生成式AI内容安全检测与模型安全研究获进展

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/39221.html

生成式AI内容安全检测与模型安全研究获进展
。近日，中国科学院软件研究所团队聚焦多模态有害内容识别、跨模态语义检索、大模型安全防护等问题，在生成式AI内容安全检测与模型安全研究方面取得系列进展。

针对网络模因有害内容隐蔽、且在形式、主题与时间上持续演化，研究提出了RepMD方法，依托攻击树理论构建设计理念图DCG，通过对历史有害模因进行设计步骤复现和图剪枝，提炼有害模因设计流程，并利用该图指导多模态大模型进行有害模因检测。这是从有害模因图的“设计理念”角度建模有害模因的生成逻辑，为溯源和分析恶意用户的攻击行为提供支撑。实验结果显示，RepMD检测精度达81.1%，在类型迁移与时间演化场景下均保持稳定性能。人工评估显示，该方法可提升审核效率，使单个模因的判别时间缩短15至30秒。

针对短视频中仇恨信息隐蔽性强、模态干扰问题，研究提出了从特征融合转向决策仲裁的SAGE框架。SAGE设计了相互解耦的模态专家网络，保留各模态的独立语义表达，并通过全局专家协商与实例级“仲裁庭”机制，根据证据显著性动态做出判断。在经典数据集上，SAGE优于现有主流框架，准确率提升6.64%至21.23%。

针对生成式检索语义区分能力不足、对齐偏置和闭集检索限制等问题，研究提出了SIGMA框架，构建了分层语义标识符体系。SIGMA通过多粒度层级标识符，保证图像表示的唯一性与语义一致性，并提出渐进式“语义内化”训练策略，引入语义软标签刻画细粒度图文对应关系，使模型具备对未见样本动态标识符分配的能力，实现开放集检索。在经典数据集上，SIGMA在Recall@1、5、10指标上分别提升10.65%、8.50%和7.00%。

针对大语言模型面临的提示注入攻击风险，研究提出了InstruCoT方法，构建多样化攻击数据合成机制，并引入指令级Chain-of-Thought微调策略，使模型能够显式识别、推理并拒绝恶意指令。研究从行为偏移、隐私泄露和有害输出三个维度进行实验评估。结果显示，InstruCoT在四种主流大模型上均优于基线方法，并在安全增强的同时保持了模型原有的实用性能。

相关论文被自然语言处理领域顶级会议ACL 2026接收。研究工作得到国家重点研发计划的支持。

研究团队单位：软件研究所