GAURAV GUPTA

gauravkumargupta6789@gmail.com +91 8143470609 LinkedIn GitHub LeetCode

Summary

AI Software Engineer with around 2 years of experience, previously worked at Infosys (Specialist Programmer). Skilled in working with LLMs, LoRA image models, and WAN video models, with a strong focus on improving inference efficiency through improved memory layouts, vectorized execution, and low-level performance tuning. Experienced in building RAG pipelines and LangChain-based workflows. Brings a solid foundation in systems engineering, distributed environments, GPU parallelism, compiler-level optimization, and cloud-native development practices.

Experience

DHEYO AI, AI Software Engineer (Hyderabad) April 2025 - Present

- Optimized the WAN video model by replacing fused Triton kernels with custom low-level implementations, achieving 20–25% faster inference in production scenarios.
- Designed and maintained Mojo-based execution graphs for multiple model architectures, improving memory efficiency by ~15% and ensuring more stable execution.
- Enhanced multi-model serving pipelines, enabling multiple LLMs to run within a shared runtime and increasing end-to-end throughput by 30%.
- Worked hands-on with LLMs and LoRA image models, focusing on improving inference through better memory layouts, vectorized execution, and targeted performance tuning.

INFOSYS LIMITED, Specialist Programmer (Hyderabad) Aug 2024 - April 2025

- Developed a LangChain-driven retrieval pipeline with FAISS and custom document processing, improving factual accuracy by ~25%.
- Streamlined LLaMA-70B model inference and contributed to a custom LLM compiler by improving tokenization, graph loading, and memory layouts.
- Created and deployed FastAPI services with Redis caching and Docker, reducing overall query latency by 35–40%.

JUSSEE AI, AI Software Intern (Hyderabad) Jan 2024 - Aug 2024

- Fine-tuned a LLM compiler in C and Rust enhanced for x86_64, leveraging Intel AMX and AVX-512 to accelerate matrix multiplication and improve inference performance.
- Improved memory efficiency by reducing STRIDES, adding data prefetching, and optimizing kernel-level matmul operations, which accounted for 85% of runtime.
- Achieved ~30% overall efficiency gains compared to Intel's baseline through custom kernel optimizations and performance-driven cost modelling.

Projects

CONTEXT-AWARE DOCUMENT Q&A SYSTEM USING LANGCHAIN & RAG [Project Link]
(Compact RAG pipeline for fast retrieval over enterprise documents)

- Built a lightweight RAG system using LangChain and FAISS with custom chunking, embeddings, and metadata tagging, improving retrieval accuracy and reducing irrelevant responses by ~20%.
- Designed a FastAPI backend with endpoints for querying and document updates, containerized with Docker to enable fast, consistent deployment and easy integration.

Tech Stack: Python, LangChain, FAISS, FastAPI, Docker, HuggingFace Embeddings

EXPERIMENTING WITH THE CLOUD CHARACTERISTICS IN KUBERNETES CLUSTER [Project Link]
(Auto-Scaling Reddit Clone on Kubernetes by using AWS service)

- Implemented a scalable, fault-tolerant Reddit-clone infrastructure using AWS, Kubernetes, Docker, and Minikube, improving deployment consistency by 30–40% in test environments.
- Delivered an auto-scaling setup that reduced manual intervention by 50% and improved overall reliability and uptime across the application.

Tech Stack: Git, Docker, Kubernetes, AWS

Skills

AI & ML	:	LLMs (LLaMA), LoRA (image-based models), WAN (video model), Deep Learning (DL), RAG (Retrieval-Augmented Generation), LangChain, Transformers, Inference Optimization, Model Compilation.
Programming	:	C, C++, Rust, Python, SQL.
Systems & Optimization	:	AVX-512, Intel AMX, CUDA (basic GPU programming), Kernel Optimization, Memory Optimization, HPC concepts, Triton, Mojo.
DevOps & Cloud-Native	:	Docker, Kubernetes, CI/CD, Linux.
Tools & Technologies	:	Git, GitHub, RabbitMQ.
Core CS Concepts	:	Data Structures & Algorithms, Operating Systems, DBMS, Object-Oriented Programming (OOP).

Education

University of Hyderabad, Hyderabad, Telangana, India 2022 - 2024
Master of Computer Application, CGPA: 08.50

Bundelkhand University, Jhansi, Uttar Pradesh, India 2019 – 2022
Bachelor of Computer Application, 80.99%

Achievements

- Gold Medallist in Under Graduation.
- Indian and National level in NCSC (National Children Science Congress) awarded by CM in Lucknow.