Nikhil Jain

PROFILE

Professional software engineer with 4 years of industry experience in distributed systems and compilers, pursuing graduate work in performance-critical systems for LLM inference and AI acceleration.

EDUCATION

University of California, Santa Cruz – M.S. Computer Science • 4.0 / 4.0 GPA • Sept 2025 – May 2027
University of California, Berkeley – B.A. Computer Science • 3.6 / 4.0 GPA • Aug 2017 – May 2021

OPEN SOURCE / CURRENT RESEARCH EXPERIENCE

llama.cpp • *Dec 2025 – present*
https://github.com/ggml-org/llama.cpp

• Enabled WebGPU backend multithreading by refactoring shared state into per-thread/per-context state; reduced contention and clarified concurrency model.
• Implemented/fixed resource lifetime management in asynchronous WebGPU execution (device/queue/buffer lifecycle, teardown/free paths) to prevent leaks/crashes and improve robustness across browsers.
• Parameterized buffer pools for efficient per-thread resource utilization

TinyLLM: Inference Runtime for Edge Applications • *Dec 2025 – present*
https://github.com/nikhilJain17/tinyLLM

• Building inference runtime from scratch in C++/CUDA
• Implemented a trie-based tokenizer; benchmarking throughput and latency across vocabulary sizes
• Implemented GGUF weight loader with instrumentation for page faults, RSS, and cold vs warm start behavior

PROFESSIONAL EXPERIENCE

Uber – Software Engineer II • *San Francisco, CA | Aug 2022 – Mar 2025*

• Contributed to a year-long rewrite of a tier-1 service to a geo-sharded architecture, reducing infrastructure costs and eliminating bespoke hardware dependencies.
• Designed and implemented global resource allocation automation, saving >1,000 servers and eliminating manual operational toil.
• Built large-scale streaming batch jobs processing hundreds of millions of driver location updates daily.

Sambanova Systems – Compiler Engineer • *Palo Alto, CA | Aug 2021 – Aug 2022*

- Implemented 13 hardware features into high-level domain-specific language, raising abstraction level from assembly to high-level code.
- Developed compiler passes for network bus assignment and register allocation, reducing programmer errors and manual effort.
- Contributed to compiler-wide testing infrastructure, eliminating 20+ production bugs.
- Contributed to experimental JIT compiler enabling interactive, notebook-style development.

Microsoft – Compiler Engineer Intern • *Jun 2020 – Aug 2020*

• Researched methods to parallelize Visual Studio's C++ linker, achieving a 25% speedup.
• Conducted extensive performance profiling and scoped future optimization work.

SKILLS

C++ • Go • Python • Java • WebGPU (runtime) • MLIR • Multithreaded & asynchronous execution • Inference systems