

蔡逸文

(+86) 189-2876-9793 | 北京, 中国 | caiyiwen.cs@foxmail.com | github.com/yiwei-cai | yiwen-cai.github.io

教育经历

北京邮电大学(BUPT)

硕士学位 (推免), 计算机技术

北京, 中国

Sep 2026 — Jun 2029

北京邮电大学(BUPT)

学士学位, 计算机科学与技术

北京, 中国

Sep 2022 — Jun 2026

- GPA: 3.76 / 4.0
- Grade: 89.37 / 100.0
- Rank: 59 / 389 (15.16%)
- English: CET6 552 / 710

实习经历

LLM 高性能推理框架调度及高性能算子优化研究

Jul 2026 — Present

腾讯混元 AI Infra 部

北京, 中国

- 负责 LLM 推理服务的推理框架调度优化与高性能算子优化
- 开发算子包括 Blackwell 架构 BF16 paged KV attention prefill、nvfp4_blockwise_gemm、per_group MXFP8 量化、fused_silu_per_token_quant 融合算子等
- 主要开源代码库: [Tencent/hpc-ops](https://github.com/Tencent/hpc-ops)
- 部分贡献代码已并入 vLLM main 分支

项目经验

第十届华为 ICT 大赛中国总决赛挑战赛

Apr 2026 — Apr 2026

华为

华中科技大学, 武汉

- 在 64 支各高校队伍中排名第 8 进入决赛
- 决赛 16 支队伍中总排名第 3 获得一等奖
- 作为队伍唯一开发者拿到算子迁移优化赛题所有队伍最佳优化
- 成绩: 全国一等奖, 最高算子迁移优化奖
- 项目 (部分):
 - [fused_add_rmsnorm_kernel](#)
 - 一个 triton 编写的 RMSNorm 和 add 融合算子
 - 多行 program 映射
 - hidden_size 分档调度
 - 性能相比 pytorch 加速比 19.28x
 - [xllm-ictfinal](#)
 - 基于 xllm 推理引擎进行 Qwen3.5 模型推理适配
 - 实现 GDN 线性注意力适配
 - 实现 MTP 投机采样推理
 - 解决 KVCache 占用异常
 - 成功运行 32k 输入推理

cuda-gemm

Jun 2025

- 从零手写 8 个 CUDA GEMM kernel, 覆盖 FP32 CUDA Core 和 FP16 Tensor Core WMMA 两条优化线
- 渐进式引入 Shared Memory Tiling、Bank Conflict Avoidance、Double Buffering、cp.async 异步拷贝等优化
- SGEMM v3 达 cuBLAS 的 95.1% (4096²: 37,898 GFLOPS; 8192²: 36,799 GFLOPS)
- HGEMM v3 Tensor Core 达 213 TFLOPS at 4096², 含 Nsight Compute/Systems profiling 分析

新加坡国立大学夏令营 NUS Summer Workshop

Jul 2024 — Aug 2024

National University of Singapore

新加坡

- 参加了云计算与大数据课程, 并完成了小组设计的项目
- 成绩: A+

- 排名: 第 1 名
- 项目: [UPLION](#)
 - 一个统一各种 AI Agent 的平台, 用于综合 AI 操作和部署
 - 云原生设计, 支持可扩展性和弹性
 - 智能任务分配和资源优化
 - Kubernetes 集成, 用于高效的工作节点管理

Jul 2024 — Aug 2024

[FastGNN](#)

May 2026

- 一个基于 TensorCore 的图神经网络训练系统
- H100 full-batch GNN
- 自研稀疏算子后端
- Voltrix block-sparse SpMM
- Fused3S sparse attention
- Rabbit 节点重排
- 端到端 benchmark 闭环
- DGL/PyG 对照验证, 加速比 1.8x-3.3x

奖项

一等奖, 最高算子迁移优化奖, 第十届华为 ICT 大赛全国总决赛挑战赛	Apr 2026
A+, NUS Summer Workshop	Jul 2024
二等奖, 第十五届蓝桥杯全国软件和信息技术专业人才大赛北京赛区	Apr 2024
大三学年校级奖学金	2025
大二学年校级奖学金	2024
大一学年校级奖学金	2023

技能

- 编程技能: Python, C/C++, CUDA, Triton, Rust, Go
- 研究方向: LLM 训练推理优化, 算子优化, 分布式计算与云计算, 高性能计算