杨晓晨

yyyxcleo123@gmail.com

150-6642-6729 yyyxcleo123

教育背景

清华大学（Top2，硕士） 计算机技术（研究方向：**多模态大模型与智能体**） **2026.09 ~ Present**

哈尔滨工业大学（C9，本科） 软件工程（专业第 1，**保送**至清华大学） **2022.09 ~ 2026.06**

- **学术成果**：英语六级 590 分；在 CCF-A/清华-A 类会议发表文章 4 篇，在投文章 1 篇。
- **荣誉奖项**：国家奖学金 (2023、2025)、哈工大一等奖学金 (2023、2024)、蓝桥杯算法大赛省级一等奖。

论文产出

- *Looking Back and Forth: Cross-Image Attention Preference Learning for Multi-Image Hallucination (ECCV26)*
- *Revisiting LLM Backdoor Attacks: A Stealthy Poisoning Framework via Harmless Inputs (EMNLP26-Main)*
- *CHILLGuard: Chinese LLM Guardrail with Data-Model-aware Preference Alignment (EMNLP26-Finding)*
- *Naming to Learn: Class Incremental Learning for Vision-Language Model with Unlabeled Data (ICLR26)*
- *ToolHallu: Mitigating Hallucination in Agentic Tool Calls via Reinforcement Learning (ICLR27 underreview)*

科研项目

基于偏好强化学习的多图视觉语言模型幻觉缓解 **ECCV 2026 (1st Author)** **2025.10 ~ 2026.03**

项目介绍：针对多图视觉语言模型因**跨图信息交互不足**与**跨图证据偏好缺失**导致的视觉幻觉问题，提出跨图注意力校准与偏好学习框架，增强模型的跨图实体关联与视觉证据利用能力。

1）**跨图注意力校准：**设计**选择性跨图注意力机制**，基于视觉 Token 的 **Embedding Energy** 筛选关键 Token，解除不同图像间的因果注意力约束，并采用**交替 Mask** 策略增强跨图信息交互与实体对齐能力。

2）**跨图偏好学习：**基于跨图交互程度构造正负偏好样本，利用 **DPO** 强化模型对跨图视觉证据的依赖，并结合 **NLL Loss** 提升正确回答的生成稳定性。

3）**实验效果：**在 **Qwen2.5-VL**、**InternVL2.5**、**GLM4.1VBase** 三类主流 MLLM 上验证方法有效性，相较对应基线最高分别提升 **4.49/3.58/2.73** 个百分点，同时保持单图任务与通用能力基本稳定。

ToolHallu: 工具调用幻觉缓解 **ICLR27 Underreview(1st Coauthor)** **2026.04 ~ 2026.09**

项目介绍：针对 LLM Agent 工具调用过程中存在**工具选择**、**参数生成及重复调用幻觉**，首次提出细粒度幻觉监督 **ToolHallu** 强化学习框架。

1）**细粒度幻觉奖励：**将工具调用幻觉细分为**工具名称**、**参数与重复调用**三类，并进一步建模参数缺失/冗余、类型错误及值错误，设计结构化 Rule-based Reward，为不同调用错误提供细粒度、可解释的训练信号。

2）**课程式奖励调度：**针对不同幻觉类型的学习难度差异，设计**余弦退火奖励调度策略**，训练前期重点优化工具选择，随后逐步提升参数级幻觉惩罚，实现从粗粒度工具选择到细粒度参数精度的渐进式优化。

3)**实验效果：**在多个模型与工具调用基准上均降低幻觉率，相较 SOTA 方法在 **API-Bank** 与 **StableToolBench** 上的幻觉指标平均相对下降 **8.00% / 20.25%**，同时持续提升任务成功率。

基于干净样本的后门越狱攻击 **EMNLP26-Main (1st Coauthor)** **2025.05 ~ 2025.09**

项目介绍：针对传统 LLM 后门攻击依赖**毒化样本**、易在模型微调前被安全检测器过滤的问题，提出仅利用**干净样本**构建隐蔽后门的方法，在保持模型正常能力的同时实现特定条件下的稳定行为切换。

1）**干净样本后门构造：**利用干净 QA 数据，通过**肯定性前缀与结构化内容**建立触发条件与目标行为的隐式关联，在不引入显式异常数据的情况下完成后门植入，提升攻击隐蔽性。

2）**通用触发器增强：**引入**梯度引导的通用触发器优化策略**，增强触发器在不同输入下的激活稳定性，提升后门触发的泛化能力。

3）**实验效果：**经安全过滤后，**Rule-based Judge 平均 ASR 达 98.13%**，多个模型达到 **100% ASR**，最高较无攻击基线提升 **96.67** 个百分点。

CHILLGuard: 中文大语言模型细粒度安全护栏 **EMNLP26-Finding (Coauthor)** **2025.11 ~ 2026.03**

项目介绍：针对现有 LLM 安全护栏在中文场景下存在**风险分类粒度不足**、**隐式/变体风险覆盖有限及高质量安全数据稀缺**等问题，构建面向中文场景的**细粒度安全护栏模型**，提升复杂中文风险的识别与泛化能力。

1）**规模化安全数据构建：**设计**多阶段数据构建流水线**，结合 RAG、Prompt Engineering 与多模型投票校准，覆盖 **5 个宏观、31 个细粒度风险类别**，构建 **405k 训练数据与 52k 测试数据**。

2）**Model-aware 偏好对齐：**设计**生成器-分类器协同训练框架**，引入 **MDPO** 动态调整 KL 约束，强化模型对隐式、混淆及边界风险的识别与对齐能力。

3）**实验效果：**8B 模型在 **CHILLGuardTest** 上 F1 达 **89.77**，较 **Qwen3Guard-8B-Strict** 提升 **15.92%**。