毛楹智

15738183237@163.com 157-3818-3237 Homepage

教育背景

中国科学院软件研究所，计科硕士在读；中文信息处理实验室　　2024.09 – 预计 2027.06

主要研究方向： *LLMs*、*Post-Training*、*Safety Alignment*、*Agent*　　**导师：**韩先培研究员、孙乐研究员

郑州大学，计算机科学与技术学士　　2020.09 – 2024.06

实习经历

蚂蚁集团 • 支付宝事业群-Jinni Studio 技术部 大模型应用算法实习生　　2026.05 – 至今

- **Judge Model 训练与强化学习优化：**面向 **AI 支付宝（Jinni）Service Agent** 服务决策场景，基于线上决策数据构造包含判定依据的 CoT 样本完成 Judge Model 冷启动，并**结合强化学习优化服务决策评估与错误归因能力**；模型判定一致率达到 90+%，错误归因与 GT 识别准确率达到 80+%，实现线上 Case 自动审核，人工排查量降低约 80%。
- **Agent 多轮能力评测与缺陷分析：**针对 Service Agent 多轮服务决策能力，构建覆盖顺承、意图切换及相似意图区分的 Benchmark 体系；基于 LLM 解析多轮交互关系，结合 Embedding 与 **HDBSCAN 聚类**实现线上数据自动归类，量化不同能力维度表现并定位模型短板，驱动模型优化迭代。
- **个性化能力 Benchmark 建设：**设计覆盖 **18 类场景**与 **5 个评价维度**的个性化 Benchmark 体系，涵盖长期记忆、会话事实理解、上下文利用与个性化响应能力；构建 1000 条专项评测样本，量化模型在多源信息融合与优先级决策场景下的个性化能力表现，推动 Service Agent 个性化能力提升 **23%**。

科研经历

LRMs 安全对齐：Self-Jailbreak 诊断与防御 *(第一作者, ACL 2026 Findings)*　　2025.04 – 2026.01

- 分析 LRMs 在多步推理中的安全失效问题，提出 *Self-Jailbreak* 现象：模型虽能在早期识别风险，但会在后续推理中自行推翻安全判断；基于该现象构建**三阶段推理分析框架**，定位安全失效发生的关键推理步骤，并基于轨迹级分析构造针对性的安全对齐数据。
- 提出 **Chain-of-Guardrail** 轨迹级训练框架，通过对导致不安全输出的关键推理步骤进行重写或回溯修正，在基本保持原有推理能力的前提下提升模型安全性；在 Qwen3-32B 上将 ASR 从 39.49 降至 **6.08**，推理平均分仅由 85.78 下降至 84.91，安全性与推理能力权衡优于现有强基线 SafeKey。

项目经历

PPTAgent *(核心贡献者)*　　2026.01 – 2026.03

- **训练数据构造与质量控制：**针对 PPT Agent 生成任务中复杂规划轨迹难以获取的问题，围绕整体风格、页面内容、页数规划等 **6 类排版约束**生成端到端训练样本；结合 **Vision-based Oversight** 与 **Critic-Follow** 构建双层轨迹质量控制机制，通过状态回溯过滤低质量轨迹。
- 基于 Qwen3.5-9B 在 51K 上下文长度下进行 SFT，针对超长输入场景通过并行训练与梯度检查点优化显存占用，解决训练过程中的 OOM 问题并保证稳定训练；在 PPTEval 上取得接近 GPT-5 表现。

基于强化学习的安全审核模型 *(项目负责人)*　　2025.11 – 至今

- **网络安全审核模型优化：**面向国内网络安全审核场景，针对复杂语境下风险识别问题，设计 CoT 样本并结合强化学习优化模型安全判别能力；基于评测集与线上数据分析模型失效模式，利用规则引擎**召回高价值难例并增量训练**，提升模型泛化能力，最终 Precision / Recall 达到 **86.35% / 91.88%**。
- **AIGC 内容安全审核模型：**面向生成式内容安全场景，构建 Harm 与业务双域训练数据，设计 Prompt/CoT 冷启动样本并结合生成式难例扩增提升长尾风险覆盖；基于 **GRPO** 设计非对称奖励机制优化安全决策边界，最终 Accuracy / Precision / Recall / F1 达 **95.16% / 93.51% / 93.50% / 92.06%**。

论文发表

- **[ACL 2026 Findings, First Author] Yingzhi Mao**, Chunkang Zhang, Junxiang Wang, et al. When models outthink their safety: Unveiling and mitigating self-jailbreak in large reasoning models.
- **[ACL 2025 Demo]** Xinyu Lu, Dong Xu, Chunkang Zhang, Xinyan Guan, Junxiang Wang, Qingyu Zhang, Pengbo Wang, **Yingzhi Mao**, et al. AUTOALIGN: Get Your LLM Aligned with Minimal Annotations.