Weihan Long

LLM Systems & Algorithms Engineer

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

Focus Areas: LLM Inference Optimization / Agentic RAG / High-Performance Kernels

Summary

- **Full-Stack Engineering:** Specialized in Large Language Model (LLM) inference efficiency and RAG reliability, combining low-level system optimization with frontier NLP algorithm design.
- **RAG Governance:** Developed robust evidence governance layers for Agentic RAG, solving "wrong-but-repeated" hallucinations and optimizing context window usage.
- **System Optimization:** Deep expertise in Tokenizer and Softmax kernel optimization, utilizing prefix caching and bit-level manipulation to achieve multi-fold speedups.
- **Global Perspective:** Visiting student at the University of Cambridge; capable of delivering solutions from bare-metal kernels to high-level agent strategies.

Education

University of Cambridge Expected 2026 – 2027
Visiting Student (CSC Full Scholarship) Cambridge, UK

University of Electronic Science and Technology of China (UESTC) 2024 – 2027
MPhil Electronic Science and Technology (Grade A+) | ***GPA: 3.86/4.0*** Chengdu, China

University of Electronic Science and Technology of China (UESTC) 2020 – 2024
Bachelor in Electronic Science and Technology (Grade A+) | ***GPA: 3.82/4.0*** Chengdu, China **Honors:** Honors Degree (Top 2 in College), Outstanding Graduate.

Technical Skills

- **Core Domains:** LLM Inference Optimization, RAG (Retrieval-Augmented Generation), AI Agents.
- **Languages:** Python (Expert, System Architecture), C/C++ (Kernel/System Level), CUDA (Basic).
- **Frameworks:** PyTorch, HuggingFace Transformers, LangChain, vLLM, DeepSpeed.
- **Fundamentals:** Numerical Stability (IEEE-754), Tokenization (BPE), Vector Databases (Milvus/Faiss).

Key Projects

OverSearchGuard: Conflict-Aware Evidence Thinning for Agents *Independent Developer*

- **Problem:** Addressed "over-searching" and "context window pressure" in Agentic RAG, where conflicting or repeated low-quality evidence leads to hallucinations.
- **Algorithm (CACT):** Pioneered *Conflict-Aware Candidate Thinning* to filter evidence based on source reliability and recency. Improved response accuracy from **14.3% to 97.0%** on high-noise benchmarks (Google Flan-T5).
- **Cost Optimization:** Implemented *Budgeted Evidence Accumulation (BEA)* to iteratively query LLMs only when necessary, reducing average token consumption by **∼78%** and optimizing the TPC (Total Tokens per Correct Answer) metric.
- **Architecture:** Designed a drop-in governance layer (Retrieval → Thinning → Generation) robust to "wrong-but-repeated" duplicate attacks.

FlashToken: Tokenizer-Side Prefix Caching System *Independent Developer*

- **Problem:** Mitigated latency in ReAct loops and multi-turn chats where re-tokenizing massive system prompts or tool outputs creates CPU bottlenecks.
- **Solution:** Developed a prefix caching library with *FixedPrefix* and *AppendOnly* strategies, utilizing Trie

structures to reuse stable token IDs.

- **Performance:** Achieved **27x − 37x speedup** (1580ms → 57ms) for long context reuse scenarios.
- **Reliability:** Guaranteed **Zero-Mismatch** correctness (token-for-token equality) with OpenAI's `tiktoken` standard, compatible with KV-cache workflows.

Turbo-Softmax: High-Precision Fast Softmax for MCUs *Independent Developer*

- **Core Optimization:** Implemented a blazing fast Softmax kernel for generic MCUs using *Range Reduction* (2^i via **IEEE-754 bit-level manipulation** and e^t via 5th-order polynomial approximation).
- **Speedup:** Achieved **4.0x − 4.2x speedup** over standard `math.h` implementations on hardware without SIMD/FPU acceleration.
- **Stability:** Maintained strict numerical stability with Max Error $< 10^{-6}$ and negligible KL-Divergence, ensuring consistent Float32 inference results.

Honors & Awards

- **Global Top 5%**, IEEEXtreme Programming Competition, 2021.
- Honors Research Certificate, Outstanding Graduate Award (Top 2 in College).
- **Patent Pending:** A System and Method for AI-Based Semiconductor Device Design Optimization.