Asc Ho

个人主页 ascendho 602 | 1.4k ascendho@outlook.com

教育背景

西南石油大学 · 计算机与软件学院	软件工程	硕士在读	2024.09 – 2027.06
西南石油大学 · 计算机与软件学院	计算机科学与技术	本科	2020.09 – 2024.06

实习经历

北京知道创宇信息技术有限公司 2026.04 – 至今

面向 ToB 私有化部署的 DeepSeek 双 DGX Spark 分布式推理与性能优化

项目背景：面向 AiPy 企业智能体的本地模型服务需求，评估双 DGX Spark 作为 1～5 用户小规模企业部署节点的可行性，基于 **DeepSeek-V4-Flash-0731** 构建并测试高性能推理服务。

- 负责 **2× NVIDIA DGX Spark（GB10）**的机房硬件部署与节点互联，完成双机管理网与计算网私网的 IP 配置。
- 基于 vLLM 搭建双 DGX Spark 分布式推理架构，采用 TP=2 部署 DeepSeek V4 Flash，`rank0` 对外提供 OpenAI-compatible API 并负责请求调度，`rank1` 作为 headless 计算节点参与跨机并行推理。
- 通过 QSFP + RoCE + NCCL 构建双机高速通信链路，优化 Tensor Parallel 跨节点集合通信性能，将 NCCL 总线带宽从 98 Gb/s 提升至 161 Gb/s（+64%）。
- 对齐模型 *YaRN* 校准上限实现 **1M 长上下文推理**，基于 `nvfp4_ds_mla` 降低 KV Cache 显存占用，并结合 PagedAttention、Chunked Prefill 与 Prefix Caching 优化长 Prompt 处理与前缀复用，提升内存效率与并发能力。
- 基于 EvalScope 开展 1～20 并发多轮 Agent 压测，DSpark 在目标 1～5 并发下较无投机解码实现单会话 Output Throughput 提升 71%～87%、TPOT 降低 42%～47%，5 并发达到 106.84 tok/s 聚合吞吐与 1.05 s TTFT。

基于 AiPy 智能体集市的项目视觉生成 Skill 设计与开发

项目背景：面向公司 AiPy 智能体集市生态建设需求，独立设计并开发项目视觉生成技能，实现智能体对项目内容的自动理解与视觉方案生成，支持 Logo、项目封面、社交分享图等物料制作，并完成产品化与集市接入。

- 设计涵盖项目分析、视觉规划、用户确认、渲染生成与结果校验的 Agent 工作流，自动解析 README、项目配置、入口代码及已有视觉资产，提取项目定位、核心功能与视觉语义，并据此规划多套视觉方案。
- 设计并实现 **Human-in-the-loop** 多阶段 Agent 工作流，沙箱目录产出 3 套 Logo 概念对比方案后硬性阻塞等待用户决策，未确认的品牌资产禁止流入下游产物，解决 Agent 在主观决策场景下擅自拍板的问题。
- 构建基于 LLM 文案、SVG 模板与 Chromium 截图的确定性渲染管线，通过实测文本宽度自动缩字，支持 CJK/RTL 多语言布局并实现像素级可复现输出，且不依赖生图模型，**可适配任意 LLM**，有效规避图像幻觉与模型绑定问题。
- 完成 Skill 的 AiPy 智能体集市适配与发布，并因实际效果**获得负责人认可**，作为 Showcase 用于公司内部及对外演示。

