龙卓涵

loongnanshine@gmail.com | 15574383908
nanshine.github.io | Nanshine | nanshine.me

教育背景

硕士：复旦大学 · 大数据学院 · 统计学专业 · **2024–2027**
研究指导：魏忠钰副教授、博士生导师 · 复旦大学数据智能与社会计算实验室（Fudan DISC）
研究方向：AI Agent 评测与训练、大模型安全
本科：同济大学 · 软件学院 · 软件工程专业 · **2020–2024** **绩点/排名**：4.80/5.00 / 9/200

实习经历

阿里巴巴｜算法工程师实习生（AI Agent） 2026.06–至今

Alibaba Token Hub（ATH）事业群 · 医疗问诊 Agent 训练与优化

- **训练与评测：**面向医疗问诊场景，搭建覆盖问诊、药品推荐与健康 / 药品检索四类动作的 Agent 训练与评测流程，构建训练数据及规则与 LLM Judge 联合奖励体系，开展单步 RL 与 SFT 训练，并探索 Agentic RL。
- **成果：**训练后的 27B 模型在问诊任务评测中超过 Qwen 3.8 Max，并最终上线线上业务。

论文发表

Strong Reasoning Isn't Enough: Evaluating Evidence Elicitation in Interactive Diagnosis
Zhuohan Long, Zhijie Bao, Zhongyu Wei
ACL Findings 2026

How Jailbreak Defenses Work and Ensemble? A Mechanistic Investigation
Zhuohan Long, Siyuan Wang, Shujun Liu, Yuhang Lai
EMNLP Findings 2025

Benchmark Self-Evolving: A Multi-Agent Framework for Dynamic LLM Evaluation
Siyuan Wang*, **Zhuohan Long***, Zhihao Fan, Xuanjing Huang, Zhongyu Wei
COLING 2025

From LLMs to MLLMs: Exploring the Landscape of Multimodal Jailbreaking
Siyuan Wang*, **Zhuohan Long***, Zhihao Fan, Zhongyu Wei
EMNLP 2024

项目经历

交互式医疗问诊智能体评测与优化 2025–2026

Strong Reasoning Isn't Enough: Evaluating Evidence Elicitation in Interactive Diagnosis, 2026

- **方法设计：**面向信息不完整、需主动追问的交互式医疗问诊场景，构建包含患者模拟、检查请求、证据收集与诊断决策的多轮评测流程，提出 Information Coverage Rate (ICR) 衡量模型对必要证据的获取完整性，并构建 EviMed benchmark。
- **实验发现：**在 10 个大模型上系统评测后发现，强诊断推理能力并不必然带来强信息获取能力；进一步提出基于诊断验证反馈的 REFINE 策略，提升信息覆盖率与诊断成功率。

大模型越狱防御机制分析与组合策略研究 2024–2025

How Jailbreak Defenses Work and Ensemble? A Mechanistic Investigation, 2025
From LLMs to MLLMs: Exploring the Landscape of Multimodal Jailbreaking, 2024

- **机制拆解：**围绕 LLM 与 MLLM jailbreak 风险，从 safety-helpfulness trade-off 视角系统比较 28 类防御方法，归纳出 Safety Shift 与 Harmfulness Discrimination 两类核心机制。
- **方法结论：**进一步研究 defense ensemble，发现同机制组合更适合强化单一目标，跨机制组合更有助于兼顾安全性与有用性，为防御方法选择与组合提供机制化依据。

多智能体驱动的评测数据集自演化 2024–2025

Benchmark Self-Evolving: A Multi-Agent Framework for Dynamic LLM Evaluation, 2025

- **框架设计：**面向静态 benchmark 易过时、易受数据污染影响的问题，设计由 pre-filter、instance creator、instance verifier 与 option formulator 组成的多智能体流程，通过六类 reframing 操作从 scalable、robust、fine-grained 三个维度生成并验证新评测样本。
- **实验发现：**在四类任务上扩展 benchmark 并重新评测多种模型，发现动态评测能显著降低多数模型得分、拉开模型差距，并更有效暴露泛化、鲁棒性与子能力短板。

Mira 多模态美妆智能体 2025

欧莱雅首届“有意思青年—科技大 Fun 颂”美妆科技黑客松突出创意奖

- **产品构想：**围绕“AI 化妆陪伴”设计多模态美妆 Agent，探索面向实时指导、情感陪伴与互动反馈的“魔镜式”智能体产品形态。
- **功能实现：**基于通用多模态智能体架构实现视频输入、语音回复与工具调用，集成肤质分析、产品识别与检索、个人产品库管理及化妆教学模式等能力。