

温 辉

手机: 15030385844 | 邮箱: 15030385844@163.com | 意向地: 天津 / 北京 | 到岗: 可尽快

个人博客: <https://margin-blog.pages.dev/> | GitHub: github.com/wh-hd

求职意向: AI Infra / 推理优化工程师 (实习 / 校招)



教育经历

天津科技大学 | 软件工程 | 硕士研究生

2024.09 - 2027.07

加权平均分 91.38 (专业排名第三); 研究方向: 大模型应用、多模态模型与医学图像智能分析

南昌工程学院 (现江西水利电力大学) | 软件工程 | 本科

2019.09 - 2023.07

专业技能

- 编程与框架:** 熟悉 Python 算法开发, 掌握 NumPy、Pandas、Scikit-learn 等数据处理工具; 熟悉 PyTorch 模型构建、训练、验证与推理流程, 具备深度学习实验调试与模块联调经验。
- 大模型与 Transformer:** 熟悉 Transformer 与 Attention 机制, 理解 Self/Cross/Multi-Head Attention、Causal Mask、Position Encoding、KV Cache 等; 理解 Attention 计算复杂度与 KV Cache 显存模型, 熟悉长上下文推理的瓶颈构成 (显存带宽、KV 膨胀、attention 占比); 了解 BERT、GPT、LLaMA、Qwen 等模型的架构差异与推理范式。
- RAG 与 Agent:** 熟悉 RAG 与 Agent Workflow, 掌握医学知识检索、Prompt 构造、任务规划、工具调用、结果总结与评估; 熟悉 Function Calling、Planning、Reflection、Evaluator 等机制。
- 多模态模型:** 熟悉 Qwen-VL、BioMedCLIP 等视觉语言模型与图文语义对齐, 理解 Vision Encoder + LLM、视觉问答与多模态推理范式, 具备医学影像多模态实践经验。
- 深度学习:** 熟悉 CNN、U-Net、ResNet、TransUNet 等算法; 掌握 LoRA / PEFT 高效微调与 Unsloth 量化微调, 以及 Dice Loss、Cross-Entropy Loss、Dice、IoU 等损失函数与评估指标。
- 推理优化与部署:** 熟悉 Linux、Git、Docker; 具备 llama.cpp / vLLM 源码级集成经验, 掌握 KV Cache 压缩、块稀疏注意力、投机解码的实现与实测方法; 熟悉 PPL 无损验证、纯算子基准分离等度量方法论。

实习经历

OpenBayes 贝式计算 · 算法实习生

2026.06 - 2026.09

【背景】本地大模型推理存在三类结构性开销: 长上下文下 **KV Cache** 线性膨胀挤占显存、**decode** 逐 token 生成受**显存带宽**制约、超长上下文中 **attention** 计算占比失衡。

【任务】负责在 **llama.cpp** 官方基线上完成三类推理优化的**源码级集成与实测** (**KV 选择性压缩** / **块稀疏注意力** / **检索式投机解码**), 并独立完成**本地部署性能评测与模型能力基准**。

① **llama.cpp 长上下文推理优化 (KV 压缩 / 投机解码 / 块稀疏注意力)**

- KV 选择性压缩 (FastKV):** 以 **K-近似-Q 注意力** 打分选出 top 保留集 + 尾部窗口, **prefill 阶段**触发 **KV Cache 物理压实**并保留绝对位置语义; 实测保留率精确 (25% / 50%), **16K 上下文 decode 提速 15-19%**。
- 块稀疏注意力 (XAttention):** 基于**反对角折叠**打分 (get_rows 抽样 + 单次 mul_mat, 评分开销降至全量 1/S) 选 **top-K 保留块**, prefill 走 **gather 紧凑 K/V + 因果掩码**; 纯 attention 层 **2K→8K 实测加速 5.1×→19.6×**, 随上下文线性放大。
- 检索式投机解码 (REST):** 以 **libsais 后缀数组**离线构建 **datastore**, 生成时检索草稿 (**token span 递减回退 + Trie 多候选**), 目标模型一次前向完成**批量树验证**; 输出分布与基线完全一致 (**PPL 6.4636 无损**), 在长程一致 **datastore** 下具备 **1 + n_acc / n_call** 的加速空间。

【结果】

• 三类优化均以 **C++** 在 **llama.cpp** 官方基线完成**源码级集成**, 贯通「参数层 → 配置传递 → KV Cache 打分与压实 → decode 触发」全链路, **可独立开关、on/off 精度无损**。

② **本地大模型部署性能评测与编程能力基准**

• 基于 **vLLM** 完成 **DeepSeek-V4-Flash** 本地部署, 以 **vllm bench** 评测**吞吐 / 延迟**, 横向对比 **MTP** 与 **DSPark** 两类**投机解码**方案。

【结果】

- 实测 **MTP 多预测 1 个 token** 配置在本环境下**吞吐最优**, 形成**推理后端与量化档位选型**依据; 同步搭建 **SWE-bench** 自动化评测流程, 覆盖**环境构建 / 补丁生成 / 单元测试回归 / 失败归因**。

