

陈嘉伟

手机: 15867066921 · 主页: <https://chenjiawei30.github.io/> · 邮箱: jiawei.ucas@gmail.com

工作经历

快手, 基础大模型与应用部, 算法实习生 2025.08 - 至今

- 研究大语言模型在用户角色扮演场景下的能力与评估方法, 探索具备行为输出的模型及其评估体系。

影刀 RPA, AI 部门, 研发实习生 2024.05 - 2024.08

- 使用 LabelStudio 标注电商网页多模态数据, 构建高质量训练集, 提升目标检测模型识别率;
- 基于影刀 RPA 平台设计自然语言驱动的自动化流程, 使用 JavaScript 开发网页动态元素捕捉算法, 结合 HTML XPath 路径优化与异步加载建模, 提升指令解析与执行的鲁棒性。

中国科学院软件研究所, LLM 实习生 2023.10 - 2024.09

- 构建面向中国法律的大模型, 设计法律三段论与多轮对话策略, 收集并整合微调数据集 (~30 万条);
- 选用 Baichuan2-7B 作为基座模型, 使用 FastChat 框架, 在 8 卡 H100 上进行全参数微调, 模型在 DISC-Eval 法律评测任务上达国内 SOTA; 额外引入 RAG 模块, 撰写相关成果获评优秀毕业论文。

教育背景

中国科学院软件研究所, 中文信息处理实验室, 软件工程硕士 (保研) 2024.09 - 至今

导师: 韩先培研究员、孙乐研究员

杭州师范大学, 计算机科学与技术, 工学学士 2020.09 - 2024.06

GPA 4.3, 排名第 1 / 342; 高等数学 100, 数据结构 99, 数据库 100, 计算机组成原理 100, 概率论 96, 算法设计 95, Python 程序设计 95, C++ 程序设计 95, Web 程序设计/服务端开发 95, 计算机网络 93

研究经历

研究方向: 多轮对话、指令生成、大模型推理

- 模型多轮对话中主题一致性退化研究 (2024 年 10 月 - 2025 年 6 月)
提出骨架引导的多轮对话生成框架, 显式建模人类真实意图与多轮对话生成方法, 增强对话跨轮主题的一致性和目标导向性。在多个基准上验证方法有效性, 显著提高了模型主题一致性和对齐能力。
- 大模型推理与部署优化: LLM 系统工程与服务架构实践 (2024 年 6 月 - 至今)
(1) 熟练使用 vLLM、LLaMA-Factory 等大模型框架, 具备多卡/多机模型部署和训练经验, 熟练使用 Docker 等工具; (2) 负责实验室 Qwen、DeepSeek 模型推理服务, 搭建统一 API 管理平台, 成功请求总次数超 500w; (3) 开发 ZServe: 一站式 OpenAI API 兼容推理服务框架, 支持动态路由、敏感词过滤、API Key 认证、流式输出与模型可见性管理, 大幅简化 LLM 服务部署与运维流程; (4) 维护 WebUI 对话服务, 服务对话次数超 100w; 针对 DeepSeek 高并发场景开发相关模块, 实现用户分级调度、资源隔离与请求排队机制, 解决 MindIE 阻塞与响应异常问题。

获奖情况

- 国家奖学金 2023 年
- 浙江省优秀毕业生 2024 年
- 校长奖章 (本科生最高奖学金, 全校 10 人) 2024 年
- 马云青春领袖奖十佳大学生 (本硕博最高荣誉, 全校 10 人) 2023 年

学术论文

- [EMNLP 2025, Oral Presentation] Jiawei Chen, Xinyan Guan, et al. *ConsistentChat: Building Skeleton-Guided Consistent Multi-Turn Dialogues for Large Language Models from Scratch*. [Preprint]
- [ICCV 2025] Qianhao Yuan, Qingyu Zhang, Yanjiang Liu, Jiawei Chen, Yaojie Lu, et al. *ShortV: Freezing Visual Tokens in Ineffective Layers of Multimodal Large Language Models*. [Preprint]
- [Submitted to ACL 2026] Yihan Chen, Jiawei Chen, Xuanang Chen, et al. *CoCoNUTS: Concentrating on Content while Neglecting Uninformative Textual Styles for AI-Generated Peer Review Detection*.