王学康

邮箱：wangxk0223@gmail.com | 电话：+86-13161189787

教育背景

北京理工大学 09/2022-07/2026

- **专业：软件工程**
- **GPA：3.72/4.0 (89.98/100)，纯成绩排名：8/109，综合排名 3/109**
- **语言：**英语熟练（CET-4: 617, CET-6: 578，**IELTS: 7.5**）
- **编程能力：ACM-ICPC 银牌**，熟练使用 Python、C++，熟悉 Pytorch 框架下的深度学习算法实现

论文

- **Xuekang Wang**, Shengyu Zhu, Xueqi Cheng. Speculative Safety-Aware Decoding.(*EMNLP 2025, Main Conference*)
- Yifei Zhu, Hengyu Zhao, Zhongxiang Lei, **Xuekang Wang**, Di Wang, Jinyan Liu. Enhancing Robustness and Privacy: Defense Against Parallel Mixed Threats in Federated Learning.（*Under Review*）
- Ruibiao Fu[+], **Xuekang Wang**[+], April Villeda Roblero, Yang Song. Chatbots on a Mid-scale Knowledge Base: Using an R2 University as an Example. 2024 ICBAIE. *(Accepted, Co-first author)*

研究经历

Speculative Safety-Aware Decoding:推理高效的 LLM 解码阶段安全对齐方法 12/2024-05/2025

指导老师：朱胜宇副研究员，中国科学院计算技术研究所

- **文献阅读：**阅读大模型安全对齐和高效推理相关论文，在复现 Deep-Align、SafeDecoding 等方法的开源代码和文献汇报的过程中注意到推理阶段的安全防御机制在节省计算资源和可迁移性上具有较大优势。
- **提出方法：**在学习 LLM Safety 相关研究的同时广泛关注其他领域，对 LLM 推理加速的投机解码技术进行了研究，意识到基于小模型的投机解码加速方法可以在调整后应用到解码阶段防御中，提出了利用深度安全对齐的小模型对未深度对齐大模型进行投机解码时的预测接受率实现快速可迁移的安全对齐算法的 idea。
- **完成实验：**编写实验代码，基于 Pytorch 框架和 Transformers 库进行多次 pilot study，在实验中根据观察到的现象逐渐完善算法的细节，最终验证了 idea 的可行性并完成了算法的设计。在 GCG、PAIR 等攻击方法上测试了方法的 safety，在 GSM8K 和 JustEval 数据集上测试了方法的 utility，通过计算 TPS 测试了方法的解码效率，并分别与现有的多种防御的 baseline 进行比较。
- **任务与成果：**作为第一作者投稿 EMNLP 2025，负责方法提出、全部实验、论文实验部分撰写与制图制表。

一种提升 AIGC 检测模型泛化性的数据选择方法 03/2025-04/2025

指导老师：夏树涛教授，清华大学深圳国际研究生院

- **文献阅读：**系统研究了当前 AIGC 图像检测的主流方法和面临的挑战，完成多篇论文阅读笔记和课题总结报告。
- **提出方法：**在复现一个基于图像语义伪影进行 AIGC 检测的检测器时，发现基于该方法训练的检测器的跨模型泛化性与训练使用图像的质量存在很大联系。因此考虑寻找一种能高效判断训练图像对检测器泛化性提升作用的数据选择方法。经过深入调研，发现一种基于超分网络的 Zero-shot 的 AIGC 检测方法具有较高准确性，但是在一部分数据集上泛化性不足。通过分析，推测该方法的泛化性不足和语义伪影存在较大联系，可在修改后用于先前复现方法的数据选择过程。
- **完成实验：**在 base 方法没有开源的情况下，基于 Pytorch 框架自主复现并对其进行改进，实现了设想的方法。
- **任务与成果：**在 ProGAN 生成图像和 Diffusion 生成图像数据集上验证了先前推测。

FedAR-BPM：一种联邦学习中针对模型混合攻击和隐私攻击的防御机制 04/2024-12/2024

指导老师：刘金艳教授，北京理工大学计算机学院

- **文献阅读：**阅读相关文献，系统研究了联邦学习中的多种模型攻击（如模型中毒、后门植入等）在混合情景下的防御策略。
- **完成实验：**复现了基于梯度反转攻击的隐私泄露模型，复现并比较了 Laplace、Gauss 等**四种差分隐私防御机制**抑制隐私泄露风险的效果。

王学康

邮箱：wangxk0223@gmail.com | 电话：+86-13161189787

- 参与设计和实现 FedAR-BPM 框架，实现了提升联邦学习在拜占庭攻击与隐私攻击场景中的鲁棒性的目的，通过 CIFAR-10 等数据集验证效果。
- **任务与成果：**负责了大部分实验的完成。参与论文的写作过程，汇总实验结果并负责了实验部分的撰写。

通过基于知识库的大语言模型开发特定领域个性化 ChatBot 的研究 01/2024-07/2024

指导老师：Yang Song 教授，北卡罗来那州立大学 2024 冬季 GEARS 科研项目

- **文献阅读：**使用 Octopus 等自动化爬虫工具抓取数据并构建了约 150,000 tokens 的数据集，为开发 ChatBot 提供数据。
- **完成实验：**设计针对知识库集成的提示词策略，结合 Retrieval-Augmented Generation (RAG) 技术，将特定领域知识注入大语言模型，在多个基础大语言模型上成功构建个性化 ChatBot。
- **任务与成果：**设计了一套系统化的大模型回复质量评估标准，基于该标准开发测试集，并对多个主流大语言模型进行对比测试，提供全面性能分析，最后作为第一作者将研究成果发表于 ICBAIE 2024 会议。

荣誉及奖励

- ACM-ICPC 国际大学生程序设计竞赛亚洲区域赛（上海），银牌
- CCPC 中国大学生程序设计竞赛全国邀请赛，银牌
- CCPC 中国大学生程序设计竞赛（济南站），铜牌
- 字节跳动 2024 Byte AI 安全挑战赛，全国决赛第八名（唯一入围决赛本科生队伍）
- 华为 2024 年软件精英挑战赛，京津东北赛区第二名，全国决赛第九名（本科生队伍最好成绩）