③ **确定性 GPU 模型交接 (Hermes Agent)**

- 将模型切换建模为「**有边界、可恢复、失败关闭**」的**前台事务**: 推理决策在 LLM 在线时冻结, 离线阶段由**确定性 Shell 调度器**执行; 以 **docker model ps + /engines/ps** 双视图三次稳定采样验证状态。

【结果】

- 实现**三层防重** (终身 claim + 并发锁 + 任务哈希双校验) 与**四层返回码**分层收敛, 完成**结构化 OCR 真实交接**, 沉淀为可复现开源方案 (github.com/wh-hd/agent_model_swap)。

天津水滴医学影像诊断有限公司 · 淋巴瘤 PET-CT 图像分割算法研发 (实习)

2024.09 - 2025.07

【背景】淋巴瘤 PET-CT 灰度图像存在目标与背景对比度低、边界模糊、前景占比小、标注有限等问题，传统 U-Net 在低对比度小样本场景下分割精度不足，难以满足影像辅助诊断勾画需求。

【任务】负责淋巴瘤 PET-CT 医学图像分割算法研发，在有限标注下提升病灶局部细节感知与全局语义建模能力。

【行动】

- 架构设计**：以 TransUNet 为基座，融合 ResNet50 编码器、Transformer 增强模块与 Attention Gate 解码器，兼顾 CNN 局部特征与 Transformer 全局依赖建模。
- 多尺度编码**：采用预训练 ResNet50 提取 5 个尺度特征并冻结早期层，保留预训练视觉知识，缓解小样本训练不稳定与过拟合。

【结果】

- 在医院测试数据集下，Dice 系数达 85%+、IoU 达 80%+，优于传统 U-Net 基线。
- 验证了 CNN 局部建模与 Transformer 全局建模结合在低对比度医学图像分割中的有效性，产出相关发明专利，并将算法集成到系统实现中。

项目 / 科研经历

多模态医疗诊断 Agent 工作流 (核心算法与 Agent 工作流开发)

2025.09 - 2026.05

【背景】医学影像辅助诊断存在知识密集、多模态融合困难、流程复杂、可解释性不足等痛点，纯生成式诊断易产生知识幻觉且过程不可追溯。

【任务】构建端到端多模态医疗诊断 Agent 工作流，覆盖医学知识检索、诊断任务规划、病例级工具调用到结构化结果总结与评估。

【行动】

- 层级化 Workflow**：构建 Task-level Planning 与 Case-level Diagnosis 两级流程，将复杂诊断拆解为知识检索、图像预处理、特征提取、定量分析、VLM 判断与结果汇总等可执行步骤，支持固定与动态计划。
- 医学 RAG**：基于 PubMedBERT 对医学知识库/诊断指南向量化检索，为 Planner 提供外部证据，降低大模型直接生成诊断流程的知识幻觉风险。
- 多模型 VLM Decider**：集成 Qwen-VL / BioMedCLIP 构建多模型决策器，用于医学图像理解、图文语义对齐与病例级诊断推理。
- 工具链封装**：实现分割工具、图像预处理、定量计算等 Function Calling 工具，形成「医学知识 + 图像证据 + VLM 推理」的诊断链路。

【结果】

- 在测试数据集完成 400 例端到端评估闭环；以 Qwen2.5 为推理模型时 F1=60%、Balanced Accuracy=73%。
- 优于 gpt-4o 直接推理 (56.4 / 21.1)，验证「检索 + 计划 + 工具 + 评估」闭环相对纯生成的价值。

论文 / 专利 / 竞赛

论文 (课题)：以第一作者提出 MMITA《基于双路径共享-私有特征解耦的医学图像-文本跨模态对齐方法》；另发表论文 3 篇。

发明专利：已发表发明专利 5 项，代表性含《一种基于强化学习的多模态诊断治疗方案推荐系统》等。

竞赛：蓝桥杯程序设计竞赛省级二等奖 / 三等奖；中国国际大学生创新大赛（原互联网+）天津赛区铜奖。

个人项目：智医助手（家庭健康管理 AI 应用）、基于 Unsloth 的 Qwen 模型 LoRA 量化微调等开源项目。

荣誉证书

CET-4 / CET-6；华为昇腾人工智能技能认证：Ascend C 算子开发能力认证（中级）。

个人优势

- 人工智能科班（天津科技大学人工智能学院硕士），聚焦大模型推理优化，具备「论文调研 → 源码级集成 → 同口径实测 → 瓶颈归因」完整闭环能力。
- 在 llama.cpp 官方基线上完成 KV 选择性压缩、块稀疏注意力、检索式投机解码三类优化集成，并给出可复现的量化收益与失效边界。
- 熟悉 KV Cache 显存模型与 attention 计算复杂度，掌握 PPL 无损验证、纯算子基准分离、prefill / decode 分测等度量方法论。
- 自驱力强，习惯以最小可复现样例定位与验证问题，并沉淀为可审计、可复现的工程方案。