# 陈致晖

电话：(+65) 80389703 | 邮箱：zhihui.chen@u.nus.edu
求职意向：大模型研发实习 | 主页：richardchenzhihui.github.io | LinkedIn

## 教育背景

**新加坡国立大学（National University of Singapore）** 2025.01 - 2028.12
人工智能博士（Ph.D. Student in Artificial Intelligence）
- 研究方向：多模态大语言模型、医疗 AI、可信与负责任 AI。

**香港大学（The University of Hong Kong）** 2022.09 - 2024.07
人工智能硕士（M.Sc. in Artificial Intelligence）

**香港中文大学（深圳）** 2018.09 - 2022.05
统计学学士（数据科学方向）

## 研究兴趣

多模态大语言模型 • 医疗 AI 与医疗基础模型 • 可信 / 对齐技术 • 面向偏好建模的强化学习

## 论文发表

· Zhihui Chen, Kai He, Qingyuan Lei, Mengling Feng. “MedForge: Interpretable Medical Deepfake Detection via Forgery-aware Reasoning”. ACL 2026 主会。
· Zhihui Chen, Kai He, Yucheng Huang, Yunxiao Zhu, Mengling Feng. “DivScore: Zero-Shot Detection of LLM-Generated Text in Specialized Domains.” EMNLP 2025 主会。
· Zhihui Chen, Mengling Feng. “Med-Banana-50K: A Cross-modality Large-Scale Dataset for Text-guided Medical Image Editing”.NIPS2026 在投

## 学术服务

担任 ACL2026, NIPS2026, IJCAI2025, IEEE AFFECTIVE COMPUTING, ACM TIST 等多个顶会顶刊审稿人

## 科研经历

**MedForge：可解释的医疗 Deepfake 检测** 2025.01 - 至今
负责人 / Lead Researcher，Emory University Health Innovation Lab
- 开发基于 MLLM 的 MedForge-Reasoner，采用“先定位、再分析”的推理机制识别医学影像篡改，准确率达到 99.23%。
- 提出 Forgery-aware GSPO，以视觉证据约束模型推理，幻觉率降低 16.2%，解释更具临床可验证性。
- 构建 MedForge-90K 基准数据集，覆盖 CXR、MRI、眼底图像等 19 类病症，共 9 万张图像及分层专家推理标注。

**Med-Banana-50K：医学图像编辑基础数据集** 2025.05 - 至今
负责人 / Lead Researcher，NUS Medicine × National University Hospital
- 构建首个大规模指令式医学图像编辑数据集 Med-Banana-50K，涵盖胸片、脑 MRI、眼底图像等 5 万组高质量编辑对。
- 设计结合 LLM-as-Judge 与医学专用 rubric 的质检流程，围绕保真度与结构合理性进行历史感知式迭代校验。
- 发布 3.7 万条含完整对话日志的负样本，为 DPO 与多模态生成鲁棒性研究提供支持。

**DivScore：零样本 LLM 文本检测** 2025.01 - 至今
负责人 / Lead Researcher，NUS Saw Swee Hock School of Public Health
- 提出 DivScore 零样本文本检测框架，利用归一化熵评分与无监督领域知识蒸馏识别专业领域中的 LLM 生成文本。
- 在医疗与法律基准上取得 SOTA，相比最佳基线 AUROC 提升 14.4%，在 0.1% FPR 下召回率提升 64.0%。
- 从理论上分析人类文本与 LLM 文本分布的差异，证明熵归一化在领域迁移场景下的有效性。

## 实习与项目经历

**字节跳动研发实习（多模态 Agent 强化学习）** 2026.06 - 2026.11
研发实习生，字节跳动新加坡
- 参与多模态 Agent 强化学习方向研发，围绕图像/视频理解、工具调用与长链路任务完成等核心能力，支持训练数据构建、奖励设计与评测闭环搭建。
- 负责轨迹数据清洗、偏好数据组织与样本质量分析，协助建设任务回放、结果校验与失败模式归因流程，为在线 / 离线强化学习迭代提供高质量数据支持。
- 配合模型评测与实验分析，围绕复杂界面操作、视觉推理与真实任务稳定性等指标输出分析结论，支持 Agent 能力优化与版本迭代。

**阶跃星辰研发实习（语音大模型）** 2025.01 - 2025.05

研发实习生，上海阶跃星辰智能科技有限公 0

- 参与语音大模型训练数据体系建设，面向 ASR、TTS 与实时语音交互等业务场景，整合开源、合作方与内部沉淀语料，完成语音-文本配对数据的标准化入库。
- 负责数据治理环节，包括音频切分、转写对齐、重复样本去重、低质样本过滤，以及说话人、语种、情绪与场景标签规范化，提升预训练语料的一致性、可控性与可复用性。
- 设计并维护预训练数据组织方案与元数据 schema，支持多语种、多方言、情绪风格控制等维度的检索、抽样、质检与版本迭代，为后续模型训练、评测和数据回溯提供稳定的数据底座。

**Quant Trading Agent：港股自动化交易智能体** 2024.05 - 2024.7

量化交易研发实习生，AQUMON（香港）

- 基于 LangChain 架构参与量化交易 Agent 开发，打通行情理解、信号生成、策略决策与执行监控等核心链路，支撑港股程序化交易场景的自动化运行。
- 对接 Futu OpenAPI 完成实时行情接入、模拟自动下单执行，协助搭建交易策略验证与联调流程，提高策略迭代与落地效率。

**Smart Word Agent：智能文档助手** 2024.02 - 2024.04

作者 / 主要开发者，GitHub 开源项目

- 基于 ReAct Agent 思路构建自然语言驱动的 Word 文档解析、编辑与重排流水线。
- 支持多文档上下文、附件推理、批量格式调整、表格操作，并以 Kimi-K2 作为核心大模型。
- 发布单文件便携 exe 版本，上线 1 周下载量超过 1,000，GitHub 获得 100+ Stars。

**社会价值投资联盟（深圳）实习** 2020.01 - 2020.06

数据分析实习生，社会价值投资联盟（深圳）

- 参与《ESG 义利 99（2020）》报告的数据整理与研究支持工作，围绕上市公司 ESG 指标、年报文本与公开披露材料开展转写、清洗与结构化处理。
- 使用 Excel 与 Python 进行指标校核、描述性统计、口径对齐与图表底稿整理，支撑报告中图表、附录与数据说明模块的撰写与交付。
- 协助建立跨来源资料的一致性核验与版本管理机制，提升研究资料的可追溯性与分析效率，支持 ESG 研究项目的标准化产出。

## 教学经历

**教学助理：《Generative AI and LLM》** 2024.02 - 2024.04

NUS Business School，Prof. Pang Yan

- 参与课程知识框架与教学内容设计。
- 编写讲义与课程项目，涵盖 PyTorch、HuggingFace、LLM 预训练、SFT、RLHF 等主题。

## 获奖情况

· 新加坡国立大学全额博士奖学金（2025 - 2029）
· 香港中文大学（深圳）优秀毕业生（前 5%）
· 香港中文大学（深圳）本科生研究卓越奖

## 技能

| | |
|---|---|
| **编程** | Python, R, C++, MATLAB, MySQL |
| **深度学习框架** | PyTorch, TensorFlow, scikit-learn, HuggingFace |
| **AI 工具** | Linux, Docker, Git, Llama.cpp |
| **语言** | 英语（TOEFL 106, GRE 320）、普通话（母语）、粤语（母语） |

兴趣：摄影、骑行、小提琴、跑步