个人项目

E-Snap：两级缓存及语义增强的智能电商客服 *code* demo

- **技术栈**：Python、FastAPI、LangGraph、LangChain、RedisVL / Redis Stack
- **项目背景**：面向电商客服中高频问题重复检索、LLM 调用成本高及动态问题错误缓存复用等问题，设计两级语义缓存增强 RAG 智能体，通过缓存复用与动态路由降低响应延迟及模型调用成本。
- 设计动态查询前置路由机制，对时间、库存、具体商品型号等不适合静态复用的请求进行拦截，避免动态问题错误命中缓存，并在系统启动阶段**预加载高频 FAQ** 以提升常见问题响应效率。
- 构建 L1 规则缓存与 L2 语义缓存**两级复用机制**，L1 基于归一化、编辑距离与子问题匹配优先复用高置信度历史答案，L2 结合向量召回与关键词检索完成语义相似问答匹配。
- 引入 LLM Validator 与 **ReAct** RAG 作为渐进式兜底机制，对缓存候选判断直接复用、补充缺失信息或回退完整 RAG，仅在缓存无法覆盖或复用收益不足时执行知识检索与答案生成，并将可复用结果按规则写回缓存形成闭环。
- **实验评估**：在 30 条离线评测请求上，系统前置拦截 4 条动态类查询，缓存覆盖率达到 42.31%，其中直接缓存复用率为 30.77%，使端到端延迟降低了 26.64%，系统吞吐量提升至原来的 1.36 倍；成本方面，相对纯 RAG 基线，平均每次请求减少 1.14 次 LLM 调用，费用整体节省 32.36%。

ChatAnchor：面向 AI Coding Agent 的跨路径会话恢复工具 *code* marketplace

- **技术栈**：TypeScript、Node.js、VS Code Extension API、Git、SQLite、Zod、Vitest、GitHub Actions、pnpm
- **项目背景**：针对 Codex、Cursor Agent CLI、OpenCode 等 AI Coding Agent 会话与项目工作目录强绑定、改名即断链的问题，独立设计并开发本地 VS Code 插件，实现 Agent 会话与项目路径解耦。
- 通过稳定 **Project UUID** 持久化项目身份，并结合 Path Alias、Git Remote、Commit SHA 等多维证据进行会话匹配，按多级置信度实现项目迁移后的历史线程自动关联、候选推荐与人工纠错。
- 统一适配 Codex App Server、Cursor Agent CLI 本地元数据及 OpenCode SQLite 数据库，封装会话发现、关联与 Resume 流程，在新工作目录下继续原有 Agent 上下文。

科研成果

PROSE：面向 LLM KV Cache 共享 CXL 内存池的对象准入机制 *code*

- **论文信息**：第二作者，已投稿至 **HPCA 2027**（在审）
- **问题定义**：面向长上下文 LLM Serving 的 CXL KV Cache 共享内存池，发现并定义 **Reclaimed Payload Exposure (RPE)**：异步读请求排队期间，若目标内存槽被回收并复用于其他对象，请求仍会返回地址合法且缓存一致、却属于错误对象版本的数据，而 CXL 协议只能保证地址有效与字节最新，无法感知应用层的对象身份。对部署系统 Mooncake Store（commit f20b706，已合入官方 2026 年修复）的源码审计与负载重放证实，该竞态在生产代码路径中仍真实存在（绕过防护时 6.15% 的传输交付了错误对象），且现有修复未能完全覆盖。
- 提出**对象准入事务**（Object Admission Transaction, OAT），在 CXL 内存扩展器（Type-3 endpoint）控制器的数据路径中新增准入逻辑，把数据真正发出前的最后一刻设一道" 准入闸"。
- **检查与加锁一步完成**：在同一原子步骤内核对对象版本并临时锁定该映射，两步之间不留缝隙，回收只能排在这步之前或之后；校验失败的请求直接拒绝、一字节不发，错误对象的数据不进入链路。
- **锁定时间最短**：锁定只覆盖传输本身，约为入队即锁方案保护时长的 **1/3**，请求排队期间槽位照常回收利用，不牺牲池化容量；拒绝仅产生 128 B 元数据、零数据载荷，可安全重试。
- **不改协议、开销可忽略**：仅使用 CXL 标准读写与通知接口实现，无需修改互联协议；新增准入逻辑仅 0.0061 mm² / 5.83 mW（不足 DDR5 级 CXL 控制器预算的 1%），准入判定仅 9 ns。
- **实验评估**：在 BurstGPT 真实负载与公开 LLM Trace 的跨层验证（SimCXL 模拟 + RTL 协同仿真 + TLA+ 形式化验证 + ASAP7 7nm 硬件结构估算）下，Mooncake 现有机制只能在" 丢弃 49% 读带宽" 与" 锁死 39.3% 池容量" 之间二选一，其 2026 年修复未覆盖的异步传输路径上仍发生 **6 次**误读事件；PROSE 在相同负载下将误读其他对象 KV Cache 的字节降为 0%，同时保持 93.8% 池容量可回收，并取得最高 3.1× 端到端加速。

