

Siddharth Sharma

✉ mail |  linkedin |  github |  LeetCode |  portfolio

EDUCATION

Thapar Institute of Engineering and Technology

Patiala, Punjab

Bachelor of Engineering in Computer Engineering (CGPA: 8.1 / 10.00)

Expected May 2027

- Relevant Coursework: Data Structures & Algorithms (C++), Object-Oriented Programming (C++), Operating Systems, Computer Networks (TCP/IP, Socket Programming), DBMS, Probability & Statistics

TECHNICAL SKILLS

Languages: Python, Go (Golang), C, C++ (C++11/14/17, STL, RAII, OOP), CUDA, Bash, SQL, JavaScript

AI/ML: PyTorch, TensorFlow, LLMs, Transformer Models, RAG, Inference Optimization, Quantization (GGUF/AWQ), Continuous Batching, Model Serving

Tools & Environments: GCC (O2/O3), GDB, Valgrind, Docker, Kubernetes, Linux (Ubuntu/Debian), Git, Prometheus, Grafana

Technologies: Apache Spark, Snowpark, LangChain, llama.cpp, FastAPI, Django, Flask, Node.js, gRPC, WebSockets

ACHIEVEMENTS

Competitive Programming | LeetCode: nowadaysidd (Knight)

- LeetCode Knight** — contest rating **1906** (top ~4% globally); **860+** problems solved primarily in C++ with STL mastery and memory-efficient implementations; **500+** day daily streak
- Participated in **19+** weekly contests; consistent solver of medium-hard problems under timed conditions

PROJECTS

ColumnarDB: In-Memory Columnar Query Engine | C++17, SIMD/AVX2, MVCC, B+ Tree | GitHub Dec 2025 – Feb 2026

- Built a C++17 columnar storage engine with a **64B-aligned slab allocator**, reducing memory allocation overhead by **3×** over `std::malloc` in micro-benchmarks; architecture mirrors low-latency data serving patterns in distributed inference pipelines.
- Implemented **AVX2-vectorized** scan kernels processing 8 int lanes per instruction, achieving **5.4 GB/s** single-core throughput for integer equality and range scans.
- Built a **SQL pipeline** (Parser, Cost-based Planner, B+ Tree) with **MVCC** and **Epoch GC** for lock-free concurrency; validated with **100+** tests and 100k+ assertions with zero memory leaks.

Alfred: Agentic Multi-Modal Knowledge Assistant | Python, Next.js, Expo, Qdrant, FastAPI | GitHub Mar 2026 – Apr 2026

- Architected an **end-to-end ML inference pipeline** integrating Deepgram STT, OCR, semantic chunking, and vector embedding into a unified ingestion service with resilient upload queuing and retry logic.
- Built a **model serving layer** using FastAPI + WebSockets for streaming LLM responses; integrated **Qdrant** for semantic retrieval and context-aware RAG, optimizing retrieval latency for dynamic business-user queries.
- Delivered the full stack across a **React Native mobile app** and **Next.js web dashboard**, shipping a maintainable, cross-platform agentic AI product end-to-end.

SatBot: Distributed Offline AI Robot | Python, C++, Vector DB, Offline RAG, Linux | GitHub Oct 2025 – Nov 2025

- Deployed a **fully offline ML inference stack** on CPU-only Linux hardware — STT, embeddings, local vector DB, and TTS — achieving **end-to-end response latency under 3 seconds** via pipeline parallelism and I/O overlap.
- Orchestrated a **microservices architecture** with Bash-based watchdog monitoring and automated recovery; applied **quantization (GGUF)** and **continuous batching** to maximize throughput within strict memory budgets.
- Engineered an extensible NLP inference layer with seamless model swapping and context-window management; automated service startup and zero-downtime restarts across all modules.

EXPERIENCE

Data Structures and Algorithms Mentor

Jun. 2025 – Aug. 2025

CCS Coding Community, Thapar Institute of Engineering and Technology

Patiala, Punjab

- Mentored **50+** students in DSA concepts using C++ (arrays, graphs, DP, trees); conducted daily doubt-clearing sessions and tracked individual student progress

Core Team Member

Jul. 2024 – Present

Creative Computing Society, Thapar Institute of Engineering and Technology

Patiala, Punjab

- Co-organized **HackTU National Hackathon** with **500+** participants; managed logistics, promotions, and provided onsite dev support