SHREYAS SMARTH

smarth.shreyas@gmail.com linkedin.com/in/shreyas-smarth github.com/Shrot101

SUMMARY

CSE student (CGPA 8.77) with hands-on experience in ML systems, deep reinforcement learning, and LLM inference optimization. Built end-to-end AI infrastructure spanning LLM middleware, adaptive ML pipelines, and RL-based controllers. Interested in AI systems engineering and distributed ML infrastructure. Also experienced in Android development and open-source contribution.

EDUCATION

B.Tech Computer Science Engineering Oct 2024 – Present
Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded CGPA: 8.77/10

TECHNICAL SKILLS

Languages: Python, Java, Kotlin, C++, C
AI/ML: PyTorch, Deep Learning, Reinforcement Learning (A3C), CNNs, Transformers, NLP
LLM & AI Systems: LLM Routing, Semantic Caching, Inference Optimization, LLMLingua, Optuna (Bayesian Opt.), Vector Search, Sentence Transformers
Backend & Infra: FastAPI, REST APIs, Redis, Docker, Client-Server Architecture
Mobile: Android (Java/XML), Flutter **Core CS:** DSA, OOP, OS, DBMS, Computer Networks
Tools: Git, Linux CLI, HuggingFace, scikit-learn, NumPy, Matplotlib

EXPERIENCE

Software Engineering Intern — Sumago Infotech Pvt Ltd (Remote) Sumago

- Applied structured version control workflows and modular solution design within a production codebase.
- Strengthened understanding of scalable application architecture through collaborative engineering practices.

PROJECTS

LLMOpt — Adaptive LLM Inference Optimization Framework llmopt

- Engineered a multi-stage AI middleware pipeline routing LLM queries by semantic complexity via NLI-based analysis (DistilRoBERTa), Gradient Boosting complexity estimation, and Bayesian weight optimization (Optuna).
- Built a Redis-backed semantic cache using Sentence Transformers (all-MiniLM-L6-v2) with cosine similarity; cache hits return responses at $0.00 cost and near-zero latency, eliminating redundant API calls.
- Integrated LLMLingua-2 for semantic token pruning (up to 40% prompt compression) with V1 heuristic fallback; deployed via FastAPI with BYOK support and per-request explainability endpoints.

ABREngine — Deep RL for Adaptive Video Streaming abrengine

- Implemented an end-to-end A3C (Asynchronous Advantage Actor-Critic) agent in PyTorch, reproducing Pensieve (SIGCOMM '17) with log-scale QoE reward formulation modelling perceptual diminishing returns.
- Designed a multi-branch actor-critic network with parallel Conv1D towers for throughput, download time, and chunk-size inputs; entropy annealing (0.5→0.1) stabilized convergence across 5,000 training episodes.
- Achieved **95.45 QoE reward** (+9.5% vs. best heuristic) and **+18.6% higher avg. bitrate** (2,797 vs. 2,359 kbps) on 100 held-out episodes; evaluated against four baseline policies.

WellBeing — Android Social Media Platform wellbeing

- Sole designer and lead contributor; architected UI/UX and core feature set for a real-time Android social platform built with Java and XML layouts, integrating REST API-backed backend communication.

Open Source: AnkiDroid PR #18355

- Merged pull request to AnkiDroid (10M+ downloads); participated in technical code review on a large-scale open-source Android codebase.