

吴奕民

(+86) 173-2677-1745 | ymwu.ai@gmail.com | github.com/kosthi

教育经历

沈阳工业大学 | 计算机科学与技术

2021.09—2025.06

GPA: 3.4/4.0 (专业前 10%)

专业技能

C/C++、Python (PyTorch)、Go | 数据结构与算法、操作系统、计算机网络、数据库原理

竞赛获奖

- 2023 年美国大学生数学建模竞赛 国际一等奖
- 2024 年全国大学生信息安全竞赛 国家二等奖
- 2024 年 OceanBase 数据库大赛 全国二等奖
- 2024 年计算机系统能力大赛 全国一等奖 (2/325)

发表论文

LiteQG: Towards Scalable and Memory-Efficient Graph-Based Approximate Nearest Neighbor Search

ICIC 2025

- 负责图索引构建与检索算法的核心实现，包括基于 HNSW 的邻居选择策略改进与启发式剪枝优化。
- 设计并实现向量量化压缩方案 (SQ/PQ)，在标准数据集上将内存占用降低的同时保持 0.99 以上召回率。
- 完成大规模数据集上的性能基准测试与对比实验，验证了 LiteQG 在可扩展性和内存效率上的优势。

项目经历

斯坦福 CS336 大语言模型入门 | [GitHub](#)

2025.01—至今

- 从零实现完整的 Transformer 语言模型，包括 RoPE 位置编码、RMSNorm、SwiGLU 激活、因果多头自注意力等核心组件，并实现 BPE 分词器、AdamW 优化器与余弦学习率调度。
- 实现分布式训练优化，包括 FlashAttention (PyTorch/Triton)、DDP 梯度同步与分桶策略、优化器状态分片，降低多卡训练的通信开销与显存占用。
- 实现大模型对齐技术，包括 SFT 监督微调、DPO 直接偏好优化与 GRPO 群体相对策略优化，完成奖励归一化、策略梯度损失计算与 clip 损失等核心算法。

OpenACE 代码上下文检索引擎 | 独立开发 | [GitHub](#)

2025.02—至今

- 基于 Rust 实现多信号代码检索引擎，集成 Tantivy BM25 全文检索、usearch HNSW 向量相似度搜索与 SQLite 图遍历三路信号，通过 RRF 算法融合排序，单信号故障自动降级。
- 使用 tree-sitter 实现多语言 (Python/TypeScript/Rust/Go/Java) AST 解析，提取符号与调用关系，单线程解析速度超 50,000 symbols/s，结合 Rayon 并行化实现万级文件 30 秒内全量索引。
- 设计增量更新机制，基于 XXH3-128 内容哈希检测变更，配合 notify 文件监听实现 1 秒内变更感知，单文件增量更新延迟低于 500ms。
- 采用 OpenSpec 规范驱动开发，为 10 个核心模块编写机器可读的行为契约与性能指标，作为 MCP Server 集成至 Claude Code 等 AI 开发工具。

荣誉证书

- 沈阳工业大学创新创业奖学金 & 创新创业先进个人
- 2024 年腾讯犀牛鸟开源人才培养计划 3TS 课题优秀学生