LumiSense：面向工业 IoT 网关的轻量大模型边缘根因诊断框架 🤗 weights （**下载量 164**）

- **论文信息**：第一作者
- **问题定义**：针对工业 IoT 场景中设备状态数据异构、运行信息不完整以及故障类型持续演化等问题，传统分类模型受限于固定标签空间，难以应对未知故障与跨场景泛化；同时大参数量 LLM 存在部署成本高、难以适配边缘设备的问题。因此，研究大模型能力向轻量语言模型（SLM）迁移的方法，实现资源受限环境下高效可靠的故障根因诊断。
- **核心方法**：基于 Qwen3.5-0.8B 构建轻量化诊断模型，通过 LoRA-SFT 注入工业故障诊断知识与结构化推理能力，并利用 Qwen3.5-27B 教师模型进行知识蒸馏，迁移大模型诊断能力；进一步采用 *WiSE-FT* 对 SFT 与 KD 阶段 LoRA 适配器进行权重插值，提升模型在未知故障类型下的泛化能力。
- **实验评估**：在 1600 条训练数据上完成模型适配，优化后的 SLM 经 320 条独立测试样本评测，取得 0.949 Macro-F1 与 100% 严格 JSON 输出率，性能逼近 Qwen3.5-27B 教师模型，并**超越** GLM-4-32B、Gemma-4-31B 等 30B 级参数量模型。同时，在**仅引入** 0.037 Macro-F1 性能损失的前提下，完成 Q4_K_M GGUF 量化部署（529MB），并于 Apple Silicon（ARM64 CPU）环境下达到 115 tok/s 推理速度与 1.135 GiB 峰值内存占用，最后结合置信度路由构建边云协同诊断机制，在 42.5% 边缘覆盖率场景下，将 Macro-F1 提升至 0.966（+0.054）。

> **其他成果**
> 1. SCI 一区论文一篇（学生一作，在投 *EAAI*）：传感器网络数据异常检测方向
> 2. SCI 二区论文一篇（学生二作，Early Access，*IEEE Sensors Journal*）：结构健康监测数据完整性方向
> 3. IEEE MLNLP 会议论文一篇（第二作者，已发表）：物联网跨设备数据安全方向
> 4. 一项发明专利已进入实质审查阶段（202610620996.9），一项软件著作权已登记（2025SR0034269）

专业能力

- 熟悉 LLM 基础原理与推理优化技术，掌握模型部署、量化压缩、投机解码及性能评测方法
- 具备 Agent 系统设计与开发经验，熟悉 Skill 开发、工具调用、上下文管理、多智能体协作及安全机制设计
- 掌握大模型微调技术体系，熟悉 SFT、LoRA、知识蒸馏、权重插值等方法，具备模型训练、评测优化与部署实践经验
- **英语能力**：CET-4 **512** 分，CET-6 **540** 分，2024 年考研英语二 **85** 分，具备适应英文工作环境的能力

奖项与证书

- **MOOC 认证**：自主学习大模型应用开发最新技术，并在 Coursera、Udemy、Harvard 等平台取得多项课程证书
- **竞赛**："建行杯"四川省国际大学生创新大赛（2025）**铜奖**（微岩精灵-薄片微观图像全自动鉴定引领者）
- **奖学金**：2024、2025 学年研究生三等学业奖学金，本科期间获院级二等、三等奖学金各一次