

丁屹

yding834@connect.hkust-gz.edu.cn | 17280255217

应聘方向：算法工程师（日常实习）| 大模型推理 / Agent 系统 / 模型训练与评测 / RAG

实习时间：可尽快到岗，每周 5 天，可持续 6 个月以上

教育背景

香港科技大学（广州），人工智能哲学硕士（MPhil）	2025.09 - 预计 2027.09
爱丁堡大学，人工智能理学硕士	2024.09 - 2025.09
诺丁汉大学，人工智能荣誉理学学士，一等荣誉学位	2020.09 - 2024.06

论文与奖项

- EMNLP 2026 已接收（第一作者）：Yi Ding, Lijun Huang, Menglin Yang. *SCIT: Testing Causal Cache Carriers in Latent Chain-of-Thought Models*.
- EMNLP 2026 已接收：Jinchang Zhu, Jindong Li, Yi Ding, Nie Xiaojian, Rong Fu, Shuangyong Song, Haowei He, Menglin Yang. *FCPRAG: Fusion-Controller Parametric Retrieval-Augmented Generation for Stable Multi-Passage LoRA Injection*.
- 独立完成并以唯一作者发表两篇 MARIO MICCAI 2024 / DIAMOND Challenge 论文，收录于 LNCS 15503 (Springer)。
- 获 MARIO MICCAI 2024 挑战赛第三名与 Best Method。

实习经历

研究助理 深圳湾实验室 王超	2024.05 - 2024.08
• 面向 Rubisco 酶活性预测，整理蛋白序列和实验标签，用 ESM 构建特征，并对比 GP、LightGBM、CatBoost 等模型。 • 设计多核 GP，比较核函数对预测误差和不确定性估计的影响，并结合特征重要性与残差分析筛选突变位点。 • 将模型结果整理为候选位点清单、可视化分析和可复现科研报告，为后续突变实验设计提供依据。	

项目与科研经历

SCIT 延伸：答案传输与破坏性敏感性 ICLR 2027 准备中 杨萌林	2026.06 - 至今
• 提出双估计框架，分别量化答案传输与破坏性敏感性，避免将“可注入答案”误判为模型的自然计算路径。 • 在 CODI-GPT-2 前瞻性实验中，latent state 与生成前缀分别实现 15/16、16/16 的答案传输，并通过交叉覆盖识别三-token relay；在独立训练的 CODI-Llama-1B 新样本上复现同一方向性模式（16/16、14/16，反向覆盖后均为 0/16）。	
FCPRAG：多段 LoRA 融合控制 EMNLP 2026 杨萌林	2026
• 构建轻量 Fusion Controller，按样本预测 passage-LoRA 的融合权重、门控与温度，在检索不确定时自动回退至保守融合，相对主干模型新增可训练参数不足 1%。 • 利用 passage adapter 在多 LoRA 合并中的边际贡献构建融合监督；在四个 QA 数据集和三种 LLM 主干上验证，2WikiMultiHopQA 与 CWQ 的 F1 最高分别提升 4.65% 和 7.55%，并增强检索扰动鲁棒性。	
SCIT：潜在思维链因果缓存载体测试 EMNLP 2026 杨萌林	2026.03 - 2026.05
• 提出 SCIT，通过源-目标反事实、K/V 替换、语义控制与匹配扰动，识别 latent-CoT 的因果缓存载体。 • 在 CODI-GPT-2 与 Sim-CoT GPT-2 中，答案经 value-cache 后缀传递，而非 hidden state、key、答案槽或单-token 触发。 • 构建能力门控的载体图谱：GPT-2/1B 算术任务主要依赖 latent-tail value/KV，具备任务能力的 8B 与非算术任务则转向 prompt-prefix 或 full-cache K/V；基础任务未通过时不作机制判断。	

项目与科研经历（续）

因果 SAE 特征选择的目标选择边际 | 已投稿 NeurIPS 2026 | 杨萌林 2026.01 - 2026.04

- 提出目标选择边际（TSM），以目标词-干扰词概率差衡量 SAE 特征定向效用，并检验其在未见提示词上的迁移。
- 在 Qwen、Gemma、Llama 上比较目标感知与随机/通用选择策略，评估不同目标、层级和强度下的增益及副作用。
- 通过留出提示词、标签打乱和跨模型消融区分目标特异信号与一般扰动，并分开报告目标概率边际和生成变化。

大语言模型推理质量优化 | 已投稿 ICML 2026 | 杨萌林 2025.09 - 2026.01

- 提出 REWIND，以熵梯度定位高风险片段，回溯语义边界后局部重生成，并用 MH-style 过滤器保留更优轨迹。
- 在五个推理基准和三类模型上验证；Qwen2.5-Math-7B 在 MATH500 达到 77.6%（MCMC: 72.6%），DeepSeek-R1-Distill-1.5B 在 AIME2025 达到 26.67%（MCMC: 16.67%），增益主要集中于高难度数学任务。

基于 CycleCoT 的大小模型协作推理 | 已投稿 ACL | 杨萌林 2025.08 - 2026.01

- 设计双向潜空间推理框架，对齐小模型隐式状态与大模型显式推理能力，探索复杂任务中的低成本协作。
- 使用残差转换网络与循环一致性约束稳定表示映射，在数学、常识和符号推理任务上优于 SoftCoT 与 Coconut。

强化学习提示词生成器 | 王孟宇 2024.09 - 2025.02

- 构建提示词生成、自动评测与反馈迭代管线，统一组织任务输入、候选 prompt、模型回答和评分信号。
- 结合 PPO 与 LoRA 微调 FinQA，并在 PubMedQA、GSM8K 和算术任务上评估奖励设计、迁移效果与训练稳定性。

MARIO MICCAI Challenge 医学影像预测 | 唯一作者 2024.06 - 2024.11

- 面向新生血管性 AMD 的 OCT 序列诊断与进展预测，构建时序与组级建模管线，覆盖预处理、特征提取与训练评估。
- 使用 ConvNeXt V2、Siamese OCT-DiffNet 与 CLAM-SB 建模序列差异和病例风险，并以加权采样缓解类别不平衡。
- 独立完成两篇 MICCAI 2024 / DIAMOND Challenge 论文，取得挑战赛第三名与 Best Method，F1 指标超过 0.84。

BNOTE 持续测试时自适应 | 钟准 2023.08 - 2024.06

- 设计持续测试时自适应框架，结合实例感知均衡 BN 与预测均衡蓄水池采样，适配低标签与分布漂移。
- 以熵值和 margin 阈值控制伪标签筛选与自适应损失，降低噪声标签对在线更新的影响，提升连续测试稳定性。
- 在 CIFAR10-C、CIFAR100-C、ImageNet-C 等非 IID 基准上较 SOTA 降低约 10% 分类错误率，仅更新 0.02% 参数。

技能与岗位匹配

- **大模型与 Agent**：熟悉 LLM 推理、RAG、提示词优化、工具调用与 Agent 编排；使用 Codex、Claude 完成代码重构、实验自动化和分析原型开发。
- **训练与评测**：具备 PyTorch、Hugging Face Transformers、LoRA、PPO/RL 与 SFT 经验，可设计评测、消融和误差分析，权衡质量与推理成本。
- **机器学习工程**：熟练使用 Python、scikit-learn、LightGBM、CatBoost，掌握特征工程、模型对比、消融与误差分析。
- **工程与交付**：熟悉 Git、Linux/HPC、SLURM 批量实验与日志分析，能整理可复现实验脚本、评测表格和技术报告。