杜雅欣

dorothydu@sjtu.edu.cn Google Scholar GitHub Homepage

求职方向：LLM Agent / Coding LLM / Mid-training

教育经历

上海交通大学 2022.09 – 至今

博士，信息与通信工程，导师：陈思衡

电子科技大学 2018.09 – 2022.06

本科，通信工程

科研方向

代码大模型与 MidTrain 数据：面向代码大模型 MidTrain 构建多语言、多任务软件工程数据，关注数据合版、质量评估、数据选择、数据配比设计与规模化训练策略。

代码生成与软件工程智能体：构建自动化、可执行、可评测的软件开发系统，关注仓库级任务、代码运行反馈、测试驱动优化与多智能体协作。

大模型智能体与工具调用：研究智能体的工具选择、轨迹合成、自进化协作与复杂任务评估，提升模型在长链路任务中的规划、执行与验证能力。

科研经历

软件工程任务与开发智能体：SWE-Dev 2025.05 – 2026.02

- **SWE-Dev（NeurIPS 2025 DL4C Workshop，第一作者）：**现有代码智能体评测多集中在函数补全或缺陷修复，难以覆盖真实软件开发中“在已有仓库中实现新功能”的端到端需求。本文构建 SWE-Dev，将真实仓库上下文、可运行环境和开发者编写的单元测试组织成可训练、可评测的软件工程任务，使其同时支持监督微调和强化学习训练。
- 数据集包含 **14,000** 个训练样本与 **500** 个测试样本，评测覆盖 **17** 个基础大模型、**10** 个推理模型、**10** 个多智能体系统和 **8** 个工具增强智能体；OpenHands 在困难划分上达到 **56.44%** Pass@1。
- **后续拓展：**将 SWE-Dev 扩展为面向 mid-training 的多语言仓库级软件工程数据，覆盖 **8** 种编程语言，并构造约 **100 万**条智能体开发轨迹，用于提升模型在真实仓库环境中的需求理解、代码修改、测试反馈和多步工具调用能力。

Multi-agent 系统：多模态 RAG、AutoResearch、Self-evolving 2024.10 – 2026.05

- **G^2-Reader（ICML 2026，第一作者）：**长文档多模态问答需要同时理解文本、表格、图片和跨页结构，而普通 RAG 的扁平切块容易破坏文档原生结构。本文构建 G^2-Reader，通过内容图保留文档结构与跨模态语义，并用规划图追踪子问题、证据链和中间结论；在覆盖五类多模态领域的 VisDoMBench 上达到 **66.21%** 平均准确率，显著超过独立 GPT-5 的 **53.08%**。
- **DataMaster（NeurIPS 2026 投稿中，第一作者）：**机器学习系统的进一步提升越来越依赖数据侧优化，但数据发现、清洗、转换和验证仍高度依赖人工经验。本文提出任务条件化的自主数据工程设定，并构建 DataMaster，通过 DataTree、共享数据池和全局记忆组织分支式数据搜索与复用；在 MLE-Bench Lite 上相比初始方案将奖牌率提升 **32.27%**，并在 PostTrainBench 的 GPQA 上超过 instruct model（**31.02%** vs. **30.35%**）。
- **EvoMAC（ICLR 2025）：**面向传统多智能体系统依赖人工设计协作流程、难以随任务反馈持续改进的问题，参与提出自进化多智能体协作框架 EvoMAC。该方法将多智能体协作网络表示为可优化的文本结构，并利用环境反馈与文本反向传播在测试时自动更新角色分工、通信关系和执行策略。构建 rSDE-Bench，作为面向需求驱动软件级开发的黑盒执行评测基准，将多智能体代码能力评测从函数级代码扩展到完整软件开发任务；rSDE-Bench 包含 **53** 个开发任务和 **616** 条需求，自动评测与人工评测一致性达到 **99.22%**。

工具调用与多源信息检索：InfoMosaic-Bench、MCP-Flow / MCP-Persona 2025.06 – 2026.04

- **InfoMosaic-Bench（ICLR 2026，第一作者）：**真实信息检索任务往往不能只依赖网页搜索，还需要调用医疗、金融、地图、视频等领域工具并整合多源证据。本文构建 InfoMosaic-Bench，并通过 InfoMosaic-Flow 合成具备跨源依赖、可验证且不能被简单查表解决的任务；评测覆盖 **6** 类场景和 **14** 个先进智能体，GPT-5 仅依赖网页信息时准确率只有 **38.2%**，且 **22.4%** 失败来自工具选择或调用错误。
- **MCP-Flow（ACL 2026）：** MCP 生态快速扩张，但现有研究通常覆盖少量服务、依赖人工整理，难以支撑模型在真实大规模工具环境中的训练。本文构建自动化的 MCP 服务发现、数据合成与训练流程，覆盖工具理解、工具选择、参数生成和多步调用轨迹；收集过滤 **1,166** 个服务和 **11,536** 个工具，生成 **68,733** 条指令–函数调用对与 **6,439** 条调用轨迹。
- **MCP-Persona（ICML 2026）：**面向 Claude Desktop、Cursor 等 MCP 客户端在真实应用中接入大规模工具集合时的个性化调用需求，覆盖代码开发、办公协作、知识库检索、日程管理和数据分析等场景。构建角色条件化的工具调用数据与评测设定，使智能体能够根据不同用户角色和任务目标，在复杂工具集合中更稳定地完成工具检索、选择和组合。

