姜惠康

邮箱：hj2742@columbia.edu | https://huikangjiang06.github.io

教育背景

哥伦比亚大学 2025 年 9 月 – 2029 年 5 月
计算机科学与应用数学，文学学士

北京市十一学校 2021 年 9 月 – 2025 年 5 月
高中

科研与工作经历

科研助理，哥伦比亚大学 ARiSE Lab 2025 年 9 月 – 至今
- 指导教师：Baishakhi Ray 教授、Simin Chen 教授。
- AI 驱动的软件工程。可信编程智能体、监控与验证系统、智能体训练。

科研助理，哥伦比亚大学 DitecT Lab 2026 年 1 月 – 2026 年 5 月
- 指导教师：Sharon Di 教授。
- LLM 后训练。提升 LLM 与人类的行为一致性，服务于社会科学研究。

科研助理，北京大学 PILLAR Lab 2026 年 5 月 – 至今
- 指导教师：潘亮铭教授。
- 机理可解释性。表征分析与工程、自然语言自编码器（NLA）和稀疏自编码器（SAE）。

研究工程师，KAV Labs 2026 年 9 月 – 至今
- 与哥大同仁共同打造的初创项目，结合策略推断与形式化验证开发的企业级软件工程智能体。

研究成果

SIGIL: Securing Agentic Code Generation with Intent Code Graph Learning
Huikang Jiang, Nihal Singh*, Hudson Reich, Simin Chen, Baishakhi Ray*
已投稿 *ICLR 2027*，审稿中 [PDF]
- 关键词：软件工程智能体、提示注入、模型劫持、图神经网络
- 提出一种面向 Coding Agent 的提示注入检测方法。构建问题子任务与生成代码块之间的语义关联图，并利用图神经网络进行异常检测，识别非必要且具有潜在风险的子任务。

Does Steering Break Your Model? A Multi-Dimensional Evaluation Suite for LLM Steering Methods
Haotian Yang, Huikang Jiang*, Yucheng Wu*, Wen-Jie Jiang, Chenpeng Wang, Yibin Lou, Liangming Pan*
已投稿 *ICLR 2027*，审稿中 [arXiv]
- 关键词：表征工程、激活操控、表征工程的副作用
- 开发激活操控的评估框架，从有效性、表征工程的副作用、鲁棒性、泛化能力和数据效率等维度，对比多种操控方法。

荣誉奖项

国际人工智能奥林匹克（IOAI） 铜牌，全球第 99 名
国际人工智能奥林匹克中国区学术活动（NOAI） High Distinction，全国第 5 名
中国计算机学会（CCF）全国智能无人车大赛 Distinction，全国第 8 名
美国数学竞赛（AMC） 130.5 分，荣誉榜，全球前 5%
美国数学邀请赛（AIME） 12 分，全球前 5%
国际数学建模挑战赛（IMMC） Meritorious Award

学术服务

国际人工智能奥林匹克中国区学术活动（NOAI）科学委员会成员 2025 年 11 月 – 2026 年 10 月
- 设计竞赛试题与模拟试卷。

美国-北美人工智能奥林匹克（USA-NA-AIO）考官 2026 年 1 月 – 2026 年 8 月
- 设计并验证竞赛试题与模拟试卷。

编程技能

Java, Python, PyTorch

Huikang Jiang

Email: hj2742@columbia.edu | https://huikangjiang06.github.io

Education

Columbia University September 2025 – May 2029
B.A. in Computer Science and Applied Mathematics

Beijing National Day School September 2021 – May 2025
High School

Experience

Research Assistant at ARiSE Lab, Columbia University September 2025 – Present
- Advised by Professor Baishakhi Ray and Professor Simin Chen.
- AI-driven Software Engineering. Trustworthy Coding Agents, Monitoring and Verification Systems, Agent Training.

Research Assistant at DitecT Lab, Columbia University January 2026 – May 2026
- Advised by Professor Sharon Di.
- LLM post-training. Enhancing LLM-human behavioral consistency for social science research.

Research Assistant at PILLAR Lab, Peking University May 2026 – Present
- Advised by Professor Liangming Pan.
- Mechanistic Interpretability. Representation Analysis and Engineering, Natural Language Autoencoders, and SAEs.

Research Engineer, KAV Labs September 2026 – Present
- A startup based at Columbia, enterprise-scale software engineering agents with policy inference and formal verification.

Research

SIGIL: Securing Agentic Code Generation with Intent Code Graph Learning
Huikang Jiang, Nihal Singh*, Hudson Reich, Simin Chen, Baishakhi Ray*
Under review at ICLR 2027 [PDF]
- Key Words: Agents for Software Engineering, Prompt Injection, Model Hijack, Graph Neural Networks
- Construct LLM-derived semantic graphs linking issue subtasks to code-chunk diffs and classify them (via GNNs or manually extracted features) to flag injected instructions that hijack SWE agents into producing vulnerable patches.

Does Steering Break Your Model? A Multi-Dimensional Evaluation Suite for LLM Steering Methods
Haotian Yang, Huikang Jiang*, Yucheng Wu*, Wen-Jie Jiang, Chenpeng Wang, Yibin Lou, Liangming Pan*
Under review at ICLR 2027 [arXiv]
- Key Words: Representation Engineering, Activation Steering, Side Effects
- Develop a comprehensive evaluation framework for representation engineering that compares methods across effectiveness, side effects, robustness, generalization, and data efficiency over a range of intervention methods.

Award

International Olympiad in Artificial Intelligence (IOAI) Bronze Medalist, World Rank 99
China's National Olympiad in Artificial Intelligence (NOAI) High Distinction, National Rank 5
China Computer Federation, National Driverless Car Contest Distinction, National Rank 8
American Mathematics Competition (AMC) 130.5 Points, Honor Roll, Global Top 5%
American Invitational Mathematics Examination (AIME) 12 points, Global Top 5%
International Mathematical Modeling Challenge (IMMC) Meritorious Award

Service

Member, Science Committee of National Olympiad in Artificial Intelligence (China)
Nov 2025 – October 2026
- Designed contest questions and sample exams.

Examiner, USA-North America AI Olympiad January 2026 – August 2026
- Designed and validated contest questions and sample exams.

Programming Skills

Java, Python, PyTorch