龙维汉 (Weihan Long)

LLM 系统与算法工程师

(+86) 187-7369-0020 | weihanlong@std.uestc.edu.cn | https://longweihan.github.io/ai

擅长领域: LLM 推理系统优化 / RAG 架构 / 高性能算子开发

个人总结

- **全栈开发**: 专注大模型推理效率与 RAG 可靠性，拥有扎实的计算机系统底层功底、前沿 NLP 算法视野。
- **RAG 治理**: 针对 Agentic RAG 痛点，设计证据治理层，大幅提升生成准确率并降低幻觉。
- **性能优化**: 深入 Tokenizer 与 Softmax 算子底层，利用前缀缓存和位级操作实现数十倍性能跃升。
- **国际视野**: 具备从 Kernel 到 Agent 策略的全链路开发能力，拥有剑桥大学访学背景。

教育背景

剑桥大学 (University of Cambridge) 预期 2026 – 2027
联合培养硕士 *(CSC 全额奖学金)*

电子科技大学 (UESTC) 2024 – 2027
硕士 / 电子科学与技术 *(A+)* / ***GPA: 3.86/4.0*** *(保送)*

电子科技大学 (UESTC) 2020 – 2024
学士 / 电子科学与技术 *(A+)* / ***GPA: 3.82/4.0*** **荣誉:** 荣誉学士学位 (学院 Top 2)，优秀毕业生

技术栈

- **核心领域:** LLM Inference Optimization, RAG, AI Agents
- **编程语言:** Python (Expert, 复杂系统构建), C/C++ (Kernel/System Level), CUDA (Basic)
- **框架与工具:** PyTorch, HuggingFace Transformers, LangChain, vLLM, DeepSpeed
- **底层技术:** 数值计算稳定性 (IEEE-754), 分词算法 (BPE), 向量数据库 (Milvus/Faiss)

核心项目经历

OverSearchGuard: Agentic RAG 的抗冲突证据治理系统 独立开发

- **背景**: 针对多跳检索中常见的错误信息重复攻击和上下文爆炸问题，设计位于检索与生成间的中间件。
- **冲突感知稀疏化 (CACT)**: 提出基于源可靠性与时效性的筛选算法。在 Google Flan-T5 模型上，将高噪声环境下的回答准确率从 **14.3% 提升至 97.0%**。
- **成本优化**: 引入预算感知证据累积机制。将平均 Token 消耗量 **降低 76%**，显著降低 API 成本。
- **架构设计**: 实现 “检索 → 治理 → 生成” 流水线，支持人类偏好对齐，解决长文本生成 “中间迷失” 问题。

FlashToken: 零模型修改的 Tokenizer 前缀缓存系统 独立开发

- **背景**: 解决多轮对话与 Agent 思考过程中 System Prompt 重复 Tokenize 导致的首字延迟问题。
- **前缀缓存架构**: 设计 FixedPrefix 和 AppendOnly 策略，利用 Trie 树对重复前缀文本进行 ID 映射复用。
- **性能突破**: 对长 System Prompt (2000+ 字符) 场景，实现 **27x - 38x** 的编码速度提升 (1580ms → 57ms)。
- **兼容性**: 保证与 OpenAI tiktoken 的逐 Token 级精确一致性 (**0 Mismatch**)，可无缝集成至 vLLM/TGI。

Turbo-Softmax: 嵌入式/移动端的高性能 Softmax Kernel 独立开发

- **底层数学优化**: 利用 IEEE-754 浮点数特性的位级操作构建 2^i 快速近似，结合 5 阶多项式拟合 e^t。
- **极致性能**: 在无 SIMD 硬件加速的通用 MCU 环境下，相比标准库实现 **4.2x 加速**。
- **数值稳定性**: 最大误差控制在 10^{-6} 以内，KL 散度误差忽略不计，确保 Transformers 推理精度无损。

荣誉与奖项

- **Global Top 5%**, IEEEXtreme Programming Competition (IEEE 极限编程大赛), 2021.
- 荣誉学士学位 (学院 Top 2), 优秀毕业生, 荣誉科研证书.
- **已受理发明专利**: 一种基于 AI 的半导体器件设计优化系统及方法.