联邦学习与大模型数据质量 2023.05 – 2025.02

- **FedDQC（ACL 2025 Findings，第一作者）：**联邦学习可以在不共享原始数据的情况下协同训练大模型，但客户端数据质量不可见会放大低质量样本对全局模型的影响。本文提出 FedDQC，通过低成本的指令–回答对齐指标在本地评估数据质量，并设计从高质量到低质量的分层联邦训练流程；实验覆盖 **4** 个合成任务数据集和 **1** 个真实 Fed-WildChat 数据集，数据打分仅约占总训练时间的 **1%**。
- **FedLLM-Bench（NeurIPS 2024）/ OpenFedLLM（KDD 2024）** 参与构建现实大模型联邦学习基准与开源训练框架，覆盖联邦指令微调、联邦偏好对齐、**7** 种代表性联邦学习算法、**8** 个训练数据集与 **30+** 个评测指标；FedLLM-Bench 客户端规模从 **38** 到 **747**，为隐私数据协同训练提供更现实的测试平台。

实习经历

IQuest Lab 2026.01 – 至今

研究实习生

- **MidTrain 数据合成：**整合多来源、多格式、多任务数据，统一数据格式、质量过滤、去重、采样权重与版本管理，支撑模型 MidTrain 数据闭环。
- **多语言、多任务软件工程数据扩展：**构建代码生成、缺陷修复、仓库级功能开发、测试生成、工具调用等多任务软件工程数据流程，扩展语言与任务覆盖。
- **MidTrain 数据选择：**基于质量规则、模型打分、分布覆盖与训练反馈筛选高价值样本，优化数据配比；参与发布 **iQuest-V1** 7B/14B、**工业代码模型 32B** 与 **Thinking 32B** 版本。

TikTok AI Innovation Center 2025.09 – 2026.01

研究实习生，Mentor：刘乾

- 通过强化学习训练 **co-evolving code agents**，使代码生成器与验证器在代码–测试交互中协同提升。
- 扩展代码强化学习环境与合成数据规模至 100K 量级，支持更大规模的代码智能体训练。

项目经历

SciMaster：面向高难度科学任务的数据构建 2025.03 – 2025.08

- 负责数据团队管理与流程设计，从本科与研究生教材中收集 **1.33B** 高质量训练数据，面向 HLE-Bench 等高难度科学任务构建训练语料。
- 参与数据清洗、章节结构解析、题目与解答抽取、质量筛选和版本管理，提升科学推理、跨学科知识问答与复杂题解任务的数据覆盖。

AI Study Buddy：科研协作与学习助手 2024.12 – 2025.02

- 整理 **100K** 科研对话数据，覆盖论文理解、研究讨论、实验设计与学术写作等场景，支撑面向科研用户的对话式智能助手训练。
- 基于监督微调与偏好优化微调 72B 大模型，使单轮回答准确率提升 **29%**，并提升模型在科研问答中的事实性和可用性。

代表论文

DataMaster: Data-Centric Autonomous AI Research（NeurIPS 2026 投稿中，共同第一作者）

InfoMosaic-Bench: Evaluating Multi-Source Information Seeking in Tool-Augmented Agents（ICLR 2026，第一作者）

G^2-Reader: Dual Evolving Graphs for Multimodal Document QA（ICML 2026，共同第一作者）

SWE-Dev: Evaluating and Training Autonomous Feature-Driven Software Development（NeurIPS 2025 DL4C Workshop，共同第一作者）

FedDQC: Data Quality Control in Federated Instruction-tuning of Large Language Models（ACL 2025 Findings，第一作者）

Enhancing Data Quality in Federated Fine-tuning of Large Language Models（ICLR 2024 Workshop，共同第一作者）

MCP-Persona: Benchmarking LLM Agents on Personalized MCP Tools and Tasks（ICML 2026）

MCP-Flow: Automating Tool-Based Agentic Workflows on MCP（ACL 2026）

NTK-Selector: Selecting Auxiliary Data via Neural Tangent Kernels for Low-Resource Domains（ICML 2026）

InCoder-32B: Code Foundation Model for Industrial Scenarios（arXiv 2026）

Self-evolving Multi-agent Collaboration Networks for Software Development（ICLR 2025）

MAS-GPT: Training LLMs to Build LLM-based Multi-Agent Systems（ICML 2025）

Federated instruction tuning of LLMs with domain coverage augmentation（EMNLP 2025）

BrowseMaster: Towards Scalable Web Browsing via Tool-Augmented Programmatic Agent Pair（NeurIPS 2025 Workshop）

OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning（KDD 2024）

FedLLM-Bench: Realistic Benchmarks for Federated Learning of Large Language Models（NeurIPS 2024）

Fake It Till Make It: Federated Learning with Consensus-Oriented Generation（ICLR 2024）

荣誉奖项

国家奖学金　　四川省优秀毕业生

专业技能

编程与工程： Python、C++、PyTorch、Git、Linux、Docker；熟悉大模型训练、数据构建、自动化评测。

英语： CET-6（619），IELTS（7.0）。