技术；GPA 前 10%
- **密歇根大学安娜堡分校,** 信息学院, 应用数据科学 | 硕士研究生 2024.03 - 2025.06(预计)
  - 主要课程：自然语言处理、推荐与搜索算法、数据库架构与技术、云计算；GPA 前 10%
- **北京大学,** 力学与工程科学系, 理论与应用力学 | 理学学士 2016.09 - 2020.06
  - 相关课程：计算概论、数据结构与算法、程序设计与算法、概率与数理统计

专业技能

- 熟悉 Python、C/C++ 和 SQL 等，掌握 PyTorch、Tensorflow 和分布式训练框架（DeepSpeed、vLLM）
- 熟悉 GPT、LLaMa、GLM、Qwen 等大模型基座，掌握 LoRA、RLHF、DPO 等微调方法
- 熟悉 RAG、Agent 等大模型应用框架，具备知识库问答、多模态内容理解、内容生成等实际应用经验

实习与科研

百度在线网络技术（北京）有限公司 2024.11 - 至今

- **搜索内容技术部, 内容供给组** | 大模型算法实习生 (百度 AI 妙笔专项)
  - 负责搜索下多个垂域的多模态内容理解，通过 COT 和多次素材抓取生成优质内容并微调文心大模型，将在手机百度 APP 上线
  - AI 润色模型训练：收集搜索垂域数据并构建偏好数据集，利用 DPO/KTO 做偏好对齐
- **移动生态战略规划部, MEG 投资管理组** | AI 战略研究实习生
  - 负责调研全球 AIGC 前沿技术及国内外一级市场融资情况，分析不同 AI 赛道商业化趋势；调研搜索相关竞品与统计不同类别 AIGC 应用的流量
  - 参与撰写 AI 及 AIGC 创投报告

阿卜杜拉国王科技大学, PRADA Lab | 大模型研究员 2024.04 - 至今

- **多模态大模型知识编辑研究** (Enhancing CLIP for Improved Multimodal Retrieval)
  - 结合 MMVP benchmark、GPT-4 和 DeepFloyd IF 构建合成图文数据，对比 DINOv2 和 CLIP 嵌入表示并筛选 CLIP 盲对，用于 LLaVA-1.6-13B 模型进行推理
  - 分析 CLIP-ViT 模型中 MLP 层神经元的二阶效应，通过 PCA 近似其主要响应方向并利用 OMP 稀疏分解技术映射为文本关键词，提升模型在处理 badcase 时的解释能力；构造 CLIP 误分类的语义对抗样本对 LLaVA 进行全参数微调，LLaVA benchmark 分数提升至 89%
  - 识别 CLIP 激活神经元并提取盲对概念，采用 PiSSA 技术微调其权重，增强嵌入表示对盲对的区分能力，测试集上准确度提升至 85%，并在图文检索任务中实现 15% 的 Recall 提升
- **跨文化大语言模型表征优化与对齐** (Cross-Cultural LLM based on Representasion Engineering)
  - 基于 VSM 数据集和 Hofstede 文化维度构造 prompt，分别使用 PCA 和对比提示提取各文化维度的表征向量，将其嵌入到 LLaMa-3-8b 模型中，在 CultureLLM benchmark 上均分达到 0.75
  - 利用 LoRA 框架对齐模型表征与目标概念表征，计算加权 adapter 并微调模型，对齐效果较 zero-shot 提升 17%，CAT 分数达到 71%，在小参数模型下表现接近 GPT-4
  - 通过 Global Opinions 等数据集训练不同国家的奖励模型，在表征空间训练值函数，推理阶段添加控制信号以最大化奖励，对齐分数优于 PPO 与 DPO，基于 GPT-4 的 win rate 提升至 71.3%

- **社交应用多模态内容安全风控系统** 2024.07 - 2024.09
  - 使用 LLaVA-NeXT、InternVL2 和 GLM4-V-9B 等模型集成打标，结合二次人工标注构建图文数据集，过滤文本和图片，应用 MinHash 与 FAISS 加速的聚类算法实现去重，并结合 CLIP 相似度清洗数据
  - 借助 DeepSpeed 框架和 LoRA 微调优化 InternVL2-2B 模型，选用违规图片及 CLIP+ 图像掩码对模型进行对抗训练，高风险内容识别准确率提升至 98.4%；采用 DPO 对模型进行 RLAIF-V 训练，降低模型幻觉率，从而提升了模型的安全性
  - 采用 AWQ 4bit 量化技术轻量化模型，经 lmdeploy 部署后识别精度高达 96.1%，并实现 5 倍推理加速

竞赛与论文

- **Kaggle**. LMSYS - Chatbot Arena Human Preference Predictions (2024). **49th out of 1849 (Silver)**
- **Kaggle**. Image Matching Challenge 2024 - Hexathlon. **22th out of 929 (Silver)**
- **Huang, Z.**, Zhou, X., Jin, W., & Li, S. (2024). Shape and space filling: a review on particle packing. *Science Bulletin*, Under Review. (IF: 18.8, Q1)
- **Huang, Z.**, Deng, W., Zhang, S., & Li, S. (2023). Optimal shapes of disk assembly in saturated random packings. *Soft Matter*, 19(18), 3325-3336. (IF: 2.9, Q2)
- **Huang, Z.**, Deng, W., Yuan, Y., Liu, L., Wang, Y., & Li, S. (2022). Determining the equivalent packing diameter of two-dimensional shapes. *Powder Technology*, 396, 565-577. (IF: 4.5, Q2)
- Zhou, X., **Huang, Z.**, & Li, S. (2024). The wall effect on packing densities of monodisperse and polydisperse spheres, *Powder Technology*, Under Review. (IF: 4.5, Q2)

学生工作

- **班长,** 北京大学工学院 20 博力行 2 班 2021.09 - 至今
- **会长,** 北京大学学生台湾研究会 2022.03 - 2023.06
- **主编,** 北京大学党委宣传部学生记者团 2018.03 - 2023.06
- **副部长,** 北京大学工学院团委组织部 2018.09 - 2019.09
- **组长,** 北京大学广播台新闻组 2017.09 - 2018.09