Sandeep Kumar

Research Scientist @ Intel, Bengaluru, India | Ph.D. CSE

sandeep007734@gmail.com | (+91) 8277361995
sandeep007734.github.io | github.com/sandeep007734 | linkedin.com/in/sandeep007734

Cover Letter

Why I'm excited about this role
I am excited by the engineering challenges at the intersection of large-scale model training/inference and high-performance systems—especially making multimodal LLMs fast, reliable, and cost-efficient on highly interconnected GPU clusters. I enjoy work that requires tight feedback loops between profiling, algorithmic choices, and low-level systems optimization.

Relevant experience I bring
I am a systems researcher and performance-focused software engineer at Intel Labs, working on AI/ML efficiency with a particular focus on memory behavior (memory tiering for LLMs and RAG systems) and production-grade systems work (including upstream Linux kernel memory-management contributions). I regularly profile and debug across the stack (OS/kernel, drivers/accelerators, userspace) and use telemetry to translate bottlenecks into targeted optimizations.
I have hands-on experience integrating and tuning hardware accelerators (Intel IAA/DSA) for real workloads, and I am comfortable making kernel- and driver-adjacent changes when those are the source of performance or correctness issues. During my PhD, I also worked on low-level systems around Intel SGX (e.g., enclave performance/benchmarking and OS-facing mechanisms), which strengthened my ability to debug across privilege boundaries.
I build and optimize Python-based AI workflows (including PyTorch training/inference) and go deeper, when needed, into native code and systems configuration. For DNN training, I have worked directly with PyTorch-based training code, including profiling/instrumentation and targeted changes to reduce memory pressure and improve efficiency. I am comfortable reasoning about distributed execution fundamentals and excited to apply that to GPU-centric stacks (e.g., PyTorch Distributed with NCCL) and modern large-scale training/inference frameworks.

How I would contribute in your stack

- **Scale-out training:** Improve parallel efficiency by profiling communication/memory bottlenecks and validating scaling strategies with rigorous measurement.
- **High-throughput inference:** Tune multimodal LLM serving for throughput and tail latency, leveraging my experience optimizing LLM inference and RAG systems.
- **Deep performance work:** Debug and optimize low-level behavior (kernels, drivers, OS configuration) to remove system bottlenecks.

I would welcome the opportunity to discuss how my systems performance background in AI infrastructure can help you push the boundaries of what is computationally possible.

Sincerely,

Sandeep Kumar
Research Scientist @ Intel, Bengaluru, India
Ph.D